Tell, Don't Show: Leveraging Language Models' Abstractive Retellings to Model Literary Themes

Li Lucy[1] **Camilla Griffiths**[2]
Sarah Levine[2] **Jennifer L. Eberhardt**[2] **Dorottya Demszky**[2] **David Bamman**[1]
[1]University of California Berkeley [2]Stanford University
lucy3_li@berkeley.edu

Abstract

Conventional bag-of-words approaches for topic modeling, like latent Dirichlet allocation (LDA), struggle with literary text. Literature challenges lexical methods because narrative language focuses on immersive sensory details instead of abstractive description or exposition: writers are advised to *show, don't tell*. We propose `Retell`, a simple, accessible topic modeling approach for literature. Here, we prompt resource-efficient, generative language models (LMs) to *tell* what passages *show*, thereby translating narratives' surface forms into higher-level concepts and themes. By running LDA on LMs' retellings of passages, we can obtain more precise and informative topics than by running LDA alone or by directly asking LMs to list topics. To investigate the potential of our method for cultural analytics, we compare our method's outputs to expert-guided annotations in a case study on racial/cultural identity in high school English language arts books.

1 Introduction

A common task in text analysis is the identification and exploration of latent themes or topics across documents, and unsupervised topic modeling has sustained its popularity in cultural analytics (Boyd-Graber et al., 2017). Latent Dirichlet allocation (LDA) remains popular (Blei et al., 2003; Luhmann and Burghardt, 2022; Fontanella et al., 2024; Sobchuk and Šeļa, 2024), though the strength of language models (LMs) have raised new possibilities (e.g. Pham et al., 2024; Wan et al., 2024). While the rise of high-performing LMs in natural language processing (NLP) has opened new avenues for cultural analytics research, it has also raised barriers for access, especially since researchers in the humanities may be constrained by API costs and compute resources.

Given this interdisciplinary context, we examine how one may perform topic modeling on literary passages with "small", resource-efficient LMs. We use these LMs to address a key challenge of computational literary analysis: literature often contains low-level depictions of characters' actions, thoughts, and dialogue, and so higher-level themes may evade classic text analysis tools reliant on lexical surface forms. This challenge of identifying patterns across documents is especially conspicuous for literary text: the golden rule of narrative writing advises authors to *show, don't tell* (Lubbock, 1921; Burroway et al., 2019). That is, good storytelling should allow readers to experience a narrative through low-level sensory detail instead of high-level abstractive exposition.

To address this issue, we ask LMs to *tell* us what book passages *show* to better surface topics and themes shared across them. We propose a novel yet simple approach for running topic modeling on literary texts: instead of applying LDA on the original texts, we apply it on LMs' abstractive retellings of them (Figure 1). We show that the topics we

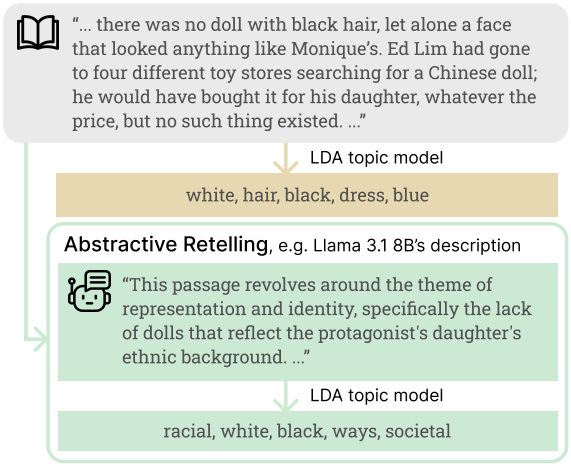

Figure 1: Language models (LMs) can support traditional topic modeling pipelines, e.g. LDA, by translating low-level sensory detail (*showing*) into high-level abstractive exposition (*telling*). This example book excerpt is from *Little Fires Everywhere* by Celeste Ng.

Findings of the Association for Computational Linguistics: ACL 2025, pages 22585–22610

then obtain are more useful for identifying cross-cutting themes than topics outputted by LDA alone and than those obtained by directly requesting topic labels from the same LMs (§5).

Aside from evaluating our approach across a variety of topics, we closely examine its potential for supporting cultural analytics research in a case study. We apply our method on books taught in American high schools, and work with in-domain experts to identify passages that mention and/or discuss racial/cultural identity. Our approach is able to predict passages' topic distributions in a manner that corresponds with human annotators' codes, and identifies more topics and passages of interest than baseline methods. Overall, our conceptual framework of *abstractive retelling* holds promise for literary content analysis. We release our code at `https://github.com/lucy3/tell_dont_show`.

