# **Evaluating Large Language Models for Narrative Topic Labeling**

Andrew Piper Sophie Wu Department of Languages, Literatures, and Cultures McGill University, Montreal, Canada

#### Abstract

This paper evaluates the effectiveness of large language models (LLMs) for labeling topics in narrative texts, comparing performance across fiction and news genres. Building on prior studies in factual documents, we extend the evaluation to narrative contexts where story content is central. Using a ranked voting system with 200 crowdworkers, we assess participants' preferences of topic labels by comparing multiple LLM outputs with human annotations. Our findings indicate minimal inter-model variation, with LLMs performing on par with human readers in news and outperforming humans in fiction. We conclude with a case study using a set of 25,000 narrative passages from novels illustrating the analytical value of LLM topic labels compared to traditional methods. The results highlight the significant promise of LLMs for topic labeling of narrative texts.

### 1 Introduction

Topic modeling has been and continues to be one of the most popular ways of interpreting and understanding documents within large digital repositories. Whether for the purposes of discourse analysis (Jacobs and Tschötschel, 2019), literary studies (Jockers and Mimno, 2013; Uglanova et al., 2020), media framing (Ylä-Anttila et al., 2022), or understanding semantic change (Hall et al., 2008; McFarland et al., 2013), successfully extracting high-level topics has been central to the digital humanities and the large scale study of history and culture (for a review see Alghamdi and Alfalqi (2015)).

Until recently, the principal way that researchers have derived topics from texts has been through the use of unsupervised learning approaches such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its various updates (Blei and Lafferty, 2006; Boyd-Graber and Blei, 2012; Roberts et al., 2013; Thompson and Mimno, 2018).

These methods, however, face well-known limitations, ranging from the ambiguity of topic labels, to their sensitivity to parameter choices (most notably the number of topics), and the oversimplification of textual content through the use of bag-of-words modeling.

Recent work has begun to show how LLMs can potentially enhance or even replace traditional topic modeling. LLMs have been used to facilitate topic labeling (Rijcken et al., 2023) and topic evaluation (Stammbach et al., 2023). And they have been used in lieu of topic modeling, demonstrating far greater semantic alignment with known human labels on sets of fact-based articles (Pham et al., 2024) and expert judgments (Lam et al., 2024).

In this paper, we build on prior research by applying large language models (LLMs) to automated topic labeling, with a focus on narrative texts as a complement to studies centered on fact-based documents. Narrative texts, a cornerstone of cultural traditions, have long been a subject of interest in digital humanities research due to their complexity and richness. Unlike information-driven texts, narratives often depend on implicit context, figurative language, shifting perspectives, and intricate temporal structures, all of which pose unique challenges for topic extraction. By evaluating LLM performance on the automated topic labeling of narratives-both fictional and factual-this study aims to enhance the methodological tools available to digital humanities researchers. To this end, we analyze two distinct narrative sub-genres: factual reporting in news articles and creative storytelling in novels.

Second, while previous research has focused on the comparative similarity between automated and human-generated labels (demonstrating that LLMs significantly outperform LDA (Pham et al., 2024)), our study evaluates the preference for LLMgenerated labels over human labels. Following a methodology similar to Lam et al. (2024), we use a crowd-sourced voting approach to determine whether independent readers (N=200) find LLMgenerated labels equal to or more favorable than human-generated ones. This methodology not only provides a robust evaluation of label quality but also offers a practical measure of how well LLMs meet the expectations of general readers. Our question is: Can LLMs label narrative topics as effectively as humans across different sub-genres, and how do they compare to well-established topic modeling techniques?

Finally, while prior studies have primarily focused on the functionality of a single model (e.g., GPT), we broaden the scope by evaluating GPT alongside a range of smaller, open-weight models. This comparative analysis aims to provide researchers with greater confidence in the utility of LLMs for topic labeling in narrative texts. To support future research and benchmarking, we publicly release all annotations generated in this study.<sup>1</sup>

## 2 Prior Work

Topic modeling has experienced wide-spread use across numerous fields (Alghamdi and Alfalqi, 2015). Despite its ubiquity, considerable research has foregrounded its methodological limitations. Traditional topic models often produce topics that are statistically coherent, for example, but lack semantic interpretability, making it challenging for human analysts to derive meaningful insights (Chang et al., 2009; Mimno et al., 2011). They also involve numerous pre-processing steps that increase researcher degrees-of-freedom that can impact replicability (Hecking and Leydesdorff, 2019; Mantyla et al., 2018).

