Story Morals: Surfacing value-driven narrative schemas using large language models

David G Hobson[1], Haiqi Zhou[2], Derek Ruths[1], Andrew Piper[2]
[1]School of Computer Science
[2]Department of Languages, Literatures, and Cultures
McGill University

Abstract

Stories are not only designed to entertain but to encode lessons reflecting their authors' beliefs about the world. In this paper, we propose a new task of narrative schema labelling based on the concept of "story morals" to identify the values and lessons conveyed in stories. Using large language models (LLMs) such as GPT-4, we develop methods to automatically extract and validate story morals across a diverse set of narrative genres, including folktales, novels, movies and TV, personal stories from social media, and the news. Our approach involves a multi-step prompting sequence to derive morals and validate them through both automated metrics and human assessments. The findings suggest that LLMs can effectively approximate human story moral interpretations and offer a new avenue for computational narrative understanding. By clustering the extracted morals on a sample dataset of folktales from around the world, we highlight the commonalities and distinctiveness of narrative values, providing preliminary insights into the distribution of values across cultures. This work opens up new possibilities for studying narrative schemas and their role in shaping human beliefs and behaviors.[1]

1 Introduction

As the esteemed literary critic Wayne Booth once wrote, "All stories teach" (Booth, 1998). While we may associate the idea of story morals with short didactic fiction (i.e. fables), narrative theorists have long argued that all stories encode value-driven schemas in the form of lessons or morals (Booth, 1998; Nussbaum, 1998). As Russell and Van Den Broek (1992) argue, "Narrative schemas enable individuals to organize and represent experiences and/or events as meaningful wholes that function as the bases for comprehension and behavior." Stories provide templates for how to live.

In this paper, we propose a new task of narrative schema labeling based on the concept of "story morals." A story's moral can be thought of as a short, compressed lesson (often memorable) that a story aims to convey to its audience (see Table 1 for examples). The goal of communicating such lessons is one of storytelling's oldest known functions, dating back over 2,500 years (Gregory, 2010).

As a narrative schema, story morals differ from other frameworks because they focus on the *values* and *intentions* of the storyteller. The focus on a story's moral shifts the locus of attention from the question, "what happened?" to "why was this told?" or "what basic lesson am I supposed to learn from this?" Importantly, while story morals are related to questions of moral beliefs, story morals need not explicitly communicate moral sentiments (e.g. "Kindness is good"). Rather, they can address general lessons that may draw from, reinforce, or challenge larger moral frameworks. It is not uncommon, for example, for individuals to come away from reading a book or watching a movie and reflect on what the writer or director was "trying to say" (as opposed to asking, "What was the moral framework the author was using?") While this interpretive process depends on the specific narrative events, a story's moral is extrapolated from content-level features.

Story morals thus represent an important, yet challenging labeling exercise. As can be seen in Table 1, different readers may bring different judgments about the central lesson of a story, just as similar judgments may be rendered using different, yet semantically-related words. At the same time, story morals are a ubiquitous way of interpreting stories that are embedded in a variety of everyday cultural documents such as movie and book reviews and classroom support material.

[1]Code available at https://github.com/davidghobson1/llm-story-morals

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12998–13032

The Lost Camel: *A king is initially skeptical of four people who are accused of stealing a camel. Impressed by their keen observations and a defence that absolves them of guilt, he recompenses the man who lost the camel, and promotes the four individuals to be his ministers.*	The US's budget deal is a victory for Mike Johnson. But for how long? : *US House Speaker Mike Johnson juggles new demands about a budget bill from members within his own party and those of the opposing party.*
What is the moral of the story?	
Good benevolent leadership pays off	**Your own team can turn against you**
A ruler must be just and righteous	**Conflict can occur within the same group**
Intelligence will be rewarded	Compromises are important to ensure effective conflict resolution

Table 1: Sample human morals from a folktale (left) and a news article (right). We provide a brief summary of the article here, while the full story is used in our prompting scenario. Morals in bold are more similar, however all morals can be seen as correct. Note that the goal of the training here was not to achieve agreement among the annotators (as differing interpretations were desirable) but to ensure consistent understanding of the definitions.

Given the importance and cultural prevalence of story morals along with the highly derivative nature of their production, we propose that LLMs are an ideal candidate for testing the task of automated story moral labeling. While LLMs continue to struggle with the problem of hallucination (Xu et al., 2024), the derivative nature of story morals, i.e. that a story's message is rarely explicitly stated in the story but must be derived from it, makes for a potentially good fit with LLM capabilities. Additionally, we assume that given the prevalence not only of stories, but also story-moral-like statements on the web, that this concept is likely well represented in LLM training data, with potential cultural biases that need to be assessed.

The ability to extract story morals across different kinds of text genres, cultures, and historical time periods can play an important role in supporting the project of computational narrative understanding (Zhu et al., 2023; Piper, 2023). With sufficient quality datasets, the ability to extract story morals can help us better understand the distribution of values across cultures and highlight where there are core differences and where humans share common ground as evidenced by the myriad stories they tell (Graham et al., 2013).

In this paper, we make the following contributions: first, we define the task of story moral labeling and create a dataset of human annotated story morals across a diverse set of story types (800 story morals across 6 genres and 144 unique stories including one foreign language (Mandarin)). Second, we implement both automated and human-level validation schemes for one candidate LLM (GPT-4) to illustrate the feasibility of using LLMs for this task. Third and finally, we illustrate the analytical affordances of LLM-assisted story-moral labeling on a diverse dataset of folktales from different regions of the world that have been translated into English.

2 Related Work

The organization of stories into broad, overarching categories is deeply rooted in the field of narratology (Genette, 1992; Thompson, 1955; Brewer and Lichtenstein, 1980; Campbell, 2008; Frye, 2020; Propp, 1968). Despite addressing narratives at varying levels of abstraction, these models converge on a fundamental premise: stories inherently share common elements, and their selection is orchestrated by higher-level schemas that shape the narrative's construction and interpretation.

In the field of NLP, work related to labeling narrative schemas ranges widely across a diverse set of approaches. Early work by Chambers and Jurafsky (2009) focused on narrative schema detection focused on identifying related event chains (Yan et al., 2019; Vauth et al., 2021; Sims et al., 2019). The chaining together of event schemas has been integral to operationalizing the concept of "plot" (Kukkonen, 2014), including plot summaries and plotlines (Rashkin et al., 2020; Anantharama et al., 2022).

Other work has focused on detecting higher-level schemas such as "conflict" and "resolution" (Fr-

ermann et al., 2023), turning points (Ouyang and McKeown, 2015), folktale motifs (Karsdorp and van den Bosch, 2013), narrative archetypes such as "rags-to-riches" (Reagan et al., 2016; Fudolig et al., 2023), and the more traditional concept of "genre" (Kundalia et al., 2020; Wilkens, 2016; Dai and Huang, 2021).

Closer to our work is research focused on identifying narrative intentions, i.e. surfacing the aims guiding why stories are told (Lukin et al., 2016; Zhu et al., 2023). Fu et al. (2019) have explored the detection of narrative intentions in advice-seeking contexts online. Their approach seeks to leverage common-sense reasoning to infer the implicit questions and aims embedded in personal stories (see also Mou et al., 2021). Antoniak et al. (2019) look at birth stories as a salient genre in which to recover author intentions in online public forums. Related to intention, Zhou et al. (2024) recently used the concept of story morals in their study of narrative messaging related to climate change.

Attention to story morals implicitly draws connections to related work on detecting moral sentiment in texts. Much of this work derives from the theoretical foundations of Graham et al. (2013), and has been applied to the study of texts such as tweets (Rezapour et al., 2019; Roy and Goldwasser, 2021; Roy et al., 2023), folktales (Wu et al., 2023), and across multiple domains (Liscio et al., 2022). Vida et al. (2023) provide a useful overview of the use of "morals" as a concept within NLP research.

In the next section we highlight how our work aligns and departs from this prior work.

3 Methods

3.1 Task Definition

We define a "story moral" as *a general lesson that the narrator wishes to impart to the audience about the world*. Central to this concept is the focus on a higher order value: lessons are meant to encourage or discourage certain behaviors, impart general wisdom to the reader, or influence their beliefs or worldview.

A central distinction here is that story morals understood as lessons mean that they are not strictly synonymous with the idea of moral "sentiments" (Vida et al., 2023). Instead they focus on forms of behavior and belief that may be integrated into or derived from pre-existing moral frameworks.

Another important aspect of this definition is that it does not rely on a pre-existing taxonomy of moral sentiments or foundations. As Dundes (1962) long ago highlighted, externally imposed narrative taxonomies are problematic because they deform or overlook important local knowledge. Our work seeks to derive narrative lessons from the bottom-up, which can then be integrated into existing systems or aggregated to produce novel frameworks of cultural values.[2]

Story morals thus reduce and distill a long text into a brief passage or statement, much like the task of narrative summarization (Ouyang et al., 2017). Unlike summarization, however, morals do not aim to reproduce the salient events of narrative content (i.e. "what happened"). Rather, they provide a high-level synthesis of the central message or lesson implicit in those events. While morals are dependent on narrative events and agent roles, different events can lead to similar morals. For example, Aesop's fable *The Tortoise and the Hare* and the movie *The Karate Kid* can both be interpreted as reinforcing the lesson of taking things slow and trusting the process ("slow and steady wins the race").

Nevertheless, the process of distilling an abstract lesson from a narrative document is open to interpretation. Similar to other subjectively informed NLP tasks (Basile et al., 2021), different readers may respond differently to the identification of a story's moral. Also similar to other subjective tasks, while there may not be a single right answer, there can be better or worse answers.

In Table 1, we show examples of student responses to the task of story moral extraction across two different kinds of narrative documents (*folktales* and *news*). First, we see that student annotations exhibit a high degree of relevance. After minimal training (meant to ensure consist interpretations of definitions), annotators are able to distill core narrative lessons across different kinds of genres. This highlights that story morals are a well-understood concept for human annotators.

Second, answers are almost never lexically identical but are often semantically similar. "Good benevolent leadership pays off" and "A ruler must be just and righteous" contain no lexical overlap but convey semantically similar messages.

Third, some answers can be semantically very different but still be relevant (e.g. "intelligence will be rewarded" is a reasonable moral for this story). Some answers, however, may be better than others

[2]The LLM's prior knowledge is an important mediator in this process, which we discuss in the Limitations section.

