jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Towards a Perspectival Moral History of the Novel Using LLMs

Andrew Piper[1]

1. Languages, Literatures, and Cultures, McGill University, Montreal, Canada.

Citation
Andrew Piper (2025). "Towards a Perspectival Moral History of the Novel Using LLMs". In: *Journal of Computational Literary Studies* 4 (1). 10.48694/jcls.4168

Date published 2025-11-14
Date accepted 2025-09-29
Date received 2025-01-30

Keywords
novels, large language models, fiction, narrative archetypes, ethical criticism, world literature, wikidata

Reviewers
Hans Ole Hatzel, Anonymous Reviewer

Note
This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 4th Annual Conference of Computational Literary Studies at Krakow, Poland, in July 2025.

Abstract. This paper introduces a new framework for studying the moral history of the novel through the lens of large language models (LLMs). Drawing on over 9,000 Wikipedia plot summaries of 20th- and 21st-century novels, it demonstrates how LLMs can surface the implicit life lessons – or story morals – encoded in narrative summaries at scale. Building on recent work in moral inference and narrative abstraction, the study proposes a reflexive, perspectival approach that emphasizes interpretation over taxonomy. To account for the semantic variability of LLM-generated morals, the study employs a randomized prompt assignment strategy and analyzes the resulting moral keywords using co-occurrence networks and hierarchical clustering, enabling the identification of latent moral communities and comparison across modeling approaches and time. Taken together, the findings argue for the value of LLMs not only in extracting narrative values, but in enabling a new, culturally situated view of literary history through computational means.

1. Introduction

A long tradition of literary criticism has emphasized the fundamental importance of understanding the ethical concerns of stories. As Wayne Booth has argued, "All stories teach" (Booth 1998, 354). Indeed, the didactic function of storytelling – that stories have a moral or lesson to impart – is one of the oldest known functions of storytelling (Gregory 2010). Aesop's Fables are the best known version in the West, but similar types of tales exist in both Hindu (Panchatantra) and Buddhist (Jatakas) traditions that date back to around the fifth century BCE.

While we typically associate the concept of 'story morals' with such traditional genres, critics like Booth (1998) and Nussbaum (1998) have argued that values-driven schemas are intrinsic to narratives more generally. As Russell and Van Den Broek (1992, 344) 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." In this sense, stories need not explicitly communicate moral sentiments (e.g., "Kindness is good" or "Thou shalt not murder"). Rather, they can address general life lessons that may draw from, reinforce, challenge or extend existing moral frameworks.

This project seeks to construct a perspectival moral history of the novel by leveraging large language models to distill the central values encoded in narratives. By 'moral history' I mean the implicit or explicit general life lessons conveyed by stories and storytellers over time. What does fiction teach us? And how is this historically and culturally

inflected? I use the term 'perspectival' here to capture a sense of the interpretive nature of the project, that narrative values and lessons are not independent of observation but are *seen* and derived *from some point of view*.

Capturing story morals is thus tied to the longstanding narratological focus on understanding narrative archetypes or schemas (Brewer and Lichtenstein 1980; Campbell [1949] 2008; Frye [1957] 2020; Genette 1992; Propp [1928] 1968; Thompson 1955). As cognitive scientists have argued, schemas are crucial ways through which we process experience (Berns 2022). Where much of this earlier work focused on content-driven questions ('what happened?'), the attention to narrative morals focuses more on the *values* and *intentions* of the storyteller, i.e., 'why was this told?' Like any schema, the story moral aims to distill an organizing principle that governs the generation and selection of narrative events and narrative perspective.

Large Language Models (LLMs) offer a potentially valuable new resource for this task given the abstractive and synthetic nature of story morals. While LLMs still suffer from hallucination with respect to fact-based extraction (L. Huang et al. 2023), they have exhibited significant progress when it comes to abstractive reasoning tasks such as narrative summarization (Subbiah et al. 2024; Zhang et al. 2024) or topic labeling (Pham et al. 2024; Piper and S. Wu 2025). Indeed, deriving a story moral is in many ways analogous to the tasks of narrative summarization or topic labeling, where a model is tasked with abstracting higher-level narrative messages that are not explicitly present in the text.

Another affordance of LLMs is that given their generative nature they allow researchers to infer story morals in an unsupervised fashion, i.e., from the 'bottom-up.' Rather than apply a pre-existing taxonomy that may not account for the diversity of cultural behavior, as Dundes (1962) long ago criticized, LLMs enable researchers to potentially surface a broader array of values and practices. This does not mean, however, that LLMs are neutral observers. They are of course 'pre-trained.' They introduce yet another layer of perspective into the interpretive process that we need to account for.

In this paper, I outline a workflow for this project I am calling a perspectival moral history of the novel (Figure 1, Figure 2, Figure 3). It is crucial to remind ourselves of Underwood's dictum that we do not yet have a clear understanding of the broad outlines of literary history, including the moral landscape of the modern novel (Underwood 2019). To undertake this project I engage in a series of steps of LLM-assisted narrative interpretation that move towards increasing levels of generality and structure (Figure 1). Beginning with stories themselves as interpretations of the world, it proceeds through summarization and moralization and ends with the identification of latent moral structures using co-occurrence networks and hierarchical clustering as two possible exploratory methods. As I will demonstrate, each step involves an act of perspective-taking that we need to build into the workflow.