2 Related Work

Topic modeling allows practitioners to model large collections of text in a bottom-up, discovery-oriented manner. Many topic modeling approaches have been proposed over the years, including structural topic models (Roberts et al., 2013), hierarchical models (Griffiths et al., 2003), semi-supervised models (Jagarlamudi et al., 2012; Gallagher et al., 2017), and embedding-based or neural models (Dieng et al., 2020; Card et al., 2018; Bianchi et al., 2021; Burkhardt and Kramer, 2019; Sia et al., 2020). Still, Blei et al. (2003)'s LDA, a generative probabilistic model, remains popular as an analytical tool (e.g. Zhu et al., 2024), given its well-studied output behavior, stability, and ease of use (Hoyle et al., 2022; McCallum, 2002).

Recently, approaches that directly elicit topic labels from LMs have emerged (Pham et al., 2024; Wan et al., 2024; Lam et al., 2024), though these frameworks often involve an intricate chain of prompting steps. In contrast, our proposed improvement over traditional topic modeling requires only one prompt. Our method is most similar to prior work asking LMs to describe documents in a manner that surfaces information that is *implicit* in them (Zhong et al., 2022; Hoyle et al., 2023; Hua et al., 2024; Ravfogel et al., 2024).

Given our focus on literature, our work also contributes to computational humanities (Piper, 2018; Underwood, 2019). At this junction, computational text analysis supports "distant reading" (Moretti, 2013; Jänicke et al., 2015), though some recent work has also examined the use of LMs for close, interpretative reading (Sui et al., 2025). Topic models have been applied to a range of textual forms, including books (Jockers and Mimno, 2013; Thompson and Mimno, 2018; Erlin, 2017), historical newspapers (Newman and Block, 2006; Yang et al., 2011), folklore (Tangherlini and Leonard, 2013), and poetry (Navarro-Colorado, 2018; Rhody, 2012). Topic modeling's appeal in cultural analytics stems from its potential to identify content of interest and relate texts to each other (Tangherlini and Leonard, 2013). Our evaluation approach and case study also consider these use cases (§5, §6).

3 Methods

Our task is to inductively discover latent topics across literary passages. We experiment with instruction-tuned LMs designed to be resource-efficient within various model families: GPT-4o mini, Llama 3.1 8B, Phi-3.5-mini (3.8B), and Gemma 2 2B.

3.1 Our Approach: `Retell`

Our proposed approach for running topic modeling on literary passages first asks an LM to "retell" the contents of a literary passage. We define a passage to be a sequence of paragraphs from a book, containing up to 250 white-spaced tokens in total. We experiment with three retelling `VERB`s – *describe*, *summarize*, or *paraphrase* – in a short prompt:

> In one paragraph, `[VERB]` the following book excerpt for a literary scholar analyzing narrative content. Do not include the book title or author's name in your response; `[VERB]` only the passage.
>
> Passage:
> `[PASSAGE]`

The first two verbs, *describe* and *summarize*, encourage abstraction, while *paraphrase* does not. Since abstraction differentiates *telling* from *showing*, experimenting with several verbs allows us to see if more abstractive variants of `Retell` yield better results. We run each LM with default parameters (Appendix A.1).

We run latent Dirichlet allocation (LDA) using Mallet[1] (McCallum, 2002) on models' passage

[1]Release version `202108`.

retellings. For LDA, we lowercase all text and remove words with fewer than 3 characters, frequent words used by more than 25% of all retellings, and rare words used by fewer than 5 retellings; these text preprocessing steps match those of Thompson and Mimno (2018). We also remove characters' names detected by spaCy's `en_core_web_trf` named entity recognition model. We prompt models to not mention book titles or author names and remove character names from outputs to avoid producing topics that only cluster terms associated with individual books (Jockers, 2013). On average, LMs' retelling lengths range from Gemma 2 2B `Retell`-*summarize*'s 105.5 words to GPT-4o mini `Retell`-*describe*'s 170.0 words.[2]

3.2 Baselines

Latent Dirichlet allocation. Our first baseline runs LDA on the original passages. Like in §3.1, we run the same vocabulary filtering steps and remove characters' names from passages. We compare `Retell` to default LDA with various numbers of topics k.

TopicGPT-lite. Our second baseline is based on Pham et al. (2024)'s TopicGPT framework, which directly elicits topic labels from an LM in a two-stage process. First, during topic generation, an LM suggests topics over a sampled subset of N documents, accumulating a set of possible labels. Second, during topic assignment, the LM assigns topics from the topic set to all documents. Unlike with LDA, k is not preset, but instead refers to the number of unique topics assigned by the LM. In other words, k is determined by the outcome of TopicGPT itself.