Category	Prompt
Summary	Can you summarize this story? State your answer as a single paragraph.
Agent	Who is the protagonist of this story? State your answer as a single name.
Agent	Is the protagonist a hero or a villain (i.e., are they portrayed positively or negatively), or are they a victim? You may choose more than one. If none, say none.
Agent	Who is the antagonist of this story? State your answer as a single name. If there is none, say none.
Topic	What is the central topic or issue of this story? State your answer as a single keyword or phrase.
Valence	Is this story more negative or positive? State your answer as a single number between 1 and 5 where 5 = very positive, 1 = very negative, 3 = neutral.
Moral	What is the moral of this story? State your answer as a single sentence.
Positive Moral (Moral+)	What is the moral of this story? State your answer as a single word or phrase followed by "is good behavior".
Negative Moral (Moral–)	What is the moral of this story? State your answer as a single word or phrase followed by "is bad behavior".

Table 2: Prompts used for narrative comprehension and story moral labelling.

in terms of their relevance to the text or the clarity with which they convey their message. Thus our validation exercise must attempt to understand the range of human responses, the capacity of LLMs to reproduce similar responses to that range, and finally the preferences of human readers with respect to potential candidate story morals.

3.2 LLM Implementation

We formulate our task as producing machine-generated story morals that are consistent with the distribution of human morals for the same story, that is, the range of responses given by humans.

Here, we offer a first solution to this task: we focus on the use of large language models (LLMs), specifically GPT-4, and the development of an appropriate prompting pipeline. While other decoder or encoder-decoder-based architectures, especially those based on summarization, could potentially be employed for this task, the lack of large-scale fine-tuning datasets with morals make such approaches difficult. The advantage of LLMs is that the understanding of the concept of "story morals" is likely well-encoded due to the concept's cultural ubiquity.

To surface the story moral for a given narrative input, and test general narrative comprehension, we employ the zero-shot prompting sequence given in Table 2. We first extract a summary to help the model focus on key narrative elements. We then identify principal agents, such as the protagonist and antagonist, and the central topic of the story to simulate a chain-of-thought style prompting sequence before extracting a free-form moral. To help narrow the space of possible moral outputs, we further extract moral keywords assuming a positive/negative valence based on the template "____ is good/bad behavior." We refer to these as the positive and negative morals, respectively (denoted for brevity as Moral+ and Moral–). While this prompting scheme is not explicitly motivated by literary theory, it is intended to promote chain-of-thought reasoning in the model.

All prompting exercises were done using GPT-4 (specifically, 0125-preview) through OpenAI's API and using a temperature of zero for consistency.

3.3 Validation

For the purposes of validation, we use a combination of automated metrics and human assessment based on Zhou et al. (2024). In the body of the paper we focus on the primary task of story moral labeling, and leave further discussion of the other narrative comprehension tasks to Appendix A.3 for interested readers.

3.3.1 Validation Data

In order to validate the task of story moral extraction, we use a test dataset of 144 documents drawn from five different narrative genres and cultural settings including: folktales, book and movie/TV summaries, personal stories from Reddit, and political news. The mean length of documents is 839 words with a minimum of 200 and a maximum of 2,200. The full breakdown by genre and source can be found in Appendix A.1.1.

To annotate our documents for the story moral task, a group of undergraduate student annotators were hired to provide answers to the prompts listed in Table 2 for each passage, however with the summarization question omitted. Student annotators were used, as opposed to domain experts, since they are educated but general readers, and since morals are intended to be interpretable by a general audience. Annotators were provided with a codebook of category definitions and examples (which can be found on our code repository), and underwent at least one round of practice annotations to affirm consistency of interpretations to the definitions. All human responses to the questions were open-response and made independently of each other and GPT-4. For more details on the annotators, see Appendix A.2.1.

3.3.2 Automated Validation

The purpose of automated validation is to provide an overview of the inter- and intra-group similarity between human and machine-labeled morals.

As we expect different human annotators to respond with different morals, we seek to model the semantic distribution of morals both among humans and between humans and GPT. As such, we primarily use techniques from semantic textual similarity (STS) and compare morals by the cosine similarity between their sentence embeddings. We employ pretrained embedding models from the SentenceTransformers library (Reimers and Gurevych, 2019) for this purpose; specifically averaged GloVe word embeddings (6b-300d) (Pennington et al., 2014), and the stsb-mpnet-base-v2 and nli-mpnet-base-v2 models (Song et al., 2020). As additional points of comparisons, we include standard evaluation metrics from summarization, including ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2019) which were both implemented using HuggingFace. BERTScore was implemented using the deberta-xlarge-mnli model (He et al., 2021). Although metrics like BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are also used in summarization, we only seek to use symmetric metrics as semantic equivalence is a symmetric relation.

	human-human	**human-GPT**	**GPT-GPT**
Rouge-L	0.00	7.41	52.94
BERTScore	56.17	57.94	81.81
GloVe	47.36	56.63	88.34
STSb-MPNet	25.14	37.14	83.63
NLI-MPNet	31.59	44.07	87.49

Table 3: Median rouge and similarity scores (out of 100) of pairwise morals between the different groups of annotators in the validation dataset. P-values for a Mann-Whitney U-test (rank-sum test) were all less than 10^{-5} under a null hypothesis that the human-human and human-GPT distributions were the same.

Table 3 shows a snapshot of the automated evaluation metrics comparing human responses to those from GPT on the moral prompt. Table 11 in the Appendix has details on the other prompts. All pairs of responses were compared for a given story which were then combined to create a single distribution of pairwise similarities between annotations. The human-human column indicates the median of the distribution comparing only human to human responses, the human-GPT column indicates the median in only comparing human responses to GPT responses, and the GPT-GPT column compares GPT responses to other GPT responses for replicability. We utilized the same number of GPT outputs as human annotations for the same story.

Overall, we find that the median lexical similarity between human responses is 0 (RougeL) and that the semantic variation between humans (first column) is often significantly higher (i.e. exhibits lower similarity) than that between GPT and human responses (second column) across all metrics. This confirms two suppositions stated above: a) that human answers on this task tend not to utilize overlapping vocabulary and b) that GPT's responses are not significantly different, semantically, compared to the human responses. We also note that GPT exhibits very high similarity to itself on multiple runs for the same text (expected given a temperature of zero), though these are not identical.

While it may be surprising that human-human similarity was lower than human-GPT, for some stories, human responses were quite different from

	Agreement (%) 1	2	3	κ	GPT Maj. %
Most Applicable					
Moral	13.9	59.0	27.1	0.01	68.06*
Moral+	14.6	65.3	20.1	0	60.42*
Moral–	16.7	66.0	17.4	0.03	52.78*
Topic	9.7	61.8	28.5	0.09	67.36*
Least Applicable					
Moral	16.7	58.3	25.0	0.12	11.11*
Moral+	20.8	63.2	16.0	0.04	11.81*
Moral–	23.6	63.2	13.2	0	15.97^
Topic	13.2	61.8	25.0	0.13	7.64*

Table 4: Inter-annotator agreement and GPT selection rate among AMT workers. The first 3 columns show how often 1, 2, or 3 annotators agreed on an option. The fourth column shows Fleiss κ coefficients. The fifth column shows the observed rate at which GPT was selected by the majority of AMT workers. * and ^ indicate a p-value of less than 0.01 and 0.05 respectively under a χ^2 goodness of fit test with a null hypothesis of random selection.

one another (as in Table 1) and, additionally, GPT had a stronger tendency to highlight more than one theme within its morals (e.g. "kindness" and "courage") which may have resulted in greater average semantic overlap across all answers.

3.3.3 Human Evaluation

While our automated measures suggest that GPT is decently consistent with human answers, they cannot answer the question posed above of whether these answers are appropriate or preferable. This evaluation is of the most consequence, as they could be similar to human responses, but somehow less desirable by human readers.

For the purposes of human validation, we employed a crowd-sourced voting paradigm using Amazon Mechanical Turk (AMT). In this way, we gained insights into human preferences that were independent from the individuals responsible for the labeling exercise.

For each of our open-response categories (Moral, Moral+, Moral–, and Topic), crowd workers were presented with a narrative document and three annotations produced for that document: one from GPT-4 and two that were randomly selected from among the human annotators. The crowd workers were then tasked with choosing the "most applicable" and "least applicable" options for each category given the story / story summary. In total, three annotations were collected for each passage. For more details on the survey and the selection of crowd workers, see Appendix A.2.2.

As seen in Table 4, for each category we achieved majority agreement in 85-90% of cases. However, inter-rater agreement as measured by Fleiss's Kappa was extremely low because while two annotators may have chosen a human moral they may have chosen different ones. We observe that in a majority of all cases the GPT moral was selected as the majority vote (GPT Maj. %), which performs well above a random baseline as indicated by a χ^2 goodness of fit test. Table 10 in the Appendix shows that this preference for the most applicable answer was consistent across all genres. In no case did crowdworkers preferentially choose GPT answers as "least applicable."

Although workers did not give reasons for their rankings, qualitatively, workers appeared to favor more explicit morals. As a general observation, GPT emphasized causes/effects in its morals ("The pursuit of vengeance can lead to complex alliances and conflicts"), which might have been an element workers preferred. By contrast, human morals were mostly imperative ("Stay true to your convictions") or maxim-like ("There are always disagreements among family members").

Overall, we find that GPT's answers to story moral extraction exhibit meaningful similarity to human answers and demonstrate above average preference among random human judges.

4 Application

We now present a case study to illustrate the analytical affordances of the story morals framework. Specifically, we collect morals from the full collection of our fairy tales data source from the validation. It consists of 1,760 tales from 54 cultural groups across all six inhabited continents; a breakdown of the cultural representation is given in Appendix B.1. All stories are English translations of the originals. We then apply clustering techniques, based on BERTopic (Grootendorst, 2022) for topic modelling, to group them into clusters and compare our results to those obtained by applying the same method to the full stories directly. We show that our clusters form cohesive units, illustrating lessons and pieces of wisdom related by a common theme, and that these results are qualitatively different from those obtained from the clusters of the

	Full Sent. Moral	Full Text
0	challenges, ingenuity, overcome	daughter, king, princess
1	love, obstacles, true	boy, farmer, old
2	cunning, deceit, trickery	hen, woman, little
3	greed, downfall, fortune	father, son, said
4	courage, facing, face	asgard, gods, king
5	acts, kindness, unexpected	shark, water, wurrunnah
6	kindness, compassion, unexpected	mother, said, boy
7	redemption, possible, resilience	pony, daughter, mother
8	justice, prevail, ultimately	good, man, day
9	persistence, success, innovation	shepherd, dervish, dog

Table 5: Top 10 largest clusters of the folktales using the embeddings of the full sentence morals compared to the full stories. The given words are the top 3 most representative words for each cluster as measured by c-TF-IDF. Only words larger than 3 letters are included.

Figure 1: 2D representation of the cluster centroids for the full-sentence morals when reduced using UMAP. Each circle corresponds to a cluster. Only a few of the largest clusters in each island are colored for readability. The circle sizes are related to the sizes of the clusters.

full stories.