Additionally, determining the optimal number of topics is often a trial-and-error process, potentially leading to over- or under-fitting of the model (Wallach et al., 2009). This problem can also lead to challenges in modulating the specificity or generality of topics (Rijcken et al., 2024). Finally, these methods can perform poorly on short texts or documents with diverse vocabulary, limiting their applicability in certain domains such as social media analysis or highly specialized technical literature (Hong and Davison, 2010).

Recent work has begun to use LLMs in conjunction with topic modeling, either to label (Rijcken et al., 2023) or evaluate topics (Stammbach et al., 2023). Pham et al. (2024) have devised a prompting framework for the generation and selection of topics using GPT-4 and shown significant improvement over LDA with respect to human labels for fact-based documents such as Wikipedia articles and U.S. Congressional bills. Lam et al. (2024) have developed a workflow that they call "concept induction" to replace topic modeling to surface more critical and research-oriented conceptual frameworks for the analysis of fact-based documents.

Here we build on this prior work to apply LLMderived topic labeling to narrative texts and assess label adequacy based on independent human assessments.

#### 3 Methods

Our experimental framework consists of two main In the first, we evaluate LLMcomponents. generated topic labels against human-generated labels using a survey platform with anonymous readers. Given prior findings on the significant superiority of LLM topics over those generated by traditional topic modeling methods such as LDA (Pham et al., 2024), we exclude LDA-based topics from this stage and focus instead on assessing the ability of LLMs to match or exceed human performance. In our case study (Section 5), we shift our focus to a large sample of fiction passages, comparing LLM-derived topics directly with LDAgenerated topics. This comparison allows us to more explicitly examine the analytical advantages and limitations of LLM-derived topics relative to traditional approaches.

## 3.1 Data

We evaluate topic labeling across two narrative genres that span the fact/fiction divide. For the fiction dataset, we use a curated collection of approximately 700 open-access novels published in the nineteenth century, provided by Chadwyck-Healey. To accommodate the topic modeling process and handle long documents, we divide the novels into 500-word chunks. For the fact-based dataset, we utilize 6,722 news articles from the Global News Dataset, representing four publications from diverse geographic regions: ABC News, Al Jazeera English, BBC News, and The Times of India.<sup>2</sup> Given the average article length of 666 words, we use the full article in our analysis. For our annotation task, we sample 50 passages/articles per dataset. For our case study, we sample 25,000 passages from the novel data.

<sup>&</sup>lt;sup>1</sup>https://doi.org/10.5683/SP3/MHJRIO

<sup>&</sup>lt;sup>2</sup>https://www.kaggle.com/datasets/everydaycodings/globalnews-dataset

#### 3.2 LLM Prompting

We compare four different LLMs for our experiment: GPT-40, Llama3:8B, Llama3.1:8B, and Gemma2:9B. To generate our LLM outputs for each model and category, we use a zero-shot prompting framework. Here is an excerpt of the full prompt:

> What are the central topics of the following passage? Provide your answer as a list of keywords separated by commas. Start from the most general topic and get increasingly specific. Give three total topics.

Note that we ask for topics of descending generality to address the problem of topic scale. We also find that introducing any pre-processing of the passages, such as distillation or summarization, resulted in poorer model responses. Due to the high cost of surveys, we only test our zero-shot approach compared to human annotations.

### 3.3 Human Labels

For the human labeling step, we hired two undergraduate student annotators with backgrounds in the humanities. To guide their understanding of "topics" we provided students with a conversation transcript with chatGPT discussing the difference between topics and themes, which can be found along with the codebook in the online repository. Here is a brief excerpt:

> A topic is the specific subject matter or main focus of a piece of writing. It answers the question, "What is this about?" Topics are explicit, straightforward, and usually stated clearly within the text. They deal with facts, events, and specific issues.