Pham et al. (2024) originally developed TopicGPT with GPT-4, and have suggested that it is impractical to use with weaker LMs, e.g. Mistral-7B-Instruct. That is, during topic generation, the list of topics accumulated by weaker LMs can snowball into a large collection of overly specific labels. We observe similar issues, where Llama 3.1 8B generates 486 unique topics for only 100 documents.

We modify TopicGPT to make it feasible to run with smaller LMs, and call our adapted version *TopicGPT-lite*. TopicGPT-lite asks LMs to suggest only one topic, rather than multiple, per document during the topic generation stage. LMs are still allowed to output multiple topics during the assignment stage. Our modification mitigates the overly high k issue: Llama 3.1 8B produces $k = 206$ unique topics assigned to all passages, Phi-3.5-mini produces $k = 538$, and Gemma 2 2B produces $k = 157$. Interestingly, GPT-4o mini tends to under-generate topics rather than over-generate them like other LMs do. So we run two versions using this LM: one with multi-label topic generation, as originally intended by TopicGPT, and one with single-label generation, matching our setup for other LMs. With the former, GPT-4o mini produces $k = 89$ topics and with the latter, $k = 14$. In the main text, we present results for $k = 89$, as it turns out to perform better; results for $k = 14$ can be found in Appendix C.

Since smaller LMs struggle more with instruction following, we minimally edited TopicGPT's prompts for TopicGPT-lite to encourage simple output formatting, or one topic per line. On average, the number of topics assigned per passage ranges from 2.8 (GPT-4o mini) to 9.8 (Llama 3.1 8B). To make TopicGPT-lite's behavior more comparable to LDA's during evaluation, our topic assignment prompt asks LMs to output topics in order of their prominence in the document. Our prompts can be found in Appendix A.2.

Our aim when adapting TopicGPT for use with smaller LMs is to test how our abstractive retelling approach introduced in §3.1 compares to directly eliciting topics from an LM, *given similar resource constraints and uses*. Though their costs are similar, TopicGPT-lite requires more compute and runtime. That is, `Retell` prompts an LM once per passage and then runs LDA over them, while TopicGPT-lite involves longer prompts per passage and also runs an additional generation prompt over N passages.[3] Another practical benefit of `Retell` is that the granularity of topics (k) can be quickly adjusted without rerunning any LM. TopicGPT-lite, on the other hand, may require additional prompt engineering to steer k.

4 Evaluation Setup

We evaluate the general use of our proposed topic modeling method in two ways. First, we obtain pre-existing literary themes/tags from online resources, and use them to efficiently evaluate topic models' performance across passage sets spanning a wide range of topics (§4.1). Second, we conduct human

[2] We use `blingfire`'s word tokenization to calculate these counts: https://github.com/microsoft/BlingFire.

[3] In our experiments, we set $N = 1000$. Pham et al. (2024) recommends choosing an N based on one's compute budget, and $N > 600$ was sufficient for the datasets in their paper.

evaluation of topics outputted at the passage-level, using a setup similar to a classic topic intrusion test (§4.2). Overall, our evaluation setup in this section aims to broadly assess the interpretive utility of topic predictions.

4.1 Relatedness of Passage Sets' Labels

A useful topic model should recover latent topics that resemble gold ones. Here, we examine how well each topic modeling approach draws out latent topics across a set of passages. To create a dataset of literary passages, we draw English literature from both Project Gutenberg (pre-1925) and contemporary book lists (1923-2021), e.g. best-seller lists (Appendix B). In total, we begin with 50.7k unique titles, which eventually result in 256 books from Project Gutenberg and 476 contemporary books after filtering to those with passages matched to gold labels.

We obtain gold topic labels using readers' tagged book quotes on Goodreads, a popular book discussion and reviewing platform, and theme-labeled quotes on SparkNotes and Litcharts, two online study guide publishers. Tags and themes on these sites may reference related or duplicate concepts, e.g. *women and sexism* and *sexism and women's roles*. Thus, we manually recode scraped tags/themes into 27 general topics, e.g. `gender`, some of which appear across multiple label sources. Our dataset is a multi-label one, e.g. a passage discussing interracial relationships would belong to both `love` and `race`.

LDA requires sufficiently long input documents to yield robust topic estimates. We make the assumption that the preceding context of a quote does not deviate far from the tag/theme attached to the quote. We match quotes to book passages, where each passage includes the paragraph around the quote as well as preceding paragraphs for up to 250 white-spaced tokens. In total, we obtain 11.6k labeled passages (21.1k topic-passage pairs), with an average of 1.67 topics per passage. To encourage LDA to learn robust topics, we augment our set of labeled passages with a same-sized sample of random passages from our book collection. Our dataset of literary passages contains a total of 5.02M white-spaced tokens; Appendix B provides additional details on dataset composition.