4.1 Clustering Methods

Morals from GPT were collected using the same methodology detailed above and were post-processed to remove prefixes like "The moral is" (for full-sentence morals) and suffixes like "is good/bad behavior" for positive/negative morals. Morals were then embedded using the nli-mpnet-base-v2 model from SentenceTransformers and clustered using hierarchical clustering with a distance threshold and complete linkage. Clustering hyperparameters were chosen through joint optimization of the Silhouette score (Rousseeuw, 1987) and the Caliński-Harabasz index (Caliński and Harabasz, 1974). Details can be found in the Appendix B.2. For the full-text baseline, the same method was used except passing the full stories to the embedding model instead of the morals. For the text pre-processing there, spaCy (Honnibal et al., 2020) was used to replace all names and locations in the stories with generic ones ("Jack" for names and "The Place" for locations) to help reduce the presence of cultural cues. Final results yielded 60, 65, 129, and 99 clusters for the positive, negative, full sentence morals, and baseline, respectively.

Cluster labeling and visualization was done using the BERTopic library. In particular, cluster labels consist of the top 3 most representative words in each cluster as identified by c-TF-IDF (Grootendorst, 2022).

4.2 Results

Table 5 shows a comparison of the 10 largest clusters for the full sentence morals and the full text embedding model when compared using their most representative words. There, we see significant qualitative differences between the two groupings. The morals all center around typical virtues and vices ("courage", "kindness", "persistence"; "deceit", "greed") or higher-level ideals like "love", "redemption", and "justice". These concepts illustrate the more general values underpinning the stories' central lessons as well as the distinctive coherence of the various conceptual clusters. Table 6 shows the correspondence between the cluster keywords and the underlying moral sentences,

Cluster	Top Words	Sample Moral
Full Sentence Morals		
0	challenges, ingenuity, overcome	Ingenuity and courage can overcome even the most daunting challenges
1	love, obstacles, true	Love and perseverance can overcome even the most daunting obstacles
50	violence, war, victims	War can tragically intertwine the fates of individuals, revealing profound connections and lost opportunities for reconciliation among enemies
52	leadership, influence, guidance	True wisdom and effective leadership require prioritizing the public good over personal self-interest and recognizing the limitations of ones own perspective
71	youth, creativity, light	Facing and overcoming troubles in youth can lead to a happier and more fulfilling future
Positive Morals		_____ is good behavior
0	adversity, perseverance, face	Perseverance in the face of adversity
7	heeding, warnings, advice	Heeding warnings and respecting others' advice
8	forgiveness, redemption, repentance	Seeking forgiveness and making amends
17	nature, respecting, natural	Respecting nature
26	justice, seeking, prevails	Seeking justice
41	dead, death, inevitability	Accepting the inevitability of death
Negative Morals		_____ is bad behavior
0	deception, revenge, jealousy	Deception
2	pride, selfishness, ingratitude	Pride
6	stealing, coercion, abducting	Stealing
10	warnings, ignoring, advice	Ignoring warnings
22	overconfidence, boasting, unfamiliar	Overconfidence
32	nature, balance, disrespecting	Disrespecting nature

Table 6: Examples of moral clusters including the top 3 most representative words per cluster (as measured by c-TF-IDF), and a sample moral from each cluster. For the full list of clusters, see Tables 14 - 16 in the Appendix.

highlighting the strong association between the keywords and the underlying morals.

The full-text clusters, by contrast, are more associated with surface-level attributes such as character types ("daughter", "king", "mother", "shepherd", "dog"). While these are an important dimension of storytelling, for folktales in particular (Propp, 1968), Broadwell et al. (2017) have shown character roles are often interchangeable to support common lessons or values (while kings and daughters are associated with distinct behavior, witches and ogres can be swapped with little consequence).

Interestingly, clusters also emerge that are characterized by "asgard" (a dwelling of the gods in Nordic mythology) and "wurrunnah" (a hero in Aboriginal Australian mythology), highlighting how the full-text approach distinguishes features that are more distinctive of different cultures but overlook the commonalities between them.

One challenge with this method is that because morals often contain more than one theme ("Ingenuity and courage can overcome even the most daunting challenges"), clusters can become numerous and have decent overlap making them harder to interpret (Fig. 1). The positive and negative morals help address these concerns at the cost of restricting the morals to general keywords, though they too illustrate significant intersections (see Figs. 3a, 3b in the Appendix). Overlapping clusters are, however, often related to a common theme. Although not shown in Fig. 1, the clusters contained within the largest circle ("challenges"; red) are related to "courage", "persistance", and "adaptation", and those within the blue cluster ("love") are related to

"kindness", "selfless", and "devotion."

Even in spite of overlap, Table 6 illustrates the breadth of morals generated, including those centered on war (index 50), leadership (index 52), and youth (71). Results in the Appendix (Table 16) further show these morals cover diverse subjects such as misunderstandings (20), revenge (26), the importance of heritage (38), the supernatural (48), and even humor (100). The positive and negative morals further crystallize the sentiment behind the morals. There, it is clear that "heeding warnings" (7) and "accepting death" (41) are seen as positive attributes, and "pride" (2) is a negative one. While many of the positive and negative morals have the same but opposite meanings, they also have distinctive values among them, like forgiveness (8) and justice (26) versus stealing (6) and overconfidence (22).

The methods and framework introduced here offer significant potential to shed light on regional differences in cultural values or how cultures promote values differently through storytelling. While beyond the scope of this work, we observed statistically significant differences ($p < 0.01$) in the frequency of certain continents or cultural regions represented within specific moral clusters. For example, North American folktales had (full sentence) morals about contentment and appreciation at a higher frequency than what would be expected by chance, and morals about cleverness had a higher rate among stories from Africa.

While these findings may only apply to our dataset, which may not be representative of the stories from each given region or culture, our framework does provides a general method for exploring these questions in more depth.

5 Conclusion

In this paper we endeavored to introduce and validate a new narrative schema framework focusing on the concept of story morals. We showed that this is a well-understood concept by both human annotators and at least one candidate LLM (GPT-4), which demonstrates the capacity to produce high-value annotations in this area. We also showed that when applied to a diverse cultural dataset of folktales, coherent clusters of values and lessons emerged that cut across different cultures. Some were more strongly associated with certain regions, indicating a potential cultural preference for moral schemas within storytelling (Wu et al., 2023).

Our framework also provides the ability to explore messaging in other more contemporary genres. While we have focused our case study on the genre of folktales, our validation suggests that story moral labeling can also be successful on contemporary long form genres like novels, visual narratives such as movies and TV, personal stories from social media, and news media. While we expect each of these domains to pose specific analytical challenges, the story morals framework can be applied to these different domains and potentially surface, for example, core lessons circulated in the news media around key social issues such as climate change or the future of democracy (Zhou et al., 2024).

The higher level organization of moral clusters seen in our clustering exercise also suggests potential connections to other values-driven frameworks such as Moral Foundations Theory (Graham et al., 2013) or Schwartz's Theory of Basic Human Values (Schwartz and Bilsky, 1987; Schwartz, 2012). The value of our method lies in its pluralism, i.e. that it is open to numerous different lessons as opposed to beginning with a specified taxonomy which may or may not be sufficiently representative of cultural behavior. While the clusters we have uncovered are more fine-grained and detailed than the five basic concepts of MFT, for example, it is likely that they could fit inside that classification or even express new themes outside that scope. Further research will be needed to explore the relationship between our work and other frameworks.

Finally, the story morals framework can play an important role in the greater workflow of large-scale narrative understanding (Zhu et al., 2023; Piper, 2023). Story morals represent a foundational narrative schema that shapes how stories are told and what issues are focalized in a given story. In this, they can support the larger narratological project of surfacing both content-level and discourse-level features. Story morals provide a valuable way of thinking about narrative schema labeling to understand broader storytelling trends and behavior.

Limitations

In our study, only GPT-4 was tested in our validation, and no other LLMs were used due to the high costs of running the AMT surveys (validating additional models would have involved comparing the human responses to each LLM response separately). This is an important limitation since

different LLMs may produce differently inflected lessons, leading to different conclusions about larger cultural behavior. We leave extensive model comparisons as future work.

Another central limitation of our task is the unknown cultural orientation of GPT-4 and other LLMs more broadly. We know that most LLMs have primarily been trained on English-language texts with a Western cultural lens, though it is uncertain whether this leads to intrinsically biased story morals. Because story morals are descriptive lessons rather than moral sentiments derived from the specific events of texts, it is possible that they may represent viewpoints that are more culturally independent. For example, we did not see human morals produced by Mandarin speakers more or less preferentially voted for by crowd workers than those produced by English speakers suggesting that the readers cultural background did not influence audience judgments concerning the suitability of morals. As Graham et al. (2013) suggest, there is likely a very strong degree of cross-cultural "preparedness" that underlies the preference for particular lessons to be derived from a given set of narrative events.

Nevertheless, future work will want to concentrate on cross-cultural comparisons of story moral annotations at both the human generation and human voting stages to provide a full accounting of the potential cultural sensitivity of human perspectives on deriving story morals.

For our case study, while our folktales dataset contained stories from many world cultures, many cultures ultimately had no representation within the dataset. In addition, of those present, the majority of the folktales were of European origin and it is highly plausible that some of the collections of stories may not have been representative of different groups or regions. As such, all cultural/regional claims made in our paper are not meant to be general and are subject to these limitations. All folktales were also English translations and therefore may not reflect the cultural nuances of the originals. While this is a fairly major limitation in terms of the cultural study of folktales, the availability of clean multi-lingual datasets remains a challenge. While we provide one instance of multilingual validation (Mandarin), validating our framework on texts from additional languages is a key area for future work, though may require significant efforts in terms of time and resources.

Ethics Statement

The study of morals and narrative messaging is closely linked to culture, and has the potential to help us better understand cultural values and biases. While our case study of folk tales is purely meant for illustrative purposes, there may be those who seek to use similar cultural analysis techniques to find or advance stereotypes, and justify discrimination. When performing cultural analyses, we therefore advocate for the responsible use of AI and LLM technologies, and for the thorough assessment of datasets to ensure they accurately and adequately represent the different groups under study. We contribute these methods in the hope of furthering language technologies towards being more socially just and equitable, and towards finding common values and perspectives among peoples.

Acknowledgements

We would like to thank the reviewers for their constructive feedback and suggestions. This research was supported by the Natural Sciences and Engineering Research Council and the Social Sciences and Humanities Research Council of Canada.

References

Nandini Anantharama, Simon Angus, and Lachlan O'Neill. 2022. Canarex: Contextually aware narrative extraction for semantically rich text-as-data applications. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3551–3564.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

David Bamman and Noah A Smith. 2013. New alignment methods for discriminative book summarization. *arXiv preprint arXiv:1305.1319*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and*

future, pages 15–21. Association for Computational Linguistics.