Students annotated 100 passages in total, split evenly by genre, providing three labels per passage.

## 3.4 LDA Labels

For our LDA topics, we run LDA with Gibbs sampling over the entire collection of 175,337 novel chunks using k=20 and k=60, with an alpha parameter of 1/k, to capture two different topic size parameters. Sample topics are shown in Table 1. In order to assign topics to documents, we keep the top three most probable topics for a given passage to align with our LLM-output of three topics per passage.

## 3.5 LLM Topic Aggregation

For our case study, we randomly select 25,000 passages from the total pool of novel chunks and process them using Gemma2:9b with the topic labeling prompt described earlier.

A key challenge with LLM-generated topics is the sheer number of distinct topics produced. For instance, in our Gemma2-generated labels, we identify 3,411 unique labels that appear more than once. To address this long-tail distribution, we introduce an intermediate step of topic aggregation, reducing the labels to a smaller set of more general topics. By presenting the results of both the granular and aggregated outputs, we allow researchers to assess which approach best aligns with their specific research objectives.

For topic aggregation, we begin by supplying all topic labels that appear more than once (N=3,411) to the reasoning model, GPT-4o1. The model first resolves overlapping topics (e.g., 'naval warfare' and 'warfare') and then consolidates synonymous topics into higher-level categories (e.g., 'farming,' 'harvest,' and 'agriculture' are grouped under 'agriculture'). This process yields a total of 922 aggregated topic labels. Next, we map the complete set of Gemma2 labels onto these 922 topics using the GloVe 6B 100-dimensional Wikipedia word embedding model (Pennington et al., 2014). For each original Gemma2 label, we identify the candidate aggregate label with the lowest cosine similarity and assign the corresponding aggregate label.

## **4** Validation Results

#### 4.1 Quantitative Validation

We validate our LLMs' performance by utilizing a ranked voting survey through the popular crowdsourcing platform Prolific. We recruited 200 participants in our survey who were presented with the following: a text passage (news or novel); a brief definition of a topic; and five possible answers, which included four LLM outputs and one human annotation. Each passage was judged by two indepedent survey participants. Figure 2 in the Appendix illustrates a screenshot of the survey. The order of the labels from the different sources (models and humans) was randomized for each survey participant.

Because both models and the annotators were initially instructed to provide three answers per topic in descending order of generality, we selected only one of these answers for each passage in our

k	Topic	Topic Words
20	Seeming	seemed, appeared, moment, length, soon, stranger, passed, appearance, though
20	Philosophy	nature, character, life, world, society, common, country, often, human
20	Daily Rhythm	day, night, morning, long, away, home, hour, evening, gone
60	Connectors	course, nothing, quite, though, done, perhaps, matter, almost, also
60	Looking	looked, back, hand, looking, face, turned, head, look, eyes
60	Feelings	mind, heart, feelings, hope, melancholy, almost, tears, length, grief

Table 1: Top words associated with LDA topics used in Figure 1

survey, where the rank of the answer was preserved across models. For example, if we selected the first answer from one model, then we selected the first answer from all other models, including the human annotators, for that passage. We over-sampled the first rank by a factor of two to privilege the most general answer, while second and third levels were weighted evenly.

Survey participants were then instructed, "Please rank these labels from best to worst (1 being best, 5 being worst) in order of preference." If some outputs were identical (i.e. models outputted the same answer), participants were told to group these together, but in any order. We required participants to be fluent in English and only allowed participants to answer one passage. Where outputs by our models were identical, we normalized participants' ranks to match the lowest ranking answer of that kind (thus if one of three identical answers was a 2 then all identical answers to that one were given a 2).

To assess the degree of disagreement among participants' ranking, we calculated the median / mean deviation between the rank of each model for each pair of survey participants responding to the same passage. The median deviation among participants was 1 with greater than 80% of rankings within two or fewer ranks. This suggests a high degree of alignment between the ordering of models by different survey participants who were most often only 1 rank apart in the order they assigned to different models.