For each set of passages that share the same gold topic, we ask Prolific crowdworkers to rate on a scale of 1-3 how well the set's most prominently predicted topic relates to its gold one (Table 1). For LDA-based approaches, the most prominent topic, e.g. *school, class, college, boys, children*, is the one with the highest average probability. For TopicGPT-lite, the most prominent topic, e.g. *childhood innocence*, is the one with the highest mean reciprocal rank (Voorhees, 2000). A topic's reciprocal rank is $1/r_i$, where r_i is the rank of topic t for passage i. If the LM does not assign t to i, then $1/r_i = 0$.

In total, disaggregated across our three gold label sources, we evaluate the most prominently predicted topic of 60 passage sets, e.g. the set of passages labeled by Goodreads as `love`, or the set labeled by SparkNotes as `death`. We compare ratings provided for all three `Retell` variants (*summarize*, *describe*, *paraphrase*) and our baselines (default LDA and TopicGPT-lite), with $k = 50$ as well as for various k determined by TopicGPT-lite. Crowdworkers achieve good inter-annotator agreement (weighted Cohen's κ = 0.70). We pay workers $16 an hour, and Appendix C.2 details worker recruitment, quality checks, and annotation instructions.

4.2 Passage-level Topic Relevance

We also evaluate our methods at the passage-level, using a setup inspired by Chang et al. (2009)'s popular topic intrusion test. Here, we present annotators with a random shuffle of the top three topics predicted for a passage and one "intruder" topic. For LDA-based approaches, we sample the intruder from a passage's bottom 50% of predicted topics by probability, and for TopicGPT-lite, we sample the intruder from topics not assigned to the passage by the model. Following Bhatia et al. (2017), we require intruders to be a top-ranked topic for at least one other document, to avoid "junk" or infrequent topics that would be trivial intruders. Annotators rate topics on an ordinal scale from 1 (*Not Relevant*) to 3 (*Very Relevant*) (full task instructions in Appendix C.3). We expect a good topic model's top predicted topic to be very or somewhat relevant to the passage, and intruders should be rated as not relevant.

Since reading and understanding literature is time-intensive, this additional evaluation focuses on six approaches, spanning `Retell` and TopicGPT-lite, determined to be competitive in the previous evaluation. The approaches we select include those pertaining to one open and one closed LM: Llama 3.1 8B and GPT-4o mini. To ensure solid literary comprehension, we recruited

Method	k	SparkNotes gender	Litcharts time
Retell-*summ.*	50	*societal, women, expectations, norms, men*	*past, poignant, time, life, loss*
Retell-*desc.*	50	*women, gender, expectations, roles, norms*	*life, existential, death, existence, mortality*
Default LDA	50	*n't, got, say, man, know*	*n't, know, just, your, get*
TopicGPT-lite	206	*family*	*life*
TopicGPT-lite	89	*loneliness*	*loneliness*
Method	k	Goodreads social class	Litcharts education
Retell-*summ.*	50	*financial, economic, labor, business, work*	*importance, education, knowledge, more, approach*
Retell-*desc.*	50	*economic, financial, wealth, labor, social*	*knowledge, education, educational, intellectual, learning*
Default LDA	50	*money, work, than, business, these*	*n't, know, just, your, get*
TopicGPT-lite	206	*wealth*	*life*
TopicGPT-lite	89	*injustice*	*loneliness*

Table 1: Examples of the most highly ranked/probable topics for passage sets that share the same gold topic. In the left column, = Llama 3.1 8B and = GPT-4o mini. Highlighted predictions are semantically related to gold labels. As a reminder, TopicGPT-lite self-determines k, while Retell and LDA allows us to preset k. On aggregate, abstractive Retell-based approaches make more related predictions than baseline methods (Table 2).

	Default LDA					TopicGPT-lite				Retell-*paraphrase*							
	no language model																
Rating / k	89	206	538	157	50	89	206	538	157	89	50	206	50	538	50	157	50
✓	0.22	0.20	0.27	0.22	0.03	0.28	0.48	0.42	0.33	0.38	0.38	0.58	0.38	0.63	0.47	0.70	0.47
?	0.20	0.12	0.17	0.15	0.12	0.30	0.42	0.30	0.32	0.35	0.37	0.35	0.35	0.20	0.35	0.22	0.32
✗	0.58	0.68	0.57	0.63	0.85	0.42	0.10	0.28	0.35	0.27	0.25	0.07	0.27	0.17	0.18	0.08	0.22