Wayne C Booth. 1998. Why ethical criticism can never be simple. *Style*, pages 351–364.

William F Brewer and Edward H Lichtenstein. 1980. Event schemas, story schemas, and story grammars. *Center for the Study of Reading Technical Report; no. 197*.

Peter Broadwell, David Mimno, and Timothy Tangherlini. 2017. The tell-tale hat: Surfacing the uncertainty in folklore classification. *Journal of Cultural Analytics*, 2(1).

T. Caliński and J Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

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

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

Zeyu Dai and Ruihong Huang. 2021. A joint model for structure-based news genre classification with application to text summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3332–3342, Online. Association for Computational Linguistics.

Alan Dundes. 1962. From etic to emic units in the structural study of folktales. *The Journal of American Folklore*, 75(296):95–105.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732, Toronto, Canada. Association for Computational Linguistics.

Northrop Frye. 2020. *Anatomy of criticism: Four essays*, volume 69. Princeton University Press.

Liye Fu, Jonathan P. Chang, and Cristian Danescu-Niculescu-Mizil. 2019. Asking the right question: Inferring advice-seeking intentions from personal narratives. 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 528–541, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikaela Irene Fudolig, Thayer Alshaabi, Kathryn Cramer, Christopher M Danforth, and Peter Sheridan Dodds. 2023. A decomposition of book structure through ousiometric fluctuations in cumulative word-time. *Humanities and Social Sciences Communications*, 10(1):1–12.

Gérard Genette. 1992. *The architext: An introduction*, volume 31. Univ of California Press.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of *Advances in Experimental Social Psychology*, pages 55–130. Academic Press.

Marshall W Gregory. 2010. Redefining ethical criticism. the old vs. the new.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

FB Karsdorp and APJ van den Bosch. 2013. Identifying motifs in folktales using topic models.

Karin Kukkonen. 2014. Plot. *The living handbook of narratology*, 24.

Kaushil Kundalia, Yash Patel, and Manan Shah. 2020. Multi-label movie genre detection from a movie poster using knowledge transfer learning. *Augmented Human Research*, 5:1–9.

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

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep 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.

Stephanie Lukin, Kevin Bowden, Casey Barackman, and Marilyn Walker. 2016. PersonaBank: A corpus of personal narratives and their story intention graphs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1026–1033, Portorož, Slovenia. European Language Resources Association (ELRA).

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative question answering with cutting-edge open-domain QA techniques: A comprehensive study. *Transactions of the Association for Computational Linguistics*, 9:1032–1046.

Martha Craven Nussbaum. 1998. Exactly and responsibly: A defense of ethical criticism. *Philosophy and Literature*, 22(2):343–365.

Jessica Ouyang, Serina Chang, and Kathy McKeown. 2017. Crowd-sourced iterative annotation for narrative summarization corpora. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Jessica Ouyang and Kathleen McKeown. 2015. Modeling reportable events as turning points in narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2149–2158.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Andrew Piper. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the Big Picture Workshop*, pages 28–39.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas press.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967*.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science*, 5(1):1–12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rezvaneh Rezapour, Priscilla Ferronato, and Jana Diesner. 2019. How do moral values differ in tweets on social movements? In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, pages 347–351.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2023. Towards few-shot identification of morality frames using in-context learning. *arXiv preprint arXiv:2302.02029*.

Robert L Russell and Paul Van Den Broek. 1992. Changing narrative schemas in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, 29(3):344.

Shalom Schwartz. 2012. Toward refining the theory of basic human values. *Methods, theories, and empirical applications in the social sciences*, pages 39–46.

Shalom H Schwartz and Wolfgang Bilsky. 1987. Toward a universal psychological structure of human values. *Journal of personality and social psychology*, 53(3):550.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Stith Thompson. 1955. *Motif-Index of Folk-Literature, Volume 4: A Classification of Narrative Elements in Folk Tales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*, volume 4. Indiana University Press.

Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated event annotation in literary texts. In *CHR*, pages 333–345.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Matthew Wilkens. 2016. Genre, computation, and the varieties of twentieth-century us fiction. *Journal of Cultural Analytics*, 2(2).

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.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Xinru Yan, Aakanksha Naik, Yohan Jo, and Carolyn Rose. 2019. Using functional schemas to understand social media narratives. In *Proceedings of the Second*

Workshop on Storytelling, pages 22–33, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. Narrasum: a large-scale dataset for abstractive narrative summarization. *arXiv preprint arXiv:2212.01476*.

Haiqi Zhou, David Hobson, Derek Ruths, and Andrew Piper. 2024. Large scale narrative messaging around climate change: A cross-cultural comparison. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 143–155, Bangkok, Thailand. Association for Computational Linguistics.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are NLP models good at tracing thoughts: An overview of narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.

A Appendix: Validation

A.1 Validation Dataset

A.1.1 Dataset Breakdown

The validation dataset consisted of passages across the following genres: folktales from Eastern (Indian and Japanese) and Western (French and German) cultures, book and movie/TV summaries from North America, personal stories from Reddit, and political news from CNN, Al-Jazeera and four sources of Chinese-language news (Mandarin) were used. Mandarin news sources were selected as a preliminary step towards validating GPT in other languages. As noted in the main text, the mean length of documents is 839 words with a minimum of 200 and a maximum of 2,200.

Genre	Description	No.
Books	Book summaries (Bamman and Smith, 2013)	16
Folktales	Eastern and Western folktales (Folk Tales dataset)	32
Movies/TV	Movie/TV episode summaries (Zhao et al., 2022)	16
News	English: CNN & Al-Jazeera (Global News dataset); Chinese (Mandarin): 2 mainland, 2 offshore (Hong Kong, Taiwan) (Zhou et al., 2024)	64
Reddit	Non-fiction human stories from r/AskReddit (Ouyang and McKeown, 2015)	16
Total		144

Table 7: Breakdown of the validation data by genre and source.

A.1.2 Additional Dataset Information

The dataset by Bamman and Smith (2013) is covered under the Creative Commons Attribution-ShareAlike licence (CC BY-SA 3.0 US), and the news datasets are public domain (CC0). Licenses for the data from Zhao et al. (2022) and Ouyang and McKeown (2015) are unknown but the data was obtained from the respective authors under conditions of academic use. The terms of use for folktales dataset can be found on the Fairy Talez website. As all data used here was purely for research purposes, it did not violate the intended use from any of the sources.

While the Reddit dataset from Ouyang and McKeown (2015) contained content with adult themes or that could be considered offensive, human annotators and AMT workers were warned of such potential content beforehand.

A.2 Further Documentation involving Human Annotations

Codebooks and the HTML template for our Amazon Mechanical Turk survey are available on our GitHub repository.

A.2.1 Moral Annotations

As mentioned in the main paper, all annotators were undergraduate students. Participants were paid at an hourly wage above minimum wage, and were all aware their answers would be used for research purposes.

For English passages, 6 human annotators were used, and for the Mandarin passages, 4 annotators fluent in Mandarin were used. All annotators were raised in Western countries.

Annotators were provided with a codebook of category definitions and examples, and underwent at least one round of practice annotations to affirm consistency of interpretations to the definitions. All human responses to the questions were open-response and made independently of each other and from GPT-4. We thus record between 4 and 6 responses for each question and for each document.

A.2.2 MTurk Voting Exercise

In total, three responses were collected for each passage. Crowdworkers were given no explicit instructions about what constituted a good or bad option and were given the freedom to select based on their own preferences, so as to avoid any selection bias.

To ensure quality responses, we required workers to have a lifetime success rate of more than 95%, and workers had to correctly answer a passage comprehension question to be considered.

To partially address new concerns among researchers of crowd-workers using ChatGPT to an-

	Agree with maj.	Any agree-ment	No agree-ment	Avg. pop. vote
Protagonist				
Book	95.31	95.31	4.69	94.79
Folktale	81.25	95.31	4.69	78.65
Movies-TV	84.38	85.94	14.06	84.38
News	49.22	61.33	38.67	62.89
Reddit	87.50	93.75	6.25	91.67
Antagonist				
Book	56.25	100	0	54.17
Folktale	80.47	99.22	0.78	72.40
Movies-TV	84.38	98.44	1.56	73.96
News	71.88	96.88	3.12	69.99
Reddit	56.25	75.00	25.00	74.48

Table 8: Percent agreement of GPT responses with human responses for protagonists and antagonists. Values represent the average agreement over all GPT responses collected (equal to the number of human annotators). Agree with maj. is the percentage of the time GPT agreed with the majority vote among humans. Avg. pop. vote is the average human popular vote as a percentage.

	human-human	human-GPT	GPT-GPT
Valence			
Book	0.77	0.77	0.11
Folktale	0.75	0.86	0.16
Movies-TV	0.79	0.83	0.21
News	0.68	0.66	0.08
Reddit	0.77	0.79	0
Protagonist Type			
Book	57.92	54.40	91.04
Folktale	60.10	53.83	94.06
Movies-TV	61.04	50.43	93.54
News	44.05	42.46	89.06
Reddit	36.46	38.63	92.08

Table 9: Average standard deviations in valence responses and Jaccard index (out of 100) in protagonist type between the different distributions of responses. The human-human column compares all pairs of responses (to the same passage) among the human annotators, the human-GPT group compares all pairs of responses between human and GPT responses, and the GPT-GPT column compares all responses between GPT responses. The number of GPT responses was always chosen to be equal to the number of human annotators.

swer the questions, all passages were provided as images. For passages in Mandarin, English translations of the text were provided to the workers.

All workers were compensated at a rate of $2/HIT (USD). The estimated time per HIT was 8 minutes thus translating to an hourly wage of $15/hour. Participants were aware they were participating for academic purposes. No geographic restrictions were placed on the AMT workers, and so no cultural representation data is available for the workers who participated in our survey.

A.3 Validation Results for Other Narrative Categories in our Prompting Workflow

Automated evaluations for the categories of protagonists, antagonists, protagonist type and valence showed positive results.

For identifying protagonists and antagonists (Table 8), GPT agreed with the majority opinion of human annotators 70.8% and 71.5% of the time, respectively, compared to the average majority agreements among humans of 75.5% and 69.7%.

For protagonist type (Table 9), since multi-label selections were permissible (“hero”, “villain”, “victim”), responses were compared using the Jaccard index. When only comparing pairwise human responses, the average Jaccard index was 0.51, compared to 0.47 when comparing human to GPT responses. While this difference was significant under a Mann-Whitney U-test ($p = 0.004$), these values were still very comparable.

For valence (Table 9), the average standard deviation between human-only responses was 0.73 across all passages, compared to 0.75 when introducing GPT responses, thus showing good overall compatibility between the responses.