This project utilizes Wikidata as its principal source of data, with plot summaries in particular as the primary data object. While traditional criticism may balk at using Wikipedia for literary study (or plot summaries for that matter), recent work in computational literary studies has illustrated Wikipedia to be an important resource for the study of literature, especially comparative literature. It provides one kind of 'lay reader'

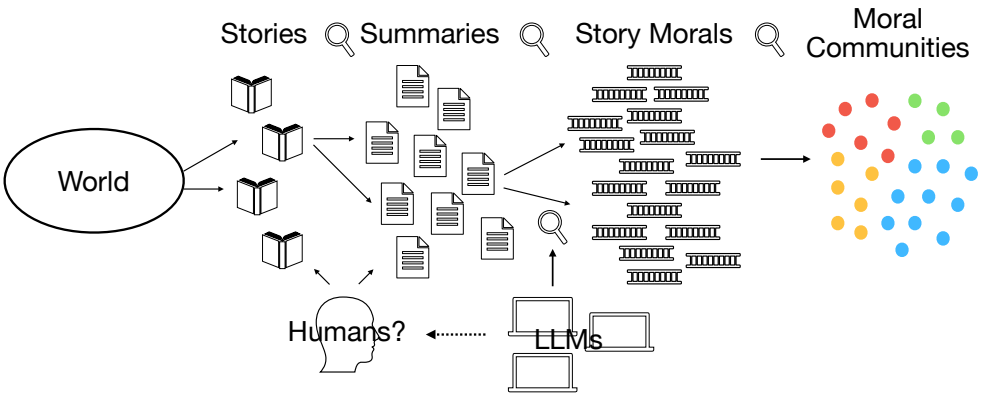

Figure 1: Overview of the story moral extraction task. Looking glasses indicate interpretive or perspectival transitions.

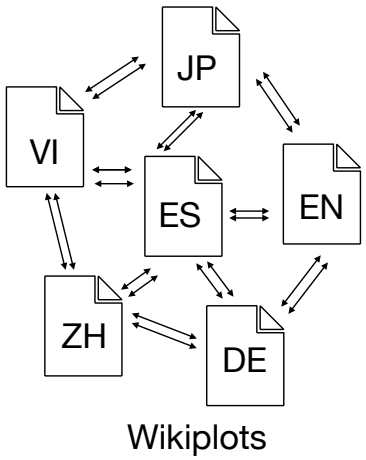

Figure 2: Schema of the many-to-many relationship of wikiplots between each language edition.

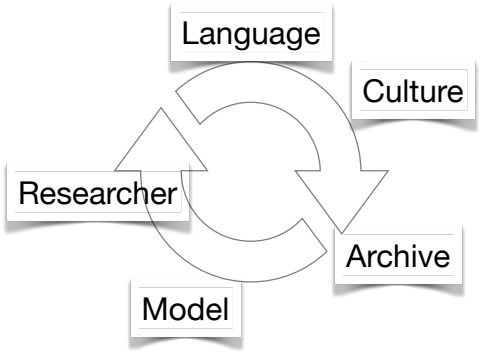

Figure 3: Schema of the LLM-based hermeneutic circle.

view of literary history. As Fischer et al. (2023, 1) write in their preface to the special issue, *Wikipedia, Wikidata, and World Literature*: "Despite the longstanding debate over the canon, what Wikipedia and Wikidata show us is that there is no monolithic canon, but many canons, depending on the data you choose to examine."

The biases of Wikipedia contributors in terms of demographic distribution, for example, are well known (Wikipedia Contributors 2024). As I show in Figure 4 (section 4), this affects the kinds of genres represented in the data, the time periods for which there is substantial data, and the choice of regions represented. But this is no less biased than a dataset generated by academic elites. Each provides a different perspective on literary history.

Wikidata is thus valuable for two principal reasons for this project. The first is the contribution of human-generated narrative summaries. Extracting 'story morals' depends on the ability to compress a long, complex narrative down to its essential components. Summarization is a key step in the workflow (Hobson et al. 2024).

The second affordance of Wikidata is its multilingual and multicultural nature, i.e., the *interconnected* nature of cultural perspectives that it contains. Plot summaries contained

in Wikidata provide insights into these cultural perspectives: both in terms of what works are chosen to be discussed and also in how the stories are reflected through the practice of summarization. Rather than provide a canonical summary of a canonical list of stories, Wikidata allows us to observe regional interpretations of story content through the practice of summarization and selection (Figure 2; Hatzel and Biemann 2024). Each language Wiki provides a perspective not only of its own cultural artifacts (English-language summaries of stories originally written in English) but also other cultures (English-language summaries of stories originally written in Japanse and vice versa). Wikidata allows us to move past the idealized 'view from nowhere' and instead contend with the idea of a 'situated world literature' (Cheah 2015; Figure 3).

For the purposes of this paper I will illustrate the workflow on a single Wikidata language set (English) and leave to future work the challenge of multilingual moral reasoning. The goal here is to demonstrate the ability of LLMs to generate common-sense based interpretations of story morals given narrative summaries as inputs and experiment with the process of moral aggregation (the final step shown in Figure 1). To do so, I build off of prior work validating LLMs' capacity to generate story morals across numerous kinds of genres (Hobson et al. 2024; Zhou et al. 2024). In this paper, my focus will be on refining this workflow for this particular data and exploring the kinds of interpretive value this produces for literary historical analysis. As I hope to show, this method can generate novel insights about the moral landscape of novels at large scale.

2. Prior Work

The organization of stories into broad, overarching categories is deeply rooted in the field of narratology (Brewer and Lichtenstein 1980; Campbell [1949] 2008; Frye [1957] 2020; Genette 1992; Propp [1928] 1968; Thompson 1955). 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.

One of the fundamental challenges for this work is deciding how to select and identify appropriate schemas as well as their level of generality. 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 (Sims et al. 2019; Vauth et al. 2021; Yan and Tang 2023). The chaining together of event schemas has been integral to operationalizing the concept of "plot" (Kukkonen 2014), including plot summaries and plotlines (Anantharama et al. 2022; Rashkin et al. 2020).

Other work has focused on detecting higher-level schemas such as "conflict" and "resolution" (Frermann et al. 2023), turning points (Ouyang and McKeown 2015; Piper 2015), folktale motifs (Karsdorp and Bosch 2013), story types such as "rags-to-riches" (Fudolig et al. 2023; Reagan et al. 2016), and the more traditional concept of "genre" (Dai and R. Huang 2021; Kundalia et al. 2020; Wilkens 2016).

The attention to story morals naturally draws connections to work on Moral Foundation Theory (Graham et al. 2013), one of the more popular frameworks in the social sciences

for thinking about the moral perspectives of cultures. MFT posits that human moral reasoning is built upon a set of innate psychological foundations shaped by evolutionary processes. These foundations – such as care, fairness, loyalty, authority, sanctity, and liberty – underlie cultural variations in moral values and guide ethical decision-making. Work in NLP has attempted to surface moral foundations in texts such as tweets (Liscio et al. 2022; Rezapour et al. 2019; Roy and Goldwasser 2021; Roy et al. 2023) and folktales (W. Wu et al. 2023), as well as identifying the potential moral foundations of LLMs (Abdulhai et al. 2023; Scherrer et al. 2023). Vida et al. (2023) provide a useful overview of the use of 'morals' as a concept within NLP research.

