

# 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 la-

bels, 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 (Stammach 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 LLM-generated 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 LLM-generated labels equal to or more favorable than









	Passage	Human	Gemma2	GPT4o	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 diplomatic issues.	international relations	international relations	us military	<b>us-china relations</b>	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	<b>prostate cancer</b>	health	health	<b>prostate cancer</b>
NEWS	AI-generated deepfake videos of Rashmika Mandanna and Katrina Kaif have raised concerns about the misuse of deepfake technology, prompting calls for stricter identification methods.	<b>artificial intelligence</b>	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	<b>nature</b>	<b>nature</b>	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.	<b>urban life</b>	<b>urban life</b>	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	<b>household</b>	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 labeling 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







- Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoidi, 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 Stambach, 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.
2.
3.
4.
5.

Worst

Figure 2: Example screenshot of our survey