	Book	Folktale	Movies-TV	News	Reddit
Most applicable					
Moral	62.50	78.12	56.25	73.44	43.75
Moral+	56.25	62.50	62.50	57.81	68.75
Moral–	56.25	50.00	62.50	51.56	50.00
Topic	75.00	65.62	37.50	73.44	68.75
Least applicable					
Moral	12.50	6.25	25.00	7.81	18.75
Moral+	6.25	12.50	12.50	12.50	12.50
Moral–	18.75	12.50	12.50	17.19	18.75
Topic	6.25	3.12	25.00	7.81	0

Table 10: Percent of passages by genre where the GPT response was selected by a majority of AMT workers.

	human-human	human-GPT	GPT-GPT	U-test
Moral				
RougeL	0	7.41	52.94	$p < 10^{-5}$
BERTScore	56.17	57.94	81.81	$p < 10^{-5}$
GloVe	47.36	56.63	88.34	$p < 10^{-5}$
STSb-MPNet	25.14	37.14	83.63	$p < 10^{-5}$
NLI-MPNet	31.59	44.07	87.49	$p < 10^{-5}$
Moral+				
RougeL	0	0	66.67	$p < 10^{-4}$
BERTScore	56.67	55.72	84.30	$p < 10^{-3}$
GloVe	25.45	34.24	84.74	$p < 10^{-4}$
STSb-MPNet	27.83	27.09	80.40	$p = 0.09$
NLI-MPNet	39.80	36.67	83.07	$p < 10^{-4}$
Moral–				
RougeL	0	0	66.67	$p = 0.06$
BERTScore	56.64	52.92	86.69	$p < 10^{-4}$
GloVe	20.67	23.65	81.33	$p < 10^{-4}$
STSb-MPNet	22.57	20.81	83.20	$p < 10^{-4}$
NLI-MPNet	33.43	28.98	84.82	$p < 10^{-4}$
Topic				
RougeL	0	0	100	$p = 0.32$
BERTScore	52.31	53.02	100	$p = 0.21$
GloVe	34.47	35.26	100	$p = 0.28$
STSb-MPNet	24.63	26.11	100	$p = 0.46$
NLI-MPNet	33.45	33.92	100	$p = 0.34$

Table 11: Median similarity (out of 100) between the different groups of annotators in the validation dataset. The human-human column is compares all pairs of responses (to the same passage) among the human annotators, the human-GPT group compares all pairs of responses between human and GPT responses, and the GPT-GPT column compares all responses between GPT responses. The p-values are calculated using a Mann-Whitney U-test (rank-sum test) with a null hypothesis that the human-human and human-GPT distributions are the same.

Moral 1	Moral 2	RougeL	BERT-Score	GloVe	STSb-MPNet	NLI-MPNet
Hard work will pay off in the end	Hard work pays off	66.67	85.13	96.00	88.48	86.76
The truth is worth fighting for	Seeking the truth is important	54.55	70.11	69.76	76.59	75.43
Investigating your past can unearth old secrets	Confronting one's past, no matter how difficult, can lead to truth, justice, and the opportunity for a new beginning	14.81	65.31	54.69	45.49	53.25
Loyalty to your loved ones will be rewarded	Justice prevails against tyranny	0	51.81	24.18	21.07	20.73

Table 12: Sample morals from the validation dataset and their pairwise similarity with respect to the different automated metrics.

B Appendix: Case Study

B.1 Folk Tales Dataset Details

Folk tales come from the Fairy Talez collection of classic and new fairy tales.[3] Specifically, we used a pre-scraped and pre-cleaned version from Kaggle (see Table 7).

In selecting the stories for the folk tales dataset, to align with our validation, only stories with lengths between 200 and 2200 words were kept. All duplicate stories were removed, and following Wu et al. (2023), all documents from the same continent whose titles had more than 80% overlap relative to the Jaccard index were also removed to avoid potential duplicates. After removals, the dataset consisted of 1760 stories with a mean length of 1033 words; 960 were from Europe (54.5%), 338 from North America (19.2%), 315 from Asia (17.9%), 66 from Oceania (3.8%), 57 from Africa (3.2%), and 24 from South America (1.4%). In addition to these continents, the nations were split into cultural regions, shown in Table 13.

B.2 Moral Clustering Hyperparameter Details

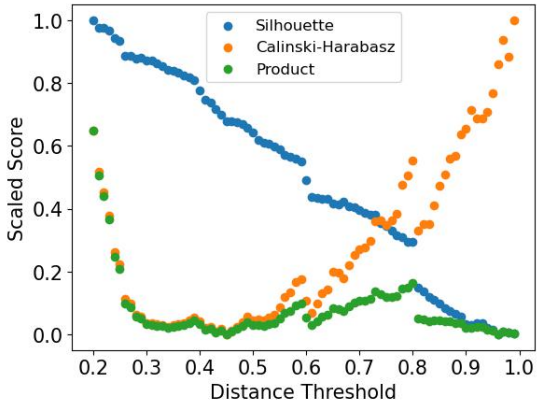
To determine the optimal clusterings, the Silhouette coefficient (Rousseeuw, 1987) and the Caliński-Harabasz Index (CHI; Caliński and Harabasz, 1974) were applied. These specific criteria were selected since they both empirically exhibited non-monotonic behavior over the clustering hyperparameters, compared to other criteria which only either favoured a single cluster or the largest number of clusters possible. The Silhouette and CHI measures were jointly optimized by scaling the minimum and maximum values to 0 and 1 and maximizing their product. Agglomerative, K-means, and HDBSCAN clustering algorithms were all investigated, but the agglomerative approach based on a distance threshold (with complete linkage) gave the most distinctive optimal points, and therefore account for all the results seen here. The plots for the hyperparameter tuning can be found in Figure 2. The optimal values for the distance threshold were $t = 0.8, 0.83, 0.62$, and 0.83 (with cluster sizes of 60, 65, 129, and 99 clusters) for the positive, negative, full sentence morals, and baseline full-text, respectively.

All algorithms, the Silhouette score, and CHI were implemented using Scikit-Learn. Agglomerative clustering was implemented using SciPy.

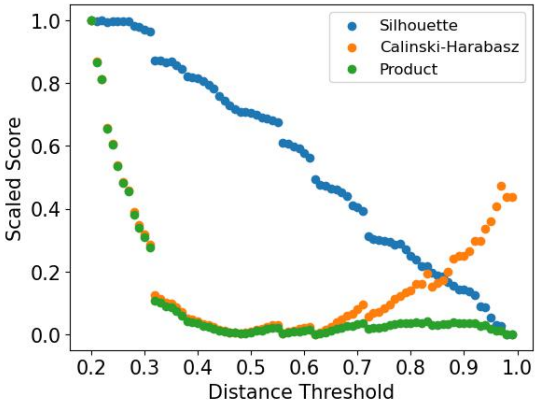
[3] https://fairytalez.com/

Cultural Region	Nation	Count	Cultural Region	Nation	Count
North Europe	german	191	North America	north_american_native	330
	english	95		canadian_native	8
	nordic	64	East Asia	japanese	63
	scandinavian	39		chinese	60
	scottish	37		philippine	57
	danish	27		korean	10
	celtic	19	South Asia	indian	82
	swedish	19		pakistani	4
	welsh	12	Middle East	armenian	17
	irish	11	(Asia)	arabic	15
	dutch	9		turkish	7
	belgian	8	Oceania	australian_ethnic	45
	icelandic	6		maori	9
	norwegian	5		new_zealand_native	7
	finnish	2		hawaiian	5
East Europe	greek	52	Africa	nigerian	36
	russian	45		south_african	13
	czechoslovak	27		tanzanian	7
	slavic	20		zimbabwe	1
	ukrainian	18	South America	brazilian	24
	polish	17			
	albanian	14			
	serbian	10			
	bulgarian	7			
	romanian	6			
	hungarian	5			
	bukovinian	3			
	croatian	3			
	estonian	1			
Romance	french	74			
(Europe)	italian	61			
	portuguese	45			
	spanish	7			
	cataloanian	1			
Total					1760

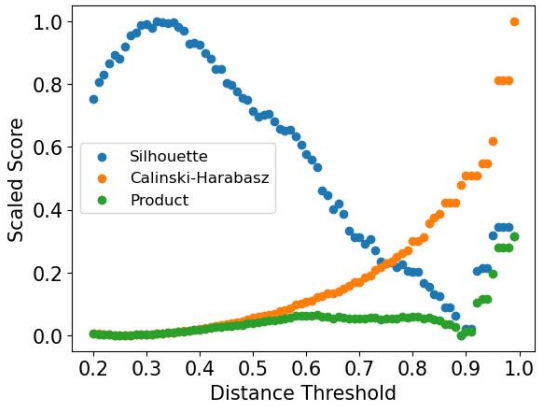
Table 13: Cultural breakdown of the folktales dataset