The key difference between the present work and prior work related to MFT or the study of narrative archetypes is the absence of a pre-defined moral taxonomy. My aim here is to uncover open-ended narrative-based moral frameworks using the generative insights of Large Language Models. As Hobson et al. (2024) have shown, LLMs like GPT produce interpretations that are both within the range of variance of human responses and also most often preferred by independent human judges. As I will illustrate in the next section, there are steps we can take to broaden the semantic variance generated by LLMs to capture a wider cultural 'perspective' from any given model. Future work will have to consider the extent to which LLMs can approximate multi-lingual and multi-cultural perspectives in their outputs. For now, however, I focus on examining LLM reasoning about narrative morals in a single language.

3. Methods: Surfacing Story Morals Using LLMs

Hobson et al. (2024, 13000) have proposed and validated a workflow for story moral extraction using LLMs. In that work, the authors "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. Story morals understood as lessons mean that they are not strictly synonymous with the idea of moral "sentiments" (Vida et al. 2023). They focus instead on forms of behavior and belief that may be integrated into or derived from pre-existing moral frameworks but are not necessarily aligned with existing moral schemas.

To generate a story moral from a text, Hobson et al. (2024) use a two-level prompting approach. They first ask the model to output the moral of a story in a single sentence and then have the model output two keywords: one negative and one positive that encapsulate the story moral. I modify this approach here in two ways that are relevant to the data: First, I ask for three keywords instead of single positive and negative keyword to allow for more overall semantic diversity; second, I include a catch for the model to not output a story moral if the input is insufficient and also forbid the use of the word empathy.[1] Table 1 (top) provides an overview of the base prompt structure.

One aspect not explored by Hobson et al. (2024) is the issue of variability in generative outputs. Large language models are known to be sensitive to prompt formulation, with

1. While the exclusion of the word *empathy* may appear subjective, I have found that models have an overwhelming and at times misleading affinity for this term. While this deserves further attention, as we will see the models have no trouble substituting synonymous keywords for this value.

Prompt Structure Overview	
Unit	**Prompt**
Level 1	`What is the moral of this story? State your answer as a single sentence. If not enough information, write NONE.`
Level 2	`Can you reduce this to three keywords? Don't use the word empathy.`
Factorial Prompt Variants	
Factor	**Levels / Description**
Information Ordering	Story summary appears in **Top** or **Bottom**.
Role Framing	**Present** or **Absent**: `Today, you are an expert story interpreter. I will give you a book summary and ask you a question about it.`
Question Phrasing	**Direct:** `What is the moral of this story?` **Interpretive:** `How might one interpret the moral of this story?`

Table 1: Base prompting structure (Top) and experimental factors used in our 2×2×2 design (bottom) to evaluate model sensitivity to moral extraction prompts.

Role = yes; Order = top; Phrasing = Interpretive	**Role = no; Order = bottom; Phrasing = direct**
`Today, you are an expert story interpreter. I will give you a book summary and ask you a question about it. Here is the summary: [SUMMARY] How might one interpret the moral of this story? State your answer as a single sentence. If not enough information, write NONE.`	`What is the moral of this story? State your answer as a single sentence. If not enough information, write NONE. Here is the summary: [SUMMARY]`

Table 2: Examples of two prompt variants used in our 2×2×2 design. The left shows all three positive changes while the right is the original base prompt.

even minor changes in phrasing often resulting in divergent outputs (Lu et al. 2022; Reynolds and McDonell 2021; Sclar et al. 2023; Webson and Pavlick 2022). This prompt sensitivity poses challenges for both the interpretability and replicability of LLM-based analyses, particularly in open-ended tasks such as narrative understanding or moral reasoning.

To assess the extent of prompt sensitivity for our moral extraction task, I conducted a controlled experiment using a random sample of 100 story summaries. Each summary was paired with eight prompting variants derived from a fully crossed 2×2×2 factorial design (N=800) (Table 1, bottom). This design systematically varied three factors that are independent of the base prompt meaning: (1) expert role framing, (2) information ordering, and (3) question phrasing. All prompts in the experiment were submitted to OpenAI's `gpt-4o-mini-2024-07-18` model via the API, using a temperature setting of 0.0 to minimize sampling variance. I show two examples of the factorial design prompt structure in Table 2.

To quantify the effects of prompt variation, I computed pairwise Jaccard similarities

between the keyword outputs generated by each prompt configuration for the same summary. This resulted in 28 pairwise comparisons across the eight prompt variants for each of the 100 summaries. The mean Jaccard similarity across all prompt pairs was 0.38, with individual pairs ranging from 0.29 to 0.58, indicating that on average less than 40% of keywords overlapped between prompting runs on the exact same story set. The most divergent combination was `no_role-bottom-interpretive_phrasing` while the most convergent combination was `role-bottom-direct_phrasing`.

This high degree of variation across prompt types gives us a good indication of the interpretive problem LLMs introduce. Even with the same model and the same temperature, we can get divergent outcomes due to prompt structure. We can also expect this to be true at the level of the models themselves. Different models will likely provide different answers. To be sure, these answers are not independent of one another (i.e., random), nor are they in some sense inaccurate because of their variability. As Hobson et al. 2024 show, LLM story morals are generally within the variance of human responses and this consensus view is most often preferred by human judges. But the variability does tell us that the semantic space of responses is wider than a single answer would indicate. It gives us a first, valuable insight into the perspectival nature of LLMs as interpretive agents.

Given this variance, I opt not to rely on a single formulation for the moral extraction task (i.e., a 'best prompt'). Instead, I aim to capture this interpretive diversity revealed by the pilot by aggregating across multiple prompt perspectives (referred to as "convergent validity" in the psychometric literature). To do this in a scalable way, I employed a randomized prompt assignment strategy: Each story summary in the full dataset was paired with exactly one prompt variant, but all eight prompt types were evenly distributed across the corpus. As in the pilot experiment, prompts were submitted to OpenAI's `gpt-4o-mini-2024-07-18` model via the API. This approach allowed me to preserve the diversity introduced by prompt wording while minimizing the computational and financial costs associated with full prompt replication across multiple models. Future work could explore additional prompt manipulation along with the inclusion of more models to capture the full range of possible interpretive diversity encoded in models. I provide examples of model outputs in Table 3.