As can be seen in Table 2, we found that for the fiction sample Gemma2 performed best and the human answers worst. For news, GPT40 performed best and Llama worst. In order to test for statistical significance among the ranking preferences between models, we performed a pairwise Wilcoxon rank-sum test with Bonferroni correction for all model pairs, including humans. We found that the only pairs that indicated statistically significant rank differences at p < 0.05 were Human-Gemma2 and Human-GPT-40 for the fiction data. There were no statistically significant differences between models for news rankings.

Model	Fiction	News	
Gemma2	2.25	2.81	
GPT_40	2.57	2.40	
Llama3.1	2.68	<u>2.84</u>	
Llama3	2.83	<u>2.84</u>	
HUM	<u>3.23</u>	2.79	

 Table 2: Average ranks of all models by genre. Bold indicates best, underline indicates worst.

#### 4.2 Qualitative Assessment

For our qualitative assessment, we provide two sample views of model outputs. The first is Table 3, which shows a list of human labels alongside the most preferred LLM label. The second (Table 4) provides summaries of sample passages with all topic labels from each model included for both fiction and news with the preferred label in bold.

In terms of survey respondent preferences, as can be seen in Table 4 we find that for news labels they generally preferred more specific labels. For example, between *real estate* and *real estate investment* readers preferred the latter or between *prostate cancer* and *health awareness* they preferred the former.

For news, our human annotators generally, though not always, provided more general labels than our models (Table 4). This was especially true in cases where the article centred around a particular celebrity (Jared Leto or Draymond Green). Depending on researcher goals this preference for specificity as it relates to news topics should be considered when applying LLMs to this task.

For the novel topics, we found that it often worked in reverse as far as survey respondents were concerned, though less clearly. For example, *urban life* was preferred over *London* while *household* 

Genre	Human	LLM
FIC	war	warriors
FIC	rivalry	respect
FIC	physical appearance	characteristics
FIC	sibling relationship	family
FIC	territory	nature
FIC	social transgression	society
FIC	survival	honor
FIC	appearance	instructions
FIC	faith	religion
FIC	marriage	social pressure
NEWS	protest	human rights
NEWS	cricket	cricket
NEWS	genetic research	genomics
NEWS	international relations	us-china relations
NEWS	health awareness	prostate cancer
NEWS	family	memorial
NEWS	us politics	us politics
NEWS	cricket	cricket world cup
NEWS	war	israel
NEWS	israel-palestine conflict	hamas attack

Table 3: Examples of human and LLM topics from a subset of passages. Bold indicates instances where the human answer was preferred, otherwise the LLM label was preferred.

was preferred over *mystery*. Here too general differences between human and LLM annotations are harder to classify. While in some cases LLM annotations appear more general (nature v. territory, family v. sibling relationship, society v. social transgression), in others the distinctions are less clear (respect v. rivalry, honor v. survival).

Despite these differences, overall we find a high degree of similarity between the labeling tendencies of our human annotators and our models. For example, we found that human annotations matched at least one model output in 50% of cases for our news data and 72% of cases for the novel data. When comparing model outputs to each other, we found that for 98% and 94% of our passages respectively at least two models generated identical outputs. This resulted in an overall matching rate of 40% across all possible LLM-generated outputs. Note this is only for exact matches, which under-estimates answers that have high semantic similarity but slight lexical differences. The overall cross-annotation similarity is also supported by our participant survey data which showed minimal statistical difference in terms of participant preferences. Models of different sizes appear to match human-level labeling capabilities for both types of narrative texts tested.

### 5 Case Study

We conclude with a case study to indicate some of the conceptual insights that can be offered by LLM-assisted topic labeling compared with traditional LDA-based topic models. Here we condition on our novel data to illustrate the most distinctive topics of the first and second half of the nineteenth-century, often referred to as the heyday of the British realist novel.