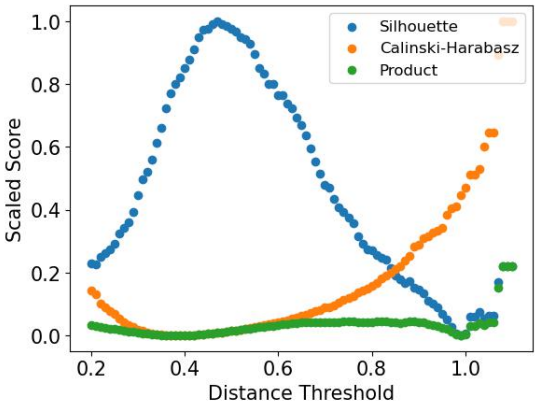
(a) Positive Moral



(b) Negative Moral

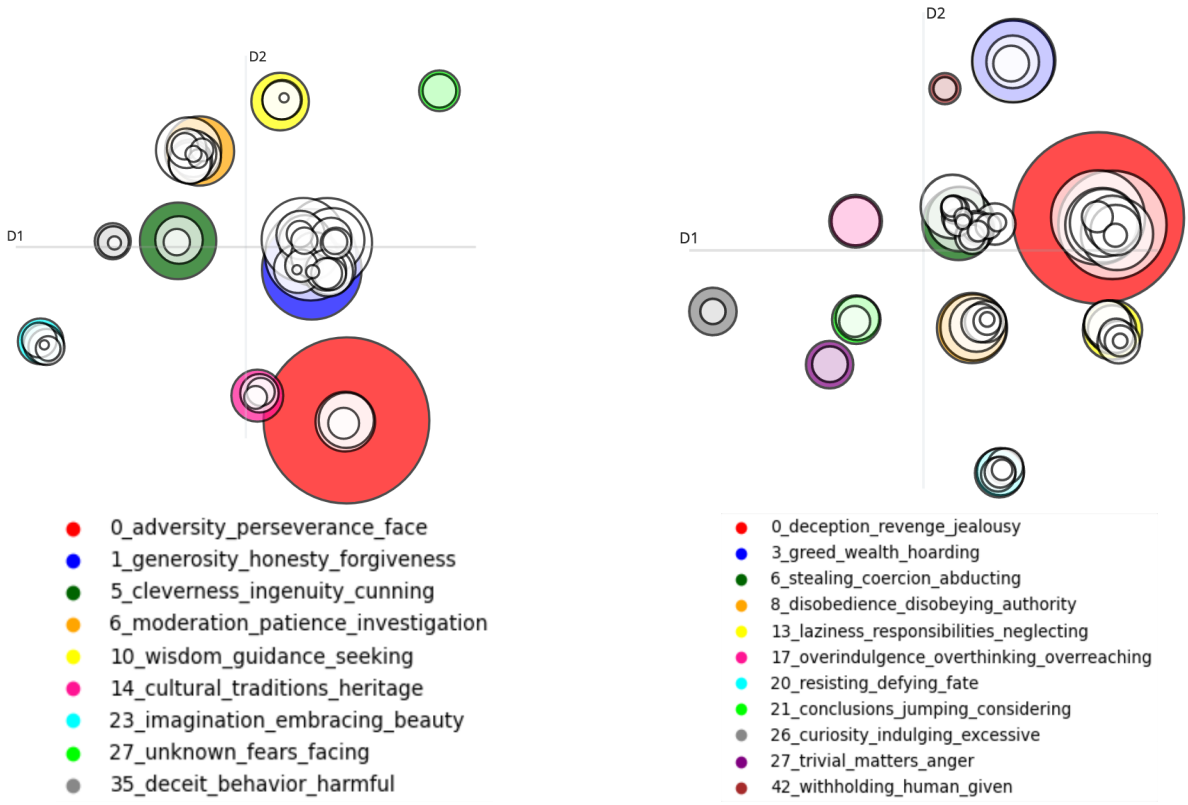


(c) Full-Sentence Moral



(d) Full-Text Clustering

Figure 2: Choosing the optimal distance threshold for agglomerative clustering by jointly maximizing the Silhouette score and Caliński-Harabasz indices. Higher Silhouette scores and Caliński-Harabasz indices correspond to better clustering. Note that since both measures have different scales, they have been re-scaled in the plot so that their minimum and maximum values correspond to 0 and 1. While distance thresholds less than 0.3 are high for both Positive and Negative Morals, they lead to more than 500 clusters are are thus too fine-grain. Similarly, while the product for the full-sentence morals grows for values above 0.9, the number of clusters becomes too coarse-grain (roughly 6 clusters with one giant cluster) to be insightful. Barring these restrictions, the optimal values are at $t = 0.8, 0.83, 0.62,$ and 0.83 respectively for positive, negative, and full-sentence morals, and full-text sentences.



(a) Positive morals

(b) Negative morals

Figure 3: 2D representation of the cluster centroids for the positive and negative morals when reduced using UMAP. Each circle corresponds to a cluster. Only the largest cluster in each island is colored for better readability. The size of the clusters are related to the sizes of the clusters. The legend names give the cluster index followed by the top 3 most representative words for each cluster as determined by c-TF-IDF.

Cluster	Top Words	Sample Moral (____ is good behavior)	Size
0	adversity, perseverance, face	Perseverance in the face of adversity	316
1	generosity, honesty, forgiveness	Generosity	114
2	integrity, honesty, treating	Honesty and integrity	96
3	humility, compassion, understanding	Humility	92
4	gratitude, kindness, showing	Kindness and gratitude	80
5	cleverness, ingenuity, cunning	Cleverness and ingenuity	67
6	moderation, patience, investigation	Moderation	55
7	heeding, warnings, advice	Heeding warnings and respecting others' advice	49
8	forgiveness, redemption, repentance	Seeking forgiveness and making amends	47
9	teamwork, cooperation, collaboration	Teamwork	41
10	wisdom, guidance, seeking	Seeking spiritual understanding and guidance	38

Continued on next page

Cluster	Top Words	Sample Moral (____ is good behavior)	Size
11	promises, keeping, honoring	Keeping promises	35
12	vigilance, work, hard	Vigilance	35
13	helping, help, vulnerable	Helping others	32
14	cultural, traditions, heritage	Respecting cultural traditions	31
15	love, pursuing, relationships	Perseverance in love	31
16	loyalty, country, true	Loyalty	28
17	nature, respecting, natural	Respecting nature	27
18	contentment, acceptance, accepting	Contentment	27
19	caution, exercising, unknown	Caution	26
20	solving, problem, innovation	Ingenious problem-solving	26
21	self, sacrifice, selflessness	Self-sacrifice	25
22	prudence, responsibly, using	Prudence	25
23	imagination, embracing, beauty	Embracing imagination	24
24	decision, making, quick	Caution in decision-making	22
25	thinking, critical, outside	Critical thinking	20
26	justice, seeking, prevails	Seeking justice	19
27	unknown, fears, facing	Respecting the unknown	19
28	material, valuing, possessions	Valuing relationships over material possessions	19
29	knowledge, teaching, luck	Sharing knowledge	18
30	communication, clear, boundaries	Clear communication	18
31	change, embracing, adaptability	Embracing change	18
32	ones, loved, memories	Protecting loved ones	17
33	curiosity, tempered, exploration	Curiosity	17
34	adapting, change, hardships	Adapting to change	15
35	deceit, behavior, harmful	Exposing deceit is risky behavior	15
36	nurturing, moments, fleeting	Appreciating fleeting moments	14
37	skepticism, supernatural, mysteries	Vigilance and skepticism	13
38	dreams, heart, pursuing	Pursuing one's dreams	12
39	peace, autonomy, innocent	Perseverance in seeking peace	12
40	greed, resisting, manipulation	Greed	11
41	dead, death, inevitability	Accepting the inevitability of death	11
42	standing, oppression, oneself	Standing up for oneself	11
43	filial, piety, divine	Filial piety	11
44	family, marriage, bonds	Loyalty to family	9
45	resolving, conflict, peacefully	Resolving conflicts peacefully	8
46	present, content, appreciating	Being content with what you have	8
47	wit, using, humor	Using wit to overcome challenges	8
48	creatures, animals, lives	Kindness towards all creatures	8
49	limitations, accepting, unpredictability	Accepting one's limitations	7
50	loss, grief, coping	Gaining perspective on personal suffering	7
51	simplicity, safety, practicality	Simplicity	6
52	elders, elderly, caring	Seeking wisdom from elders	6
53	underestimating, unwise, behavior	Underestimating others is unwise behavior	4
54	secrets, verifying, information	Keeping sacred secrets	3
55	waste, pursuits, makes	Haste makes waste	2
56	silence, response, provocation	Silence	2

Continued on next page

Cluster	Top Words	Sample Moral (____ is good behavior)	Size
57	change, accepting,	Accepting change	1
58	planting, future,	Planting for the future	1
59	leaving, authorities, judgment	Leaving judgment to higher authorities	1

Table 14: All clusters for the positive morals

Cluster	Top Words	Sample Moral (____ is bad behavior)	Size
0	deception, revenge, jealousy	Deception	275
1	deceit, death, truth	Deceit	110
2	pride, selfishness, ingratitude	Pride	83
3	greed, wealth, hoarding	Greed	64
4	greed, squandering, wealth	Greed and deceit	63
5	underestimating, value, forgetting	Underestimating others	52
6	stealing, coercion, abducting	Stealing	49
7	promises, breaking, gullibility	Breaking promises	49
8	disobedience, disobeying, authority	Disobedience	46
9	cowardice, responsibility, true	Cowardice	43
10	warnings, ignoring, advice	Ignoring warnings	41
11	mistreating, bullying, oppressing	Mistreating the vulnerable	39
12	forcing, imposing, intruding	Forcing one's will on others	38
13	laziness, responsibilities, neglecting	Laziness	33
14	mocking, mockery, deceiving	Mocking or underestimating others	29
15	disregarding, signs, ignoring	Ignoring supernatural signs	27
16	gain, personal, superficial	Deception for personal gain	26
17	overindulgence, overthinking, overreaching	Excessive talking	26
18	trust, betraying, secrets	Betraying trust	24
19	needs, indifference, small	Neglecting the needs of others	23
20	resisting, defying, fate	Resisting change	22
21	conclusions, jumping, considering	Jumping to conclusions	22
22	overconfidence, boasting, unfamiliar	Overconfidence	22
23	neglecting, communication, historical	Neglecting communication in relationships	22
24	imitation, supernatural, superstition	Envy and imitation without understanding	21
25	material, term, expense	Pursuing selfish desires at the cost of others' happiness	21
26	curiosity, indulging, excessive	Curiosity	21
27	trivial, matters, anger	Quarreling over trivial matters	21
28	recklessness, strangers, trusting	Recklessness	18
29	giving, easily, face	Giving up	17
30	traditions, sacred, cultural	Disregarding traditions	16
31	appearances, judging, showing	Judging by appearances	16
32	nature, balance, disrespecting	Disrespecting nature	14
33	harm, harming, justice	Harming others	13
34	fear, fears, facing	Succumbing to fear or evil influences	12
35	temptation, succumbing, seven	Succumbing to temptation	12
36	impatience, complaining, dissatisfaction	Impatience and rudeness	12
37	harshly, judging, disregard	Showing indiscriminate kindness to those with harmful intentions	12
38	witchcraft, unknown, wickedness	Dabbling in unknown forces	11
39	elderly, elders, desertion	Abandoning the elderly or weak	11
40	persecution, intolerance, punishment	Persecution of others for their beliefs	11
41	grief, discontentment, allowing	Neglecting the living while consumed by grief	11

Continued on next page

Cluster	Top Words	Sample Moral (____ is bad behavior)	Size
42	withholding, human, given	Withholding charity	9
43	danger, abandoning, away	Renouncing one's tribe under pressure	9
44	refusing, help, compromise	Refusing to help others	9
45	judgment, rash, hastily	Rash judgment	9
46	love, expressions, unrequited	Dismissing genuine expressions of love	9
47	dismissing, unconventional, contributions	Dismissing unconventional solutions	8
48	marriage, rejecting, dynamics	Forcing someone into marriage	7
49	rudeness, scolding, incest	Rudeness	7
50	dead, graves, disturbing	Disregarding the separation between the living and the dead	7
51	mourning, distress, incessant	Incessant mourning	6
52	judging, status, social	Judging people by their social status	6
53	solely, relying, physical	Relying solely on physical strength	6
54	obsession, clinging, obsessing	Obsession	6
55	vows, breaking, sabbath	Disobeying sacred vows	5
56	selling, permission, taking	Taking what is not yours without permission	5
57	miraculous, irreverence, enduring	Disbelief in the miraculous	4
58	trying, solve, results	Relying solely on others without first attempting to solve your own problems	4
59	impossible, setting, tasks	Setting impossible tasks	4
60	ostracizing, isolating, excluding	Ostracizing others	3
61	prioritizing, fun, solely	Focusing solely on physical appearance	3
62	blindly, following, adhering	Blindly following advice	2
63	treasure, speaking, hunt	Speaking during a treasure hunt	2
64	repetitive, lives, habitats	Confining animals to repetitive lives	2