	Retell-*summarize*								Retell-*describe*							
Rating / k	89	50	206	50	538	50	157	50	89	50	206	50	538	50	157	50
✓	0.65	0.60	0.67	0.58	0.62	0.50	0.62	0.50	0.67	0.68	0.55	0.53	0.63	0.63	0.62	0.47
?	0.22	0.30	0.20	0.33	0.23	0.42	0.28	0.37	0.22	0.20	0.35	0.40	0.30	0.32	0.23	0.42
✗	0.13	0.10	0.13	0.08	0.15	0.08	0.10	0.13	0.12	0.12	0.10	0.07	0.07	0.05	0.15	0.12

Table 2: Distributions of crowdworkers' ratings of relatedness between passage sets' most prominently predicted topic labels and their gold labels for GPT-4o mini ($k = 89, k = 50$), Llama 3.1 8B ($k = 206, k = 50$), Gemma 2 2B ($k = 157, k = 50$), and Phi-3.5-mini ($k = 538, k = 50$). On the left, ✓ = "Very Related", **?** = "Somewhat Related", and ✗ = "Not Related". Appendix C.2 includes additional results for GPT-4o mini and Llama 3.1 8B showing how Retell's better ratings generalize with $k = 100, k = 200$.

two in-house annotators, who were not informed of our methodological approach but have prior experience annotating film and literature (Appendix C.3). These annotators rated topics for 50 passages sampled from the dataset introduced in §4.1, with good agreement (weighted Cohen's $\kappa = 0.66$).

5 Evaluation Results

Retell's most prominently predicted topics for passages that share the same gold label demonstrate our approach's ability to surface cross-cutting themes in literary texts (Table 1). Retell's topics contain terms that tend to be more conceptual, e.g. *financial*, than default LDA, e.g. *money*. We also observe that TopicGPT-lite has a tendency to group a high proportion of passages under a few common, overly broad topics, e.g. Gemma's *social interaction* (21.4%) or Llama's *life* (32.9%). In Appendix C.1, we compute methods' precision and recall of passage pairs that share the same gold topic, and find that Retell is generally more precise than baseline methods. Appendix C.1 also details how Retell's higher precision generalizes across different gold label sources, and persists even when LMs are rerun.

Retell produces topics that crowdworkers often judge to be semantically related to passage sets' gold labels (Table 2). By experimenting with different instructive verbs, we wished to see how abstractive verbs that encourage more "telling" (*describe* and *summarize*) may perform compared to a verb that may retain more "showing" (*paraphrase*). As discussed in §1, we hypothesized that the former two should perform better than the latter. Averaged across all four models and choices of k shown in Table 2, we indeed see that Retell-*summarize* (✓ = 0.59) and Retell-*describe* (✓ = 0.60) perform better than Retell-*paraphrase* (✓ = 0.50).

Still, there is some variation in how different models respond to different verbs. For instance, GPT-4o mini's summaries and descriptions outperform its paraphrases, while Gemma 2 2B's

Retell variants do not share the same pattern, and its Retell-*paraphrase* actually performs well (Table 2). Upon closer inspection of these LMs' retellings, we observe that GPT-4o's paraphrases rewrite the original passage in a more direct, low-level manner. Take for example a passage from from John Steinbeck's *The Pearl*, which begins with the sentence *Kino moved sluggishly, arms and legs stirred like those of a crushed bug, and a thick muttering came from his mouth*. GPT-4o's paraphrase of this passage starts with *Kino moved slowly, his limbs resembling those of a squashed insect, as he muttered thickly to himself*. In contrast, Gemma 2 2B is less exact with instruction following, and its "paraphrases" sound more like summaries: *This passage depicts the ripple effect of a single, extraordinary event–the discovery of a legendary pearl–on a small community*. Our observation suggests that though the *paraphrase* vs. *describe*/*summarize* distinction may not be interpreted similarly by all LMs, downstream topic modeling performance may still relate to retellings' level of abstraction.

Regardless of models' idiosyncrasies, our three variants of Retell consistently outperform the baselines default LDA and TopicGPT-lite (Table 2). Frequently predicted TopicGPT-lite labels, e.g. *human nature*, are uninformative and imprecise, with more ratings as "Somewhat Related" than "Very Related" to gold labels. Across all methods, we find that the gold topics related to the social and behavioral concepts hypocrisy, ambition, and innocence are the most difficult to yield very related predictions, while the topics war, religion, and education are easiest.

Human ratings of passage-level topic relevance provide additional support that Retell is a competitive alternative to TopicGPT-lite (Table 3). Here, we compare a few of the more promising Retell approaches, as suggested by Table 2, against TopicGPT-lite. Our evaluation here includes one open model (Llama 3.1 8B) and one closed one (GPT-4o mini), and two values of k for each: $k = 50$ and the k set by TopicGPT-lite. We note that it is not possible to perform apple-to-apple comparisons when comparing the top topics drawn from a probabilistic model (LDA-style outputs) with those outputted by a direct labeling approach (TopicGPT-lite). Still, on average, intruder topics have lower relevance scores than passages' top topics for both methods. Appendix C.3 includes further discussion around the implications of differing topic formats for human interpretation.