4. Data: Wikipedia and Literary History

For the purposes of this project, I use a collection of 9,756 human-written story summaries drawn from English-language Wikipedia for books published since the beginning of the twentieth century. These summaries are based on Mark Riedl's Wikiplots dataset (Riedl 2024), which consists of 112,936 story plots extracted from any English language article that contains a sub-header that contains the word "plot" (e.g., "Plot," "Plot Summary," etc.). David Bamman has provided a filtered version of this data that subsets by "creative works" and "books" to capture the genre of novels (Bamman 2024). I further filter Bamman's version by removing any entries where the summary field is less than 40 words or overly long (> 1,325 words). Additionally, I filter for only publications published after 1900, which represents the bulk of summaries in the data. This yields summaries representing novels published in 116 countries, with the vast majority from

Title	Label	Model Output
White Fang	Moral	*The moral of the story is that resilience, loyalty, and the capacity for change are powerful forces that can transform even the most dire circumstances.*
	Keywords	Resilience, Loyalty, Transformation
IQ84	Moral	*The moral of the story is that the quest for connection and understanding transcends the boundaries of reality and time, urging us to confront our past and the complexities of our world to find true love and purpose.*
	Keywords	Connection, Transcendence, Purpose
Play it as it lays	Moral	*The moral of the story is that the search for personal identity and meaning amidst the chaos of modern life can lead to self-destruction if not anchored by genuine connections and self-awareness.*
	Keywords	Identity, Chaos, Self-Destruction
The Golden Notebook	Moral	*The moral of "The Golden Notebook" is that the quest for personal and artistic wholeness requires confronting and integrating the fragmented aspects of one's identity and experiences.*
	Keywords	Integration, Identity, Wholeness

Table 3: Examples of story morals produced by GPT-4o-mini across well-known books.

English-speaking countries. The average summary is 372 words long with an interquartile range of 130 to 556 words. This dataset is accessible in our data repository and hereafter referred to as 'Wikiplots_Novel_EN.' Figure 4 illustrates some descriptive statistics of the data.

4.1 Validating Summary Quality

One question we might ask moving forward is whether the summaries are themselves reasonable representations of the books they claim to represent. As with all summarization assessment, this is not an easy question to answer. There is no right or best summary. Indeed, my research question is not principally interested in the morals of the underlying books themselves, but rather the morals of the books *as they are captured by the human summaries in different Wikipedias*.

That being said, in addition to the quality checks mentioned above (removing overly short summaries and adding a prompt catch for low information) I also perform a small validation study to estimate the quality of the summaries' relationship to their source texts to get a rough estimate of the relationship between the summaries and their sources.

For a subset of novels for which we have both the full text from Project Gutenberg and corresponding summaries in our dataset ($N = 122$), I estimate the semantic similarity between each novel and all candidate summaries. The underlying assumption is that an accurate summary should be semantically closest to the book it describes, reflecting a reliable condensation of its most salient content.

To measure semantic similarity, I divide each novel into 500-word chunks and embed both the chunks and the summaries using the Sentence-BERT model `all-MiniLM-L6-v2`

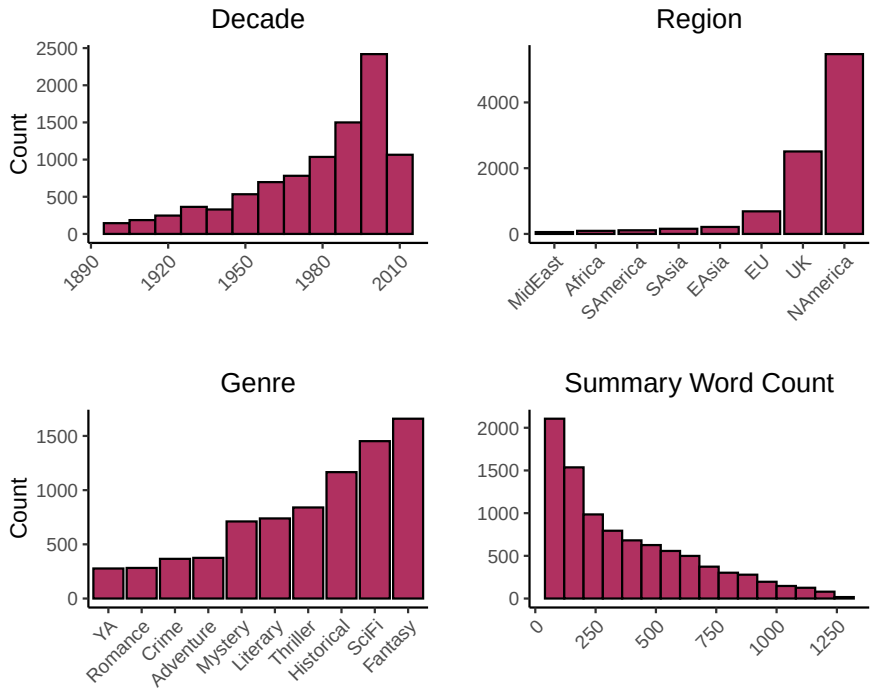

Figure 4: Overview of the Wikiplots_Novels_EN data used in this article.

from the sentence-transformers library. Each chunk is encoded into a 384-dimensional embedding vector with L2 normalization enabled (normalize_embeddings=True) to ensure comparability via cosine similarity. I then calculate the average cosine similarity between all embedded novel chunks and each candidate summary, selecting the highest-scoring match under both top-1 and top-3 conditions. The model achieves a top-1 matching accuracy of 72.80% and a top-3 accuracy of 87.20%. An error analysis of mismatches suggests that summary length alone does not account for misattribution, indicating that other factors may be influencing performance, including the coarseness of the model itself. Nevertheless, this preliminary analysis suggests that an overwhelming majority of summaries are indeed reflective of their source-texts and thus reasonable proxies for the underlying books. Future work can expand on this by studying the inter-cultural variation of story summaries on similar books.