For our experiment we use the above-mentioned sample of 25,000 novel chunks and label them two ways. For LLM-assisted labeling we use Gemma2:9b with the same prompt used for our human validation experiment. We retain two sets of labels: all original labels and the aggregated labels using the method described above. Next, we applied Latent Dirichlet Allocation (LDA) using Gibbs sampling. We set the Dirichlet prior for documenttopic distributions to  $\alpha$ =50/k, a commonly used heuristic that ensures a moderate spread of topics per document, and estimated  $\beta$  during training. The model ran for 1000 iterations with a burn-in of 20, retaining the best solution (best=TRUE). We tested two levels of k=20/60. Topic labels were then manually added by the authors as domain experts.

After labeling, we identify the most distinctive topics in passages published before and after 1850 to model large-scale shifts in topical focus within British novels. To measure distinctiveness, we use Dunning's log-likelihood statistic, a method that highlights words or topics disproportionately represented in one group compared to another based on their observed versus expected frequencies (Dunning, 1994). Figure 1 presents the most distinctive topic labels across four conditions: the specific Gemma2 labels, the aggregated Gemma2 labels, and the two k settings for our LDA models.

Overall, we observe that LLM labeling produces significantly more intelligible topics. Where several of the top topics in the LDA models are largely grammatical distinctions that transpire over the course of the century (e.g. the introduction of contractions to capture direct speech) or clusters of common verbs (such as looking or taking), LLMs produce more detailed and informative topics. "Combat," "revenge," "travel," and "revolution" in the general model tell us considerably more about the genres distinctive of the pre-1850 Romantic and post-Romantic periods in British novelwriting than topics like "seemed," "conduct," "war," "philosophy," and "religion." Similarly with the

	Passage	Human	Gemma2	GPT40	Llama3	Llama3.1
NEWS	The US military has begun buying Japanese seafood to support the industry amid China's import ban over treated Fukushima water, while tensions between the US and China continue over economic and diplo- matic issues.	international relations	international relations	us military	us-china relations	trade
NEWS	Sports presenter Steve Rider, recently diagnosed with prostate cancer, urges men to get early check-ups, sharing his own experience of catching the disease in time for curative surgery and raising awareness about its risks and symptoms.	health awareness	prostate cancer	health	health	prostate cancer
NEWS	AI-generated deepfake videos of Rashmika Man- danna and Katrina Kaif have raised concerns about the misuse of deepfake technology, prompting calls for stricter identification methods.	artificial intelli- gence	deepfakes	technology	technology	misinformation
FIC	A man gazes upon a breathtaking panorama of hills, mountains, and rivers, but his thoughts are consumed by the encroachment of white settlements, which he perceives as a tightening serpent symbolizing the inevitable displacement and doom of his people.	territory	scenery	nature	nature	civilization
FIC	Arriving in bustling London, Philip is overwhelmed by the city's impersonal crowds but finds comfort in a kind innkeeper's hospitality, renewing his resolve to pursue the work that brought him there.	urban life	urban life	london	world	traveler
FIC	At Thornfield, Jane overhears hints of a mysterious secret as preparations for an important event bring the estate to a polished splendor, while she remains in the quiet refuge of the schoolroom, awaiting the arrival of Mr. Rochester's anticipated guests.	mystery	social dynamics	mystery	household	general

Table 4: Sample topics for each model for selected passages. GPT-generated summaries are provided for each passage. Bold indicates survey participant preference.

more specific models, "faith," "slavery," "marriage," and "civil war" are far better than "school," "daily rhythms" or "communication."

To be sure, it is not the case that LDA cannot inform researchers of broad trends in fictional narratives. The emphasis on dialogue, children, and perception are all notable dimensions of post-1850 novels. Additionally, as we mention in the discussion section, there is much more testing one could do to optimize the LDA workflow to improve the labelinng procedure. The value of LLM-based labeling, however, lies first in the *topicality* of the topic labels–dialogue, perception and children all capture very different kinds of stylistic features for example, while faith, finance, and marriage are far closer to what readers understand as narrative "topics."