Method	Topic 1	Topic 2	Topic 3	Intruder
Retell-*summ.* ∞, $k = 50$	2.36	2.30	2.12	1.67
Retell-*summ.* ∞, $k = 206$	2.40	2.14	2.17	1.81
TopicGPT-lite ∞, $k = 206$	2.64	2.57	2.19	1.40
Retell-*desc.* ⊛, $k = 50$	2.81	2.51	2.23	1.63
Retell-*desc.* ⊛, $k = 89$	2.60	2.53	2.40	1.77
TopicGPT-lite ⊛, $k = 89$	2.59	2.48	2.51	1.52

Table 3: Passage-level human evaluation results in top topics' and intruders' ratings that differ significantly for each method (U-test, $p < 0.05$). Ratings are on a scale of 1 (*Not Relevant*) to 3 (*Very Relevant*) and averaged across 50 passages. In the leftmost column, ∞ = Llama 3.1 8B and ⊛ = GPT-4o mini.

Across two forms of evaluation, we show that our proposed approach, Retell, is capable of identifying cross-cutting themes in literary passages.

6 Case Study: Race in ELA Books

Now that we've evaluated our approach across a variety of LMs and topics, we conduct a deep-dive demonstration and analysis of its ability to support distant reading within a specific use case. A common first-order task for scholars working with literature is identifying passages of interest for closer reading and analysis (Tangherlini and Leonard, 2013). Bottom-up tagging of literary collections with topics can also support search and discovery (Park and Brenza, 2015). Here, we investigate the extent to which topic modeling can surface passages of interest for researchers. Specifically, we examine whether predicted topics align with expert-guided, theory-driven human annotations.

In particular, of ongoing interest in education and social psychology is the question of how race and racism is taught, or could be taught, to students (Perry et al., 2020; Ladson-Billings, 1995; Lynn and Parker, 2006; Bedford and Shaffer, 2023). Within the context of English language arts (ELA), or courses that teach reading and writing skills, researchers may investigate how literature depicts race (e.g. Adukia et al., 2023), how teachers contextualize these depictions (e.g. Beach et al., 2015), and how literature and pedagogical practice may combine to affect students' understanding of themselves and others (e.g. Byrd, 2016; Cocco et al., 2021). To conduct these types of studies, it may be useful to identify literary passages within curricular collections that substantively engage with topics regarding racial/cultural identity.

In this case study, we run topic modeling on 1,645 human-annotated passages drawn from books associated with ELA courses in the U.S. We supplement this set with 19.8k additional passages stratified sampled across these books, to yield robust topic estimates. We compare outputs produced by `Retell`-*summarize* ($k = 50$), TopicGPT-lite, and default LDA ($k = 50$), with the former two run with an open LM, Llama 3.1 8B.[4] By running three different methods juxtaposed against passage-level human annotations, we hope to illustrate how they may differ in practice.

6.1 Data

We draw passages from 396 books listed in Lucy et al. (2025)'s study of racial/ethnic representation in high school ELA.[5] This data originates from two overlapping sources: 250 titles listed in AP Literature exams essay questions in 1999-2021 (College Board, 2024), and 207 titles listed by 189 teachers who taught in schools with mainly Black and Hispanic student populations in 2016-2022 (Levine et al., 2021). This data spans two distinct views of ELA in the U.S. That is, AP Literature represents a standardized, prescriptive view of the literary canon (Abrams, 2023), while teacher-listed books demonstrate how curricula could potentially be adapted for diverse classrooms.

We focus on analyzing the topical composition of passages that *explicitly* mention racial and/or cultural identity. Though race can be implicitly referenced, defining the scope of implicit cues is worthy of more in-depth study in future work, and so we focus on explicit mentions for our demonstrative study. Our view of what constitutes "race" draws from definitions established by educators, sociolinguists, and cultural analytics researchers (Bonilla-Silva, 2015; Morning, 2011; Milner, 2017; Algee-Hewitt et al., 2020). We include inherited categories as well as those shaped by shared cultural practices; that is, we consider references to ethno-religious and geographic origin groups, e.g. *Jewish*, *South Asian*, and racial ones, e.g. *White*.