5. Results

I begin my analysis by looking at the distribution of moral keywords. The first thing we can observe is the long-tailed nature of keywords with 1,383 unique moral keywords, 586 of those appearing just once, 408 appearing more than five times, and only 133 (10%) accounting for 80% of all occurrences. Table 4 provides a snapshot of the most frequent keywords across the entire dataset.

As we can see, our model and prompts provide novel insights into the high-level values associated with the modern novel as seen through the eyes of Wikipedians. One way to think about the contribution here is to contrast this taxonomy with the more traditional kinds of abstractive information such as topics that have traditionally been extracted

Keyword	Count
consequences	1138
resilience	909
identity	875
connection	837
love	779
understanding	710
courage	623
truth	592
loyalty	580
sacrifice	554

Table 4: Top 10 most frequent moral keywords for Wikiplots_Novels_EN.

from narratives. Seen in this way, the story moral framework provides a new lens to understand the narrative concerns of fiction over the past century not captured by topic modeling or thematic concerns.

To gain a deeper understanding of these keywords, we can measure co-occurrence patterns of moral keywords for the same stories. By transforming moral co-occurrences into a network graph, we can better understand story morals at two levels of scale: 1) local semantic neighborhoods that can illustrate an individual term's meaning by identifying other terms it most often occurs with and 2) broader latent moral structures that may exist across the dataset.

To do so, I first construct a co-occurrence network from the model outputs, where nodes represent moral keywords and edges indicate how often two keywords co-occur within the same story. To improve interpretability, I trim the network by filtering low-frequency edges (< 10) and nodes (< 5) (N=72), and then apply multiple community detection algorithms to identify clusters of related moral concepts.

To assess the robustness of the detected moral communities, I apply the following five community detection algorithms to the co-occurrence network: The Louvain method yields the highest modularity (0.31) with four communities, followed closely by the Fast Greedy algorithm (0.30) which also identifies four clusters. Walktrap produces a slightly lower modularity (0.27) and divides the network into five communities. Both Infomap and Label Propagation produced only two communities and yielded the lowest modularity scores (0.17), suggesting a weaker fit to the network's structure. Overall, the convergence of Louvain and Fast Greedy on a four-community solution with relatively high modularity supports the presence of a stable latent structure within the moral co-occurrence network.

Figure 5 visualizes the co-occurrence network using a force-directed graph layout and Louvain community detection. I include the three most frequent labels for each community. The illustration helps us see greater clarity around the semantic associations of the different keywords along with larger frameworks to which they belong. If we take four communities as a reasonable estimate, we can infer high-level groupings around distinct areas of Truth/Justice, Resilience, Identity/Growth, and Compassion.

A network graph is of course only one way of surfacing latent structure within the co-occurrence matrix. Each method will shift our understanding of the moral communities

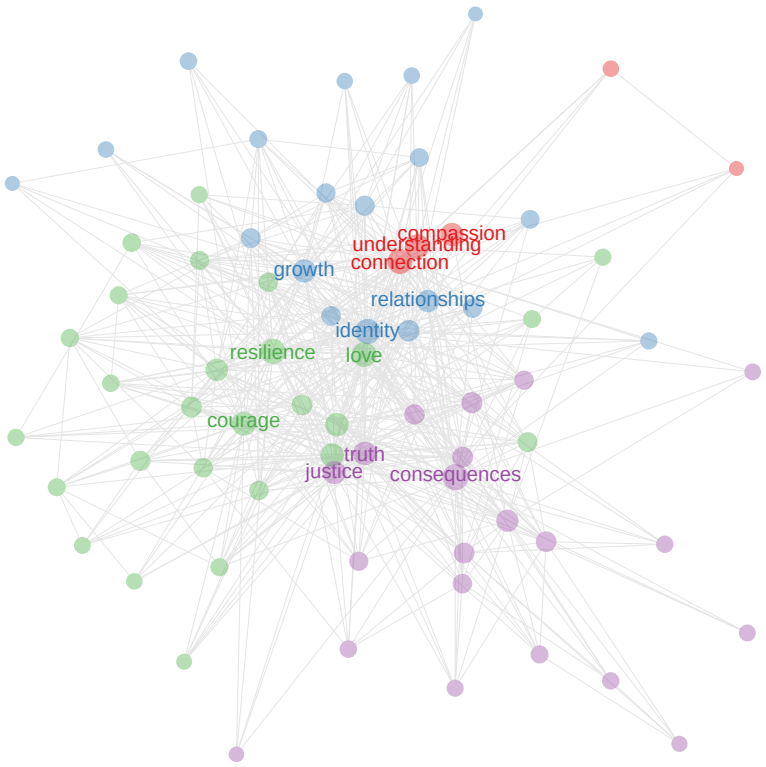

Figure 5: Co-occurrence network of moral keywords in the Wikiplots_Novels_EN corpus. Nodes represent moral concepts that appear together in story-level annotations, with edges weighted by the frequency of co-occurrence. The network is trimmed to include only edges with a frequency greater than 10 and nodes with at least five connections. Communities are identified using the Louvain method and labeled by color. Node size reflects the log frequency of each keyword, and labels illustrate the three most frequent keywords within each community.