Second, as has been widely observed LDA topics pose challenges of interpretation for readers leading to difficulties with consistency in topic labeling. While we did not experiment with this problem here, one of the challenges of LDA labeling is the labeling step itself. Third, LLM-derived topics also capture more thematic diversity than LDA methods without introducing the noise of unintelligible topics. Table 5 presents a more extended list of distinctive topics k=60 and Gemma (General) models. For example, we see far more nuance in the range of topics even in the general Gemma model, such as conspiracy, justice, strategy, diplomacy, etc. compared to LDA topics like discover, exclamation, or seafaring. These more nuanced concepts allow researchers to test broader more detailed theories about thematic changes over long stretches of literary history.

## 6 Discussion

The results of this study highlight the promise and limitations of using large language models (LLMs) for narrative topic labeling, particularly when evaluated across distinct genres like fiction and news. While prior work has largely focused on the application of LLMs for fact-based or general documents, our findings extend this understanding to narrative texts, showcasing the strengths and weaknesses of these models in a storytelling context.

One of the key findings of this study is the comparable performance of large language models (LLMs) to human annotators in narrative topic labeling. Our analysis revealed that LLMs effectively



Figure 1: Top five most distinctive topics for each period using all models.

generated coherent and contextually appropriate labels for both fiction and news genres. For example, in the fiction dataset, Gemma2 provided labels such as "family relationships" and "urban life," which aligned well with human annotations of similar passages. Similarly, in the news dataset, LLMgenerated labels like "US-China relations" and "prostate cancer" closely matched human-provided labels. Importantly, we found that pre-processing or intermediate steps were not necessary; a direct, zero-shot prompting approach performed on par with human annotations, streamlining the process without compromising quality.

One of the key challenges we encountered is the long-tail distribution of LLM-generated labels. The sheer diversity of labels produced by the models often led to an overwhelming number of distinct topics, many of which were semantically similar or redundant. To address this, we implemented a reasoning model to aggregate these labels into a more manageable set of general topics. While this approach reduced redundancy and improved interpretability, it introduced its own limitations, such as potential errors through aggregation. Other methods, such as clustering techniques or alternative aggregation strategies, may be more effective and warrant further exploration to refine the process of topic consolidation.

Another challenge lies in our evaluation framework. For the human validation component although the ranked voting survey provided valuable insights into label preferences, it also introduced potential biases, such as the influence of phrasing or vocabulary on participant choices. Additionally, our evaluation relied on the subjective preferences of general readers, which may not fully capture the utility of the labels for specific research applications. Expanding the evaluation to include task-specific downstream applications or expert assessments could provide a more comprehensive understanding of LLM performance and its alignment with user needs.

For the LDA comparison, our case study only scratched the surface of LDA optimization suggesting that future could more exhaustively test LLM v. LDA exercises, especially given the far greater computational resources necessary for LLM-assisted labeling. The models used in this study, particularly larger ones like Gemma2:9B, require substantial computational power and finan-

1800-1850		1850-1900		
Gemma2	LDA	Gemma2	LDA	
combat revenge escape travel revolution canada battle conspiracy punishment strategy america history captivity folklore novel culture betrayal justice ownership diplomacy conflict romance	feelings conduct danger return though party appearance family seafaring Nat.American discover exclamation approaching violence mystery battle religion politics philosophy nature death	relationship family marriage finance faith childhood change clergy mystery scandal love religion business horse	contractions adverbs looking dialogue school sickness marriage letters faces feelings time of day animals home colors reading remember summer village take numbers sleep eating	

Table 5: Most distinctive topics for each model by half-century. For LDA we use the k=60 condition and Gemma (General).

cial resources for both inference and aggregation tasks. These constraints can make the application of LLMs for massive labeling tasks of hundreds of thousands of passages far more restrictive. Potential solutions include leveraging smaller, finetuned models, optimizing inference processes, or exploring hybrid approaches that combine LLMs with more traditional methods to reduce resource demands.