We use a directed content analysis approach to construct a coding scheme for manually annotating passages that mention or discuss race (Hsieh and Shannon, 2005), iteratively refining face-valid and theory-driven codes. Like in §4, passages are up to 250 words in length while respecting paragraph boundaries. In early stages of annotation, we observe that identifying passages of interest resembled searching for needles in a haystack. Thus, to raise our chances of coming across positive examples, we focus on annotating passages that include keywords in an extensive seed list (Appendix D), including names of countries (e.g. *Jamaica*), racial, ethnic, cultural, or nationality identifiers (e.g. *Turkish*, *Hmong*), or terminology related to racial issues (e.g. *Whiteness*, *undocumented*).

Our annotator team consists of four undergraduates led by the second author, a social psychologist who specializes in research around racial identity. Within four months, these annotators coded 1,645 passages. Annotators found 401 passages containing cursory mentions of racial identity (`mention`), and 198 involving deeper engagement with racial identity and issues (`discuss`), including descriptions of racial/cultural groups, instances of racially-based violence, negative racist language or dialogue, and mentions of racism or racial inequality. We consider passages that do not fall within `mention` nor `discuss` to be negative examples (`neither`). Appendix D details our codebook, annotation process, and agreement levels, which were moderate (Cohen's $\kappa \geq 0.5$) to substantial (Cohen's $\kappa \geq 0.8$), depending on the code. Overall, combing through literature to identify passages of interest required much time and effort.

6.2 Results & Discussion

6.2.1 Topic Models' Behavior

We find that differences in passages' topical emphases relate to annotators' codes. That is, topics signaling racial/cultural identity are prominent in both `mention` and `discuss` passage sets (Table 4). For both default LDA and `Retell`-*summarize*, the probability of topics that strongly suggest race, or *black, people, white* and *black, racial, white* for each approach respectively, are significantly higher in `discuss` than `mention`, matching our intentions in the design of these categories (U-test, $p < 0.001$). In contrast, TopicGPT-lite's top topics do not signal race and/or culture, and are similar across all three passage sets (Table 4). Within the longer tail of TopicGPT-lite's $k = 223$ labels, we do find three topics in its generated topic pool that are suggestive of race: *Slavery*, *Racism*, and *Social Justice*. Though *Social Justice* appears in predictions for 20.7% of passages in `discuss`, it does not specify nor focus on racial/cultural identity.

[4] $k = 100$ produced similar topics and conclusions.

[5] ELA booklists in the Cultural Analytics Dataverse: https://doi.org/10.7910/DVN/WZQRRH.

Set	TopicGPT-lite (k=223)	MRR	Default LDA (k=50)	Prob	Retell-*summarize* (k=50)	Prob
Passages that only mention race (mention)	*Identity*	0.459	*black, people, white, new, american*	0.041	*cultural, identity, heritage, family, complexities*	0.062
	Family	0.357	*war, men, soldiers, two, army*	0.033	*black, racial, white, community, individuals*	0.034
	Work	0.356	*mother, father, family, children, son*	0.032	*identity, self, complexities, explores, own*	0.027
Passages that discuss race (discuss)	*Identity*	0.522	*black, people, white, new, american*	0.085	*black, racial, white, community, individuals*	0.106
	Relationships	0.352	*people, its, more, which, also*	0.057	*cultural, identity, heritage, family, complexities*	0.090
	Family	0.290	*think, want, maybe, really, how*	0.045	*identity, self, complexities, explores, own*	0.041
Other annotated passages (neither)	*Identity*	0.416	*think, want, maybe, really, how*	0.029	*rather, than, human, not, can*	0.031
	Work	0.361	*says, has, can, does, say*	0.028	*family, mother, father, dynamics, son*	0.029
	Family	0.357	*people, its, more, which, also*	0.028	*natural, world, landscape, nature, scene*	0.027

Table 4: The top three most prominent topics in passages that only mention racial/cultural identity (mention), those that discuss it (discuss), and those that do not mention nor discuss it (neither), as defined in §6.1. Topics are ordered by mean reciprocal rank (MRR) for TopicGPT-lite, and ordered by their average probability for default LDA and Retell.

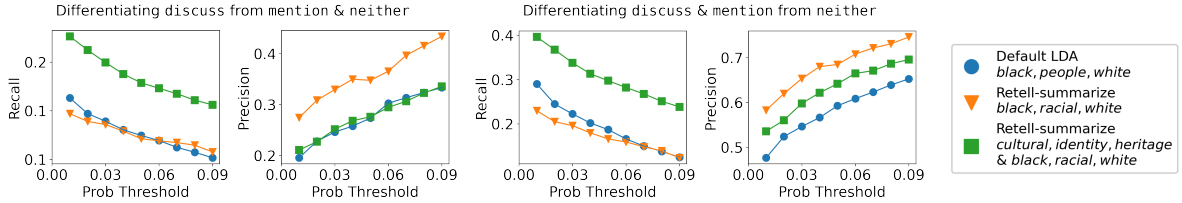

Figure 2: Precision and recall (y-axes) of using relevant Retell and LDA topics' probability thresholds (x-axes) to distinguish different types of human-labeled passages from each other (discuss, mention, neither). Examining ● against ▼ compares only one topic per method, and ▼ is more precise. Comparing ● with ■ considers all relevant topics each method surfaced, and ■ has higher recall without sacrificing precision. Left two plots: identifying passages that discuss race. Right two plots: identifying passages that either discuss or mention race.