Table 15: All clusters for the negative morals

Cluster	Top Words	Sample Moral	Size
0	challenges, ingenuity, overcome	Ingenuity and courage can overcome even the most daunting challenges	100
1	love, obstacles, true	Love and perseverance can overcome even the most daunting obstacles	74
2	cunning, deceit, trickery	Cunning and deceit may lead to temporary gain, but ultimately result in isolation and loss of trust	73
3	greed, downfall, fortune	Greed and deceit ultimately lead to one's downfall	59
4	courage, facing, face	Even in facing formidable challenges, courage and determination can lead to overcoming great threats, though sometimes with unintended consequences	59
5	acts, kindness, unexpected	Acts of kindness and compassion can lead to unexpected and rewarding outcomes, even in the face of greed and betrayal	55
6	kindness, compassion, unexpected	Kindness and perseverance in the face of adversity can lead to unexpected and rewarding outcomes	55
7	redemption, possible, resilience	Envy and malice can lead to unintended consequences, but resilience and goodness can ultimately lead to redemption and happiness	53
8	justice, prevail, ultimately	Courage, cleverness, and perseverance can overcome deceit and evil to restore justice and well-being	50
9	persistence, success, innovation	Persistence and resourcefulness can overcome obstacles and achieve one's goals	49
10	wit, triumph, strength	Wit and cleverness can triumph over brute strength and fear	40
11	rewarded, selfishness, cruelty	Kindness and humility are rewarded, while greed and selfishness lead to negative consequences	38
12	faith, new, enduring	Faith and perseverance can lead to the establishment and growth of new communities and beliefs, despite adversity and skepticism	36
13	eventually, justice, retribution	Deceit and wrongdoing will eventually be met with justice and retribution	32
14	trust, breaking, promise	Breaking a promise, especially one made to someone you love, can lead to irreversible loss and regret	30
15	tragic, desires, jealousy	The pursuit of selfish desires can lead to tragic consequences and irreversible loss	29
16	loyalty, adversity, face	Loyalty and principles can lead to personal hardship, but integrity and steadfastness in one's beliefs, even in the face of adversity, can ultimately lead to redemption and respect	28
17	advice, ignoring, warnings	Heed the advice of elders, as ignoring their wisdom can lead to dangerous or harmful situations	28
18	pride, humility, vanity	Pride and greed can lead to one's downfall, emphasizing the importance of humility and appreciating what one already has	25
19	reunite, support, family	Love and communal effort can overcome obstacles and reunite separated loved ones	23
20	misinterpretations, unintended, based	Fear and panic can often be based on misunderstandings or misinterpretations of natural events	21

Continued on next page

Cluster	Top Words	Sample Moral	Size
21	guilt, cheat, death	Attempting to cheat or defy the natural order, especially death, can lead to dire consequences	20
22	neglecting, irreversible, responsibilities	Neglecting responsibilities and failing to communicate effectively can have irreversible and far-reaching consequences	20
23	faith, acts, selfless	Hard work, kindness, and faith can bring unexpected blessings and transform lives in miraculous ways	18
24	severe, disrespecting, irrevocable	Deceit and betrayal ultimately lead to severe consequences and isolation	18
25	wit, situations, turn	Wit and cleverness can help one escape from dangerous or deceitful situations	17
26	vengeance, revenge, innocent	Acts of vengeance and hatred can lead to tragic consequences, affecting not only the perpetrators but also innocent lives and future generations	17
27	finding, contentment, happiness	True happiness and contentment come from appreciating what one has and finding joy in one's own circumstances	17
28	adventures, dreams, experiences	The pursuit of knowledge and adventure can lead to extraordinary experiences and discoveries that transcend the ordinary limits of time and space	16
29	luck, honesty, fortunes	Naivety and lack of wisdom can lead to poor decisions and loss, but maintaining a positive attitude and gratitude can make one feel fortunate despite material losses	16
30	conflicts, peace, misunderstandings	Through understanding, compromise, and a willingness to adapt, conflicts can be resolved and harmony achieved between individuals of vastly different backgrounds	16
31	fear, superstition, irrational	Fear and misunderstanding can lead to irrational decisions and tragic outcomes	16
32	protect, vigilance, prudence	Caution and vigilance can uncover deceit and protect one from harm	15
33	wish, careful, present	Be careful what you wish for, as the pursuit of desire can lead to unintended and transformative consequences	14
34	heed, advice, warnings	Heed warnings and advice, especially from parents, to avoid dangerous situations and consequences	14
35	arrogance, stubbornness, powers	Arrogance and mockery can lead to one's downfall, and that dedication to one's own path, even in the face of ridicule, is a virtue	13
36	nobility, superficial, worth	True worth and nobility come from within, and perseverance and acceptance can lead to unexpected and rewarding transformations	13
37	anger, discernment, deep	Desperate decisions made in moments of crisis can lead to profound regret, and one must carefully weigh the consequences of their actions	13
38	past, heritage, preserving	The importance of preserving and sharing local history and folklore to keep the past alive and understand the cultural heritage of a place	13

Continued on next page

Cluster	Top Words	Sample Moral	Size
39	flattery, gossip, exploit	Attempting to gain favor through insincerity and espionage is unwelcome among the noble, and authenticity is valued over flattery and deceit	13
40	deceiving, strangers, appearances	Appearances can be deceiving, and one should be cautious in their dealings, especially with strangers who may not be what they seem	12
41	natural, history, legends	Appreciate the rich tapestry of history and folklore that shapes our understanding of natural landmarks and their cultural significance	12
42	curiosity, overconfidence, disobedience	Exercise caution and respect towards the unknown, as overconfidence or curiosity in unfamiliar territories can lead to perilous situations	11
43	fulfillment, away, come	True fulfillment and happiness come from following one's heart and spiritual calling, even if it leads away from societal expectations and norms	11
44	cautious, offers, appears	Be cautious of tempting offers and to respect the unknown, as greed and disrespect for the supernatural can lead to unforeseen consequences	11
45	changes, luck, fleeting	Unexpected events can lead to significant changes in life, and what may initially seem like misfortune can ultimately bring about good fortune	11
46	respecting, nature, stewardship	Every creature has a specific environment where it thrives best, and understanding and respecting these natural habitats is crucial for harmony in the natural world	11
47	refusal, isolation, cooperate	The refusal to share and cooperate within a family can lead to division and loss	11
48	supernatural, entities, prioritize	Encounters with the supernatural can bring unforeseen consequences, highlighting the themes of curiosity, fear, and the unpredictable nature of engaging with mystical entities	11
49	comes, responsibility, wisely	With great power comes the responsibility to use it wisely and compassionately	11
50	violence, war, victims	Heinous acts of violence and cruelty leave a lasting impact on communities and the memories of victims, echoing through history as a reminder of the depths of human brutality	11
51	forgiveness, healing, improving	Mercy and forgiveness can triumph over vengeance and hatred, leading to redemption and healing	10
52	leadership, influence, guidance	True leadership and deliverance come from humility, service to a higher cause, and the courage to fight for freedom and justice	10
53	agreements, promises, breaking	Honor your agreements and be wary of the consequences of breaking promises	10
54	unpredictable, immortality, pursuit	The pursuit of immortality and exploration of the unknown can lead to unforeseen and transformative consequences	9

Continued on next page

Cluster	Top Words	Sample Moral	Size
55	sacrifices, spiritual, introspection	Extreme sacrifices made out of love and concern can lead to tragic outcomes but also have the power to transform and enlighten individuals about the sanctity of life	9
56	knowledge, forbidden, pursuit	Impatience and the pursuit of forbidden knowledge can lead to unforeseen consequences and loss	9
57	overly, relentless, loneliness	Setting unreasonably high standards and being overly proud can lead to loneliness and the loss of potential happiness	9
58	unseen, respect, mysterious	Respect and caution towards the unseen and mysterious forces of folklore are essential in navigating their potential influence on the natural and human world	9
59	imagination, storytelling, power	Appreciate the wonder and whimsy in the world, and to recognize the value of imagination and storytelling in our lives	9
60	worth, appearances, conventional	True worth and happiness are found beyond appearances and promises should be honored	9
61	content, risking, satisfied	One should value and be content with what they have, rather than risking it all for the uncertain promises of wealth and ambition	8
62	small, interactions, misunderstanding	Misunderstanding and fear can lead to tragic outcomes, highlighting the importance of compassion and understanding in interactions with others, including animals	8
63	temptation, yielding, judgment	The actions of one individual should not dictate the judgment or fate of an entire group, and that true virtue and loyalty are invaluable and should be cherished	8
64	quick, thinking, situations	Wit and quick thinking can save one from dangerous situations	8
65	assets, traditional, education	Knowledge and wisdom are the most precious assets one can possess, surpassing even the most tangible and rare treasures	7
66	unequal, favoritism, selfishness	Favoritism and unequal treatment can lead to mistrust and ultimately result in loss rather than loyalty	7
67	fleeting, outweigh, time	The longing for one's home and loved ones can outweigh the allure of even the most enchanting and idyllic surroundings	7
68	rewards, observing, attentiveness	Honoring one's promises and displaying integrity and honesty in one's actions leads to reward and prosperity	7
69	greater, suffering, actually	Sometimes what we perceive as misfortune may actually be divine mercy sparing us from greater suffering	7
70	unresolved, past, prevent	Unresolved misunderstandings and the inability to move past grief can lead to a life consumed by sorrow and regret	6
71	youth, creativity, light	Facing and overcoming troubles in youth can lead to a happier and more fulfilling future	6