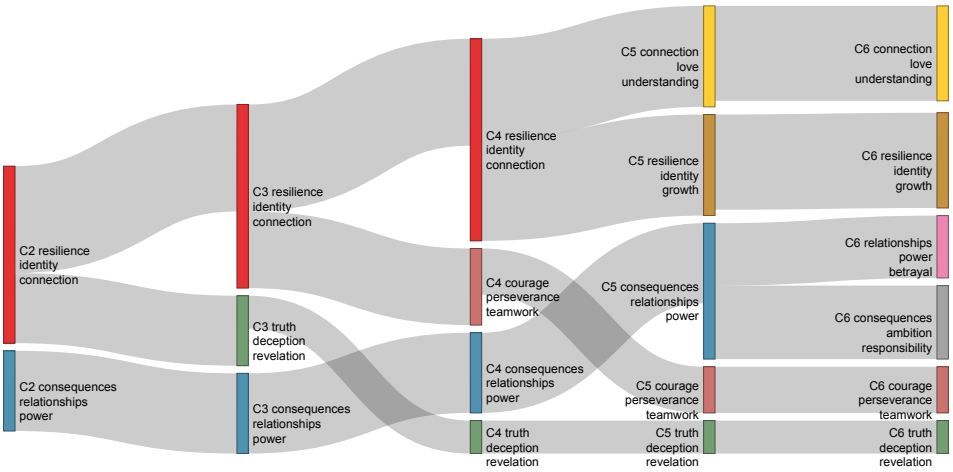

Figure 6: Hierarchical clustering of moral keywords visualized as a Sankey diagram. The diagram illustrates how clusters of moral concepts evolve across increasing levels of granularity, from k = 2 to k = 6. Each node represents a cluster of keywords identified through hierarchical clustering based on cosine distances between normalized co-occurrence vectors. Edges indicate how clusters at one level split into more fine-grained subgroups at the next. Nodes are labeled with the three most frequent keywords in each cluster. Cluster width reflects the average frequency of its top keywords.

by some degree. To explore the latent structure of moral keywords beyond discrete community detection, I also apply hierarchical clustering to the co-occurrence matrix (Figure 6). After filtering for keywords that appear in more than five stories (N=408), I compute pairwise cosine distances between normalized keyword vectors and perform agglomerative clustering using Ward's D.2 method. The resulting dendrogram reveals a multilevel hierarchy of moral groupings based on distributional similarity. To visualize how these groupings evolve across different levels of resolution, I generate a Sankey diagram showing how clusters at broader levels (e.g., k = 2) split into more refined subgroups at lower levels (up to k = 6). Cluster nodes in the Sankey diagram are labeled with their top three most frequent keywords, providing an interpretable summary of their semantic focus.

Here we see some further nuance to our network-based method. A *connection, love,* and *understanding* community emerges similar to the network, whereas *resilience* belongs to the *identity* and *growth* community rather than the *courage* and *perseverance* one. *Consequences*, the most frequent term overall, is located in a *power* and *ambition* cluster here with *truth* more squarely associated with its antonyms *deception* and *betrayal*. Each method produces slightly different insights into the data, where we might think about how to aggregate these different aggregative measures into a more holistic view.

Finally, I analyze changes in the prominence of moral clusters over time by comparing their relative frequency across decades (Figure 7). Using both network-based (Louvain) and hierarchical clustering methods, each moral keyword is assigned to a cluster and its frequency is tracked as a proportion of all moral keyword mentions in a given decade. The resulting time series visualization reveals a striking degree of stability: Despite cultural and temporal shifts, the relative ordering of cluster prominence remains largely consistent within each method. Moreover, the comparison highlights important differences in semantic emphasis. In the hierarchical model, the cluster labeled by

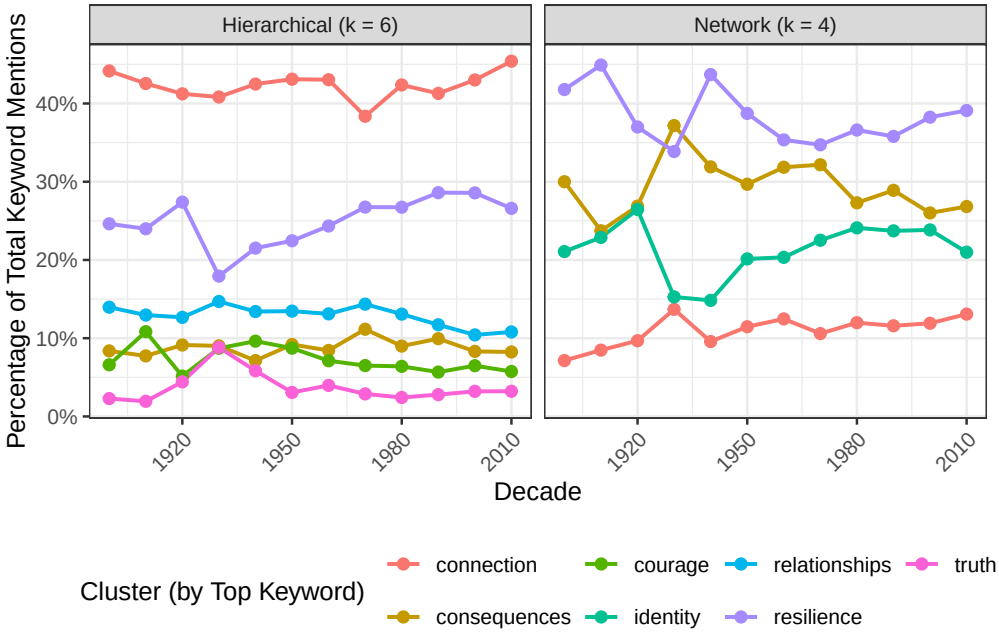

Figure 7: Relative frequency of moral clusters by decade, comparing hierarchical and network-based clustering methods. Each line represents a moral cluster labeled by its most frequent keyword, with vertical position indicating the proportion of total moral keyword mentions assigned to that cluster in each decade.

connection consistently dominates, suggesting a structurally central role for interpersonal and relational themes. By contrast, the network-based clustering foregrounds *resilience* as the most prominent and enduring cluster (with *resilience* second in the hierarchical model), pointing to a model of morality more centered on perseverance and individual strength. These contrasts illustrate how different modeling assumptions surface distinct moral contours within the same narrative data.

6. Conclusion

In this paper, I have endeavored to illustrate three salient points: The value of LLMs for extracting story morals at large scale, the value of Wikipedia for literary study, and the value of seeing literature through the lens of moral concerns. Each of these areas offers opportunities and challenges for future work.

As the work of Hobson et al. (2024) has shown and as we can see in section 5, LLMs offer us a reliable means of extracting high-level narrative representations that would have been unthinkable in the past. Nevertheless, even with the appearance of surface validity, it is worth pausing to ask in what ways LLMs interpretively orient us towards texts. Even though I have used a factorial variation approach to prompting and even though Hobson et al. (2024) show that LLM-generated morals are within the human range of labels, there are lingering questions about the overall semantic orientation of language models given their known cultural biases. Language models still *situate* us with respect to the text. Future work can focus on the effects of training data or fine tuning on the ways in which 'story moral' inference depends on prior knowledge – and more specifically 'whose knowledge.' To continue to foreground this issue of

perspectivalism, we need to continue to better understand the intrinsic perspectives encoded in LLMs.