While LLM-assisted labeling demonstrates clear advantages in interpretive depth, traditional approaches like LDA still hold value, particularly as tools for dimensionality reduction. LDA's ability to cluster and summarize large textual datasets efficiently provides complementary insights that are less focused on interpretive richness but valuable for structuring data. In contrast, LLM-based labeling excels in producing semantically rich and contextually specific labels, making it more suitable for applications where interpretive depth is prioritized. The choice between methods should depend on the specific goals and constraints of the research project.

Our case study demonstrated the thematic richness that LLM-assisted labels can bring to largescale cultural research. By analyzing shifts in topical focus within British novels across the 19th century, we showed how LLMs could generate insightful and historically significant insights, such as emerging attention to "civil war" and "slavery" in the later nineteenth century and a receding attention to topics related to "native-american culture" and "land ownership." This capability highlights the potential of LLM-assisted labeling to validate and discover new dimensions of understanding in literary and cultural studies, offering researchers a powerful tool for examining thematic evolution across time and genres.

## 7 Conclusion

This study underscores the transformative role large language models (LLMs) can play in narrative topic labeling, particularly in capturing the semantic richness and thematic complexity of both fiction and news texts. By performing on par or above human annotators across numerous passages, LLMs demonstrate their ability to produce labels that resonate with general readers while maintaining consistency across genres. Importantly, this capability not only streamlines the annotation process but also opens new possibilities for scalable and nuanced narrative analysis, particularly in contexts where traditional methods such as LDA struggle with interpretive specificity.

Our results also highlight the unique contributions of LLMs to narrative understanding beyond their technical accuracy. Unlike earlier methods, LLMs offer the ability to identify subtle thematic patterns and connect these to broader cultural or historical narratives. This ability to balance specificity with breadth positions LLMs as powerful tools for both academic research and applied settings in journalism, literature, and cultural studies.

While challenges such as label aggregation and computational costs remain, this study demonstrates the promise of LLMs as a paradigm shift in narrative topic labeling. Their ability to go beyond clustering and surface themes that align with human intuition makes them invaluable for complex narrative analysis.

### Limitations

While we compare four different open-weight and one frontier model to human answers, our results are not generalizable to all language models. Similarly, while we test two kinds of narrative genres it is possible that different genres might yield different results. The lower preference for human answers on the fiction task may also be a reflection of the quality of the human answers or, conversely, biases of the survey participants. Thus a different set of human respondents may yield more competitive human answers. Nevertheless, we believe the research here supports the assertion that LLMs are at least on par with highly educated human readers. While our survey included 200 unique responses, it is possible that with a larger sample of text passages we might observe more/less differentiation among models than in our study.

We also note limitations around our topic aggregation approach. Future work will want to explore this area as its own problem domain. One of the intrinsic challenges of topic labeling is the issue of scale, that there are different appropriate answers at different levels of generality.

### Acknowledgments

We wish to thank the Social Sciences and Humanities Research Council of Canada (435-2022-089) for funding to support this research.

#### References

- Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1).
- David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Jordan Boyd-Graber and David Blei. 2012. Multilingual topic models for unaligned text. *arXiv preprint arXiv:1205.2657*.
- Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.
- Ted Dunning. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.
- David Hall, Dan Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 363–371.
- Tobias Hecking and Loet Leydesdorff. 2019. Can topic models be used in research evaluations? reproducibility, validity, and reliability when compared with semantic maps. *Research Evaluation*, 28(3):263–272.

- Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.
- Thomas Jacobs and Robin Tschötschel. 2019. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, 22(5):469–485.
- Matthew L Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.
- Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with highlevel concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 26 of *CHI '24*, page 1–28. ACM.
- Mika V Mantyla, Maelick Claes, and Umar Farooq. 2018. Measuring lda topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement*, pages 1–4.
- Daniel A McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D Manning, and Daniel Jurafsky. 2013. Differentiating language usage through topic models. *Poetics*, 41(6):607–625.
- David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (*EMNLP*), pages 1532–1543.
- Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A promptbased 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)*, pages 2956–2984.
- Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. 2023. Towards interpreting topic models with chatgpt. In *The 20th World Congress of the International Fuzzy Systems Association.*
- Emil Rijcken, Kalliopi Zervanou, Pablo Mosteiro, Floortje Scheepers, Marco Spruit, and Uzay Kaymak. 2024. Topic specificity: A descriptive metric for algorithm selection and finding the right number of topics. *Natural Language Processing Journal*, page 100082.

- Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.
- Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357.
- Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914.
- Inna Uglanova, Evelyn Gius, F Karsdorp, B McGillivray, A Nerghes, and M Wevers. 2020. The order of things. a study on topic modelling of literary texts. *CHR*, (18-20):2020.
- Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112.
- Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2022. Topic modeling for frame analysis: A study of media debates on climate change in india and usa. *Global Media and Communication*, 18(1):91–112.

# 8 Appendix

See next page.

#### Please read the following passage carefully before answering the questions.

Thrush will go with the White Eagle," said the maiden, "and sing him to sleep, so that he shall not harm his red brother." "The Brown Thrush must go with Thayendanegea," said the chief. "No!" said his aunt, speaking for the first time, although she had been an attentive listener. "My sister's daughter now has no mother but me. My sister is dead. My sister's son, listen to my command. The Thrush shall not go with you." "My mother's sister, my ears are open. What you have said has entered them, and you must be obeyed." "My sister's son," she continued, with deliberation, "the Brown Thrush shall go with the White Eagle." "You command it. It must be so. But whither will they go? You cannot command the three thousand warriors whose chiefs have decided that my white brother shall not return to the pale-faces until the war is ended." "False, treacherous, perfidious Thayendanegea!" said Charles. "And this is the cowardly work of the one I have loved and trusted! No more my brother! Henceforth we are foes!" "My brother, do not make my blood boil over. Another had died ere the speech were finished. Thayendanegea did nothing. He knew it not until the chiefs had decided. He did not approve it, but he could not oppose it. He loves his brother still. He waits to hear his brother's next words." "Forgive me, my brother!" said Charles, with tears in his eyes. "I ask my brother's pardon." "It was the Malcha Manito, and not my brother. But what can my brother do? The warriors surrounding him, who will not declare war against his white brothers, will not oppose the decree of the chiefs. They are not ready to fight their red brothers." "I will escape. You know the White Eagle can soar above his enemies." "But whither will he direct his flight? He will not find the Antelope in the peaceful vale." "My brother speaks no fables," said Charles, pale, and deeply moved. "No. Thayendanegea cannot say what is not true. His brother's white sister has been, ere this, conveyed away. It was the decree of the chiefs, solicited by the Queen of the Senecas; but she cannot be injured. You are unhappy?" "Oh," cried the Indian maiden, "let her be brought hither, or go where we go, and I will kiss away her tears and sing her to sleep!" "Sister's son," said the aunt, "let it be so." "It will be so," he replied. "Such is the purpose of the one who decided every thing, and whose decision was merely ratified by the chiefs." "And that was old Esther," said Charles. "Queen Esther," said Brandt. "My brother," said the Delaware chief, Calvin, who had hitherto remained a silent listener, addressing Charles, "I will remain with you, or we will go together, whithersoever the great Ha-wen-no-yu, or our Holy Father, may direct our steps." "Farewell!" said Brandt, rising. "The maple-leaf is red. It has been painted by the first frosts.

#### Here is a definition for a topic.

A **topic** is the specific subject matter or main focus of a piece of writing. It answers the question, "What is this about?" Topics are explicit, straightforward, and usually stated clearly within the text. They deal with facts, events, and specific issues. For instance, a central topic of Harry Potter and the Philosopher's Stone would be "Magic."

#### Here are five possible labels for the central topic of this passage.

native american culture	brotherhood	warfare	war	warriors

Please rank these labels from best to worst (1 being best, 5 being worst) in order of preference. If some are identical just put those in any order as a group (but make sure to place the group in the appropriate rankings relative to the other options!).

If one of the options is blank, put that one last.

Best

1. Select	$\sim$
2. Select	~
3. Select	~
4. Select	~
5. Select	~

Worst

Figure 2: Example screenshot of our survey