Continued on next page

Cluster	Top Words	Sample Moral	Size
72	laziness, missed, negative	Avoiding work and responsibilities leads to negative consequences and additional burdens	6
73	solutions, solution, problems	Magical or unexpected solutions to problems can bring relief, but they also require responsibility and knowledge to manage them properly	6
74	favors, bold, fortune	Sometimes, fortune favors the bold, and success can come from the most unexpected circumstances and actions	6
75	everyday, embrace, celebrate	Embrace the wonders of childhood imagination and the adventures it brings, valuing the joy and uniqueness in life's journey	6
76	memories, rituals, collect	Devotion and respect for loved ones transcend physical presence and can be honored through heartfelt rituals and memories	6
77	exist, share, regardless	Beauty and value exist in all of creation, regardless of appearances or societal hierarchies, and that love and appreciation should be extended to all beings	6
78	traditions, beliefs, imposition	Love and humanity can soften the harshest of judgments and bring unity, but cultural and religious differences can lead to the suppression of joyful traditions and the imposition of a more somber way of life	5
79	practical, different, application	Adopting simple strategies and learning from others can significantly improve one's abilities and lead to success	5
80	forces, endeavors, appeasement	Respect for and appeasement of supernatural forces can protect and aid individuals in their endeavors	5
81	driven, selfless, act	A lifetime of stinginess cannot be redeemed by a single, reluctant act of charity, and generosity towards others is essential for spiritual redemption	5
82	grief, living, accepting	Grief and tragedy can have far-reaching effects, impacting not just the immediate victims but also the wider community and environment in a cascading chain of reactions	5
83	bond, invaluable, parents	The unbreakable bond between parents and their children can overcome great distances and challenges	5
84	emerge, harmony, order	Ambition and conflict can lead to destruction, but from such events, new beginnings and beauty can emerge, highlighting the importance of harmony and unity	5
85	mysteries, unsolved, unexplained	Some mysteries remain unsolved, leaving communities and individuals to speculate and wonder about the truth behind enigmatic events and people	5
86	appearance, intellect, physical	Never underestimate someone based on their appearance, as hidden talents and abilities can lead to unexpected success	5
87	cycles, renewal, year	The natural cycles and changes of the earth, including the harshness of frost, are part of a greater plan and serve to rejuvenate and awaken life	5

Continued on next page

Cluster	Top Words	Sample Moral	Size
88	treasure, secrecy, riches	Valuable discoveries, even if initially kept secret, are destined to be shared for the greater good of the community	4
89	limitations, stepping, station	One should accept their own limitations and roles as intended by the higher powers, and not covet the abilities or possessions of others, to avoid getting into trouble	4
90	escape, fate, matter	No matter how clever or quick one may be, it's impossible to escape one's ultimate fate	4
91	belong, taking, does	Taking what does not belong to you can lead to unforeseen and unpleasant consequences	4
92	shortcomings, similar, differently	Individuals often criticize the flaws in others while ignoring their own shortcomings	4
93	dead, avoid, invoking	Fighting and deceit lead to lasting scars and consequences, and it is best to avoid conflict altogether	4
94	marriage, important, set	Each choice in marriage comes with its own set of dynamics and challenges, and one should carefully consider these before making a decision	4
95	assuming, position, judgment	Judgment and punishment are the prerogatives of a higher authority, and assuming such roles without wisdom or right can lead to one's downfall	4
96	transcend, devotion, heavens	True love and devotion are eternal and can transcend even the boundaries between earth and the heavens	4
97	nature, advancements, design	Innovation and learning from nature can lead to significant advancements and improvements in human life	4
98	rare, attains, revenues	Even with determination and the invocation of higher powers, some pursuits may ultimately prove elusive and beyond one's grasp	4
99	adaptation, adapting, america	Adaptation and unity are crucial for survival in the face of inevitable change and threats	4
100	humor, lightening, uncertainty	Wit and humor can provide a unique perspective on life's mysteries and challenges, often lightening the mood in the face of uncertainty	4
101	longing, dissolution, prepare	While it's natural to yearn for what we lack, we must also prepare for the impermanence of joy and the inevitable changes that life brings	4
102	animals, solving, problem	Intelligence and reasoning are not exclusive to humans, as even animals can exhibit strategic thinking and problem-solving abilities	4
103	elders, treating, red	Treating elders with respect and kindness is more valuable than material wealth, and deceit and neglect towards them will ultimately lead to regret and loss	3
104	attain, merits, heaven	Every profession and way of life has its merits, and anyone can attain heaven through their own path and virtues	3
105	creation, complexities, imperfections	Curiosity and discovery are fundamental to the creation and understanding of the world and its inhabitants	3

Continued on next page

Cluster	Top Words	Sample Moral	Size
106	look, leap, thoroughly	Look before you leap, emphasizing the importance of caution and foresight before taking action	3
107	sell, goods, hoarding	Obsessively hoarding wealth without using it is a futile endeavor that can lead to its ultimate loss or waste	3
108	abusive, inevitably, rid	Abusive behavior and clever manipulation of words cannot justify wrongdoing, and such actions will ultimately lead to deserved punishment	3
109	inevitable, death, accepted	Death is an inevitable part of life that should be accepted gracefully, and the signs of aging are natural messengers of its approach	3
110	friendship, friends, unselfishness	True friendship and unselfishness are more valuable than material wealth and can lead to unexpected rewards	3
111	rule, impartial, legal	True justice is impartial and inevitable, and one must accept the natural course of life and death without seeking to alter it for personal gain	3
112	coward, tie, safe	Don't tie your fate to a coward's	3
113	age, elderly, treat	One should not underestimate the value and abilities of the elderly or judge worth by age alone	3
114	futile, execute, unfulfilling	Devising a plan is futile without the willingness or courage to execute it	3
115	absurdity, story, implausible	The story, being a nonsensical and whimsical tale, does not have a clear moral but rather serves to entertain and amuse with its absurdity	3
116	status, eyes, equally	Heavenly rewards do not reflect earthly wealth or status, but rather, everyone is valued equally in the eyes of the divine, with differences in reception highlighting the rarity rather than the value of the individual	3
117	evidence, stubborn, laughter	Stubborn belief without evidence can lead to public embarrassment and ridicule	3
118	thoughtlessness, indigenous, resources	The devastating impact of human greed and thoughtlessness on nature and indigenous cultures	2
119	disputes, benefiting, competence	Disputes should be resolved by demonstrating competence and productivity rather than through prolonged and costly legal battles	2
120	underestimating, underestimated, opponents	Underestimating the capabilities of the seemingly weak can lead to unexpected and transformative outcomes	2
121	christmas, young, books	Embrace the joy and magic that the spirit of christmas brings to hearts young and old	2
122	erased, obtained, mark	The cycle of vengeance only leads to mutual destruction and lasting sorrow	2
123	forever, hidden, kept	Secrets cannot be hidden forever and will eventually come to light, often in unexpected ways	2
124	homeland, ready, patriotism	True heroes, with their strength and vigilance, remain ever-ready to protect their homeland, embodying the spirit of resilience and patriotism	2

Continued on next page

Cluster	Top Words	Sample Moral	Size
125	royal, elevation, daughters	Having beautiful daughters can lead to unexpected fortune and elevation into the royal family, showcasing the potential for beauty to change one's destiny	1
126	pretty, stranger, marry	Never marry a stranger, no matter how pretty she may be	1
127	remember, final, resting	Failing to honor and remember the contributions and final resting places of historical figures leads to a loss of heritage and understanding for future generations	1
128	instruments, guides, belong	True creativity and honor belong to the higher power that guides the instruments, not to the instruments themselves	1

Table 16: All clusters for the full-sentence morals

Cluster	Top Words	Size	Cluster	Top Words	Size
0	daughter, king, princess	62	50	popcorn, fish, jelly	13
1	boy, farmer, old	55	51	fool, clay, coat	13
2	wo, hen, woman	51	52	eagle, wings, crows	12
3	father, son, said	45	53	teeny, tiny, church	12
4	jack, asgard, gods	45	54	place, noises, indians	12
5	jack, mr, shark	45	55	lad, steel, child	12
6	mother, jack, said	43	56	ship, shore, harbor	12
7	jack, place, pony	43	57	sun, light, night	12
8	jack, good, man	43	58	umbrella, witch, girl	12
9	shepherd, dervish, dog	41	59	fisherman, fish, crab	11
10	jack, place, says	37	60	witch, grove, girl	11
11	king, prince, wife	37	61	cinder, fairies, prince	11
12	bear, badger, deer	37	62	beetle, nose, long	11
13	said, bell, little	36	63	tin, soldier, boat	11
14	place, knight, mountain	36	64	fairies, queen, priest	10
15	shoemaker, soldier, money	35	65	princess, rings, spring	10
16	boy, bear, man	33	66	owl, pipe, rake	10
17	jack, count, place	33	67	hammer, sun, asgard	10
18	mother, maiden, children	31	68	wo, sparrow, tartars	10
19	jack, place, fafnir	27	69	bonze, rice, quails	9
20	lamb, garden, thistle	25	70	monkey, monkeys, razor	9
21	portuguese, bird, duck	25	71	crocodile, lamp, traveller	9
22	place, lodge, glooskap	23	72	oo, oom, reeled	9
23	place, monk, jack	22	73	toad, buzzard, bug	9
24	maiden, horse, king	22	74	christmas, ax, dolls	8
25	musician, parasol, owl	22	75	gipsy, devil, god	8
26	hoodie, brother, fox	22	76	abbess, convent, pouch	7
27	tiger, fox, hare	22	77	tailor, stag, town	7
28	gods, kettle, asgard	22	78	frog, croaked, stairs	6
29	cat, mouse, rooster	21	79	peasant, ouyan, flesh	6
30	dingo, horse, balloons	21	80	hedgehog, hare, exchange	6
31	king, barber, prince	21	81	ogre, cat, king	6

Continued on next page

Cluster	Top Words	Size	Cluster	Top Words	Size
32	rod, water, hen	21	82	dogs, flea, pepper	6
33	jan, sick, sack	21	83	suspicions, molasses, jug	6
34	people, hatrack, baboons	20	84	teapot, little, flea	5
35	ole, dat, ses	20	85	church, troll, christmas	4
36	fairy, step, shining	18	86	grasshoppers, dollar, brass	4
37	magician, devil, scholar	18	87	accordion, blind, toboggan	4
38	place, ferryman, river	17	88	magpies, seventh, bridge	4
39	aponitolau, hen, scissors	17	89	porcelain, hare, cage	4
40	half, wirreenun, brothers	17	90	says, skeins, noo	3
41	flower, flowers, branch	17	91	prize, oom, consideration	3
42	kite, sun, sky	16	92	magicians, instantly, ring	3
43	devil, jack, emperor	16	93	trina, honey, work	3
44	chimney, sweep, said	15	94	hat, thieves, crooked	3
45	brother, woe, rich	15	95	crabs, shrimp, waves	3
46	emperor, apprentice, man	14	96	slipper, christian, jewish	2
47	organ, open, grinder	14	97	birch, break, bulls	2
48	elephant, goblins, hippo	13	98	gringo, package, ball	2
49	snow, stove, summer	13			

Table 17: All clusters for the full text clustering. The top 10 results may differ from Table 5 because only words of length 3 or more were included, and the words “jack” and “place” were removed since they were artifacts of the entity replacements with spaCy (see section B.2).