In a similar vein, there is still much more work to do to understand the large-scale insights offered by this methodology as it relates to the history of the novel. Even if we take at face value the moral outputs as reasonable approximations of 'general' human judgments, what exactly do these commitments to 'truth,' 'resilience,' and 'connection' mean? Who are the principal agents of these stories? What are the common settings, genres, or topics that are associated with such lessons? Are there nuances to what it means to be 'resilient' or who can exemplify it? And what if we go further down the tree to understand novels of *redemption* or *sacrifice*? How many moral frameworks are there according to the novel and how can we identify a more nuanced literary history from this data? There is an opportunity here to explore methods for connecting the large-scale structural insights we've been seeing to more granular understanding of the moral concerns of novels.

Finally, to point in the other direction, how can we scale this workflow upwards to encapsulate the multilingual level? What are the limitations and potential solutions for working with less resourced languages than English when it comes to using LLMs? How well can LLMs embody 'cultural perspective'? Similarly, what limitations will we encounter in the data when we collect multiple language versions of Wikiplots?

Despite these challenges, there is a tremendous amount of promise offered by LLMs for the purpose of large-scale literary history and the moral history of the novel in particular. Stories teach. Surfacing the kinds of lesson encoded in stories is an exciting prospect. As we become less dependent on single, monolithic models, we can one day add-in a further reflexive dimension where culturally specific models provide views of culturally specific views of other cultures. Perspectives all the way down.

7. Data Availability

Data and code have been archived and are persistently available at: https://doi.org/10.5683/SP3/0EYY0T.

8. Author Contributions

Andrew Piper: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

References

Abdulhai, Marwa, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques (2023). "Moral Foundations of Large Language Models". In: *arXiv preprint*. 10.48550/arXiv.2310.15337.

Anantharama, Nandini, 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*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, 3551–3564. 10.18653/v1/2022.findings-emnlp.260.

Bamman, David (2024). *Wikiplots*. http://yosemite.ischool.berkeley.edu/david/wikiplots.txt (visited on 10/22/2025).

Berns, Gregory (2022). *The Self Delusion: The New Neuroscience of how we Invent – and Reinvent – our Identities*. Basic Books.

Booth, Wayne C. (1998). "Why Ethical Criticism can Never be Simple". In: *Style* 32 (2), 351–364. https://www.jstor.org/stable/42946431 (visited on 10/22/2025).

Brewer, William F. and Edward H. Lichtenstein (1980). *Event Schemas, Story Schemas, and Story Grammars*. Tech. rep. No. 197. Center for the Study of Reading, University of Illinois.

Campbell, Joseph [1949] (2008). *The Hero with a Thousand Faces*. New World Library.

Chambers, Nathanael 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*. Association for Computational Linguistics, 602–610.

Cheah, Pheng (2015). *What is a World?: On Postcolonial Literature as World Literature*. Duke University Press. 10.1215/9780822374534.

Dai, Zeyu 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*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 3332–3342. 10.18653/v1/2021.findings-acl.295.

Dundes, Alan (1962). "From Etic to Emic Units in the Structural Study of Folktales". In: *The Journal of American Folklore* 75 (296), 95–105. 10.2307/538171.

Fischer, Frank, Jacob Blakesley, Paula Wojcik, and Robert Jäschke (2023). "Preface: World Literature in an Expanding Digital Space". In: *Journal of Cultural Analytics* 8 (2). 10.22148/001c.74598.

Frermann, Lea, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak (2023). "Conflicts, Villains, Resolutions: Towards Models of Narrative Media Framing". In: *arXiv preprint*. 10.48550/arXiv.2306.02052.

Frye, Northrop [1957] (2020). *Anatomy of Criticism: Four Essays*. Princeton University Press.

Fudolig, Mikaela Irene, Thayer Alshaabi, Kathryn Cramer, Christopher M. Danforth, and Peter Sheridan Dodds (2023). "A Decomposition of Book Structure Through Ousiometric Fluctuations in Cumulative Word-time". In: *Humanities and Social Sciences Communications* 10 (187). 10.1057/s41599-023-01680-4.

Genette, Gérard (1992). *The Architext: An Introduction*. University of California Press.

Graham, Jesse, 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". In: *Advances in Experimental Social Psychology*. Ed. by Patricia Devine and Ashby Plant. Vol. 47. Academic Press, 55–30. 10.1016/B978-0-12-407236-7.00002-4.

Gregory, Marshall W. (2010). "Redefining Ethical Criticism. The Old vs. the New". In: *Journal of Literary Theory* 4 (2), 273–301. 10.1515/jlt.2010.017.

Hatzel, Hans Ole and Chris Biemann (2024). "Tell Me Again! A Large-Scale Dataset of Multiple Summaries for the Same Story". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. ELRA and ICCL, 15732–15741. https://aclanthology.org/2024.lrec-main.1366/ (visited on 10/22/2025).

Hobson, David, Haiqi Zhou, Derek Ruths, and Andrew Piper (2024). "Story Morals: Surfacing Value-driven Narrative Schemas Using Large Language Models". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Association for Computational Linguistics, 12998–13032. 10.18653/v1/2024.emnlp-main.723.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. (2023). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 43 (2). 10.1145/3703155.

Karsdorp, F. B. and A. van den Bosch (2013). "Identifying Motifs in Folktales Using Topic Models". In: *Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning*, 41–49. https://benelearn2013.org/pdfs/paper_28.pdf (visited on 10/22/2025).

Kukkonen, Karin (2014). "Plot". In: *The Living Handbook of Narratology*. Ed. by Peter Hühn, Jan-Christoph Meister, John Pier, and Wolf Schmid. Hamburg University. https://www-archiv.fdm.uni-hamburg.de/lhn/node/115.html (visited on 10/22/2025).

Kundalia, Kaushil, Yash Patel, and Manan Shah (2020). "Multi-label Movie Genre Detection from a Movie Poster Using Knowledge Transfer Learning". In: *Augmented Human Research* 5, 1–9. 10.1007/s41133-019-0029-y.

Liscio, Enrico, Alin E Dondera, Andrei Geadau, Catholijn M. Jonker, and Pradeep K. Murukannaiah (2022). "Cross-domain Classification of Moral Values". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Association for Computational Linguistics, 2727–2745. 10.18653/v1/2022.findings-naacl.209.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp (2022). "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Association for Computational Linguistics, 8086–8098. 10.18653/v1/2022.acl-long.556.

Nussbaum, Martha Craven (1998). "Exactly and Responsibly: A Defense of Ethical Criticism". In: *Philosophy and Literature* 22 (2), 343–365. 10.1353/phl.1998.0047.

Ouyang, Jessica 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*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Association for Computational Linguistics, 2149–2158. 10.18653/v1/D15-1257.

Pham, Chau, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer (2024). "TopicGPT: A Prompt-based Topic Modeling Framework". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies (*Volume 1: Long Papers*). Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Association for Computational Linguistics, 2956–2984. 10.18653/v1/2024.naacl-long.164.

Piper, Andrew (2015). "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel". In: *New Literary History* 46 (1), 63–98. https://www.jstor.org/stable/24542659 (visited on 10/22/2025).

Piper, Andrew and Sophie Wu (2025). "Evaluating Large Language Models for Narrative Topic Labeling". In: *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. Ed. by Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar. Association for Computational Linguistics, 281–291. 10.18653/v1/2025.nlp4dh-1.25.

Propp, Vladimir [1928] (1968). *Morphology of the Folktale*. University of Texas Press.

Rashkin, Hannah, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao (2020). "PlotMachines: Outline-conditioned Generation with Dynamic Plot State Tracking". In: *arXiv preprint*. 10.48550/arXiv.2004.14967.

Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds (2016). "The Emotional Arcs of Stories are Dominated by six Basic Shapes". In: *EPJ Data Science* 5 (1), 1–12. 10.1140/epjds/s13688-016-0093-1.

Reynolds, Laria and Kyle McDonell (2021). "Prompt Programming for Large Language Models: Beyond the Few-shot Paradigm". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Ed. by Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, and Takeo Igarashi. Association for Computing Machinery, 1–7. 10.1145/3411763.3451760.

Rezapour, Rezvaneh, 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*. Ed. by Eric Gilbert and Karrie Karahalios. Association for Computing Machinery, 347–351. 10.1145/3311957.3359496.

Riedl, Mark (2024). *Wikiplots*. https://github.com/markriedl/WikiPlots (visited on 10/22/2025).

Roy, Shamik 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*. Ed. by Lun-Wei Ku and Cheng-Te Li. Association for Computational Linguistics, 1–13. 10.18653/v1/2021.socialnlp-1.1.

Roy, Shamik, Nishanth Sridhar Nakshatri, and Dan Goldwasser (2023). "Towards Few-shot Identification of Morality Frames Using In-context Learning". In: *arXiv preprint*. 10.48550/arXiv.2302.02029.

Russell, Robert L and Paul Van Den Broek (1992). "Changing Narrative Schemas in Psychotherapy". In: *Psychotherapy: Theory, Research, Practice, Training* 29 (3), 344–354. 10.1037/h0088536.

Scherrer, Nino, Claudia Shi, Amir Feder, and David Blei (2023). "Evaluating the Moral Beliefs Encoded in LLMs". In: *Advances in Neural Information Processing Systems*. 36, 51778–51809.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr (2023). "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned

to start worrying about prompt formatting". In: *arXiv preprint*. 10.48550/arXiv.2310.11324.

Sims, Matthew, Jong Ho Park, and David Bamman (2019). "Literary Event Detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Association for Computational Linguistics, 3623–3634. 10.18653/v1/P19-1353.

Subbiah, Melanie, Faisal Ladhak, Akankshya Mishra, Griffin Adams, Lydia Chilton, and Kathleen Mckeown (2024). "STORYSUMM: Evaluating Faithfulness in Story Summarization". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Association for Computational Linguistics, 9988–10005. 10.18653/v1/2024.emnlp-main.557.

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

Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Vauth, Michael, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann (2021). "Automated Event Annotation in Literary Texts". In: *Proceedings of the Computational Humanities Research Conference 2021*. Ed. by Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert, 333–345.

Vida, Karina, Judith Simon, and Anne Lauscher (2023). "Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research". In: *arXiv preprint*. 10.48550/arXiv.2310.13915.

Webson, Albert and Ellie Pavlick (2022). "Do Prompt-based Models Really Understand the Meaning of their Prompts?" In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Association for Computational Linguistics, 2300–2344. 10.18653/v1/2022.naacl-main.167.

Wikipedia Contributors (2024). *Wikipedians*. https://en.wikipedia.org/wiki/Wikipedia:Wikipedians (visited on 10/22/2024).

Wilkens, Matthew (2016). "Genre, Computation, and the Varieties of Twentieth-century US Fiction". In: *Journal of Cultural Analytics* 2 (2). 10.22148/16.009.

Wu, Winston, 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*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Association for Computational Linguistics, 5113–5125. 10.18653/v1/2023.emnlp-main.311.

Yan, Zhihua and Xijin Tang (2023). "Narrative Graph: Telling Evolving Stories Based on Event-centric Temporal Knowledge Graph". In: *Journal of Systems Science and Systems Engineering* 32 (2), 206–221. 10.1007/s11518-023-5561-0.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen Mckeown, and Tatsunori B Hashimoto (2024). "Benchmarking Large Language Models for News Summarization". In: *Transactions of the Association for Computational Linguistics* 11, 39–57. 10.1162/tacl_a_00632.

Zhou, Haiqi, 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)*. Ed. by Dominik Stammbach, Jingwei Ni, Tobias Schimanski, Kalyan Dutia, Alok Singh, Julia Bingler, Christophe Christiaen, Neetu Kushwaha, Veruska Muccione, Saeid A. Vaghefi, and Markus Leippold. Association for Computational Linguistics, 143–155. 10.18653/v1/2024.climatenlp-1.11.