Next, we examine how various thresholds of topic probabilities compare to human annotations. That is, consider a simple classifier that predicts passages with one or more topics above some probability cutoff c to be those that discuss and/or mention racial/cultural identity. Figure 2 shows that with Retell's two relevant top topics, we're able to recover more passages of interest than with the one relevant top topic predicted by default LDA. That is, when Retell's *black, racial, white* is paired with a second prominent topic *cultural, identity, heritage*, this approach can recall more passages than LDA's *black, people, white* alone, without sacrificing precision. When considering only one topic for each method, Retell's *black, racial, white* can more precisely identify passages in discuss and mention than LDA's analogous topic can (Figure 2).

Though our case study focuses on racial and cultural identity, Retell's general ability to surface a range of different topics (§5) holds potential for the study of thematic co-occurrences. Table 5 illustrates example passages where Retell topics related to racial and/or cultural identity intermix with other socially significant topics, including ones relevant for intersectional studies. With this table, one can also concretely see how each method's outputs differ at the passage-level, and the kinds of terms that compose the topics produced by each.

6.2.2 Error Analysis

Though our approach shows promise compared to baselines, the performance in Figure 2 leaves room for improvement. To understand why Llama 3.1 8B's retellings may fall short, we qualitatively examine the 20 most severe false positives, or passages labeled as neither that have the highest probabilities of our two target topics. A few of these cases involve human annotators missing mentions of race, e.g. *our rightful place ... in Black history*, suggesting a potential use of Retell as additional verification. In other cases, passages' discussion of racial/cultural identity is more implicit, falling outside the scope of our annotation codebook. For example, eight false positives discuss language use and differences, which the LM relates to culture or heritage.

With some false positives, we observe that the LM's retelling may include additional book context that is not present in the original passage. For example, one retelling contrasts the passage's dialogue with a character's Iranian background, which is not mentioned in the passage but is accurately described by the LM. This output behavior arises

Passage excerpt	`Retell`-*summarize*	Default LDA	TopicGPT-lite
... white families in one of the wealthiest cities in the country could hire colored domestics for as little as five dollars a week ... — The Warmth of Other Suns by Isabel Wilkerson	*financial, economic, family, reveals, life* *black, racial, white, community, individuals* *labor, work, animals, also, ability*	*money, work, get, pay, dollars* *got, than, much, more, because* *people, its, more, which, also*	*Work* *Poverty* *Security*
... these protesters are so much more vehement about Canadian-born Japanese ... Lots of people have been fired from their jobs. Business on Powell Street ... — Obasan by Joy Kogawa	*black, racial, white, community, individuals* *city, new, its, town, including* *financial, economic, family, reveals, life*	*think, want, maybe, really, how* *black, people, white, new, american* *very, old, man, much, little*	*Work* *Business* *Identity*
Black boys are taught to replicate the white boy game ... I've played the game over and over again and have wounded black girls and women ... — No Ashes in the Fire by Darnell L. Moore	*black, racial, white, community, individuals* *societal, expectations, women, norms, social* *rather, than, human, not, can*	*people, its, more, which, also* *play, music, game, playing, other* *which, men, these, great, man*	*Relationships* *Identity* *Work*
There is something important about women in Guatemala, especially Indian women ... This feeling is born in women because of the responsibilities they have, which men do not have. — I, Rigoberta Menchú by Rigoberta Menchú	*societal, expectations, women, norms, social* *cultural, identity, heritage, family, complexities* *community, public, personal, including, system*	*people, its, more, which, also* *woman, women, love, girl, wife* *which, life, its, love, heart*	*Identity* *Family* *Work*

Table 5: Illustrative excerpts from example passages and the top three topics predicted by each approach. These passages originate from the unlabeled pool of additional passages included in topic modeling (§6.1). `Retell`-*summarize* with Llama 3.1 8B relates the first two passages to socioeconomic status, and the last two to gendered social expectations. Our approach also relates all four passages to racial and/or cultural identity.

from the presence of book summaries or the books themselves in pretraining data (Chang et al., 2023; Cooper et al., 2025). Like with human readers, an LM's interpretation of a passage may vary based on its familiarity with how the passage is situated within a larger narrative. Altogether, our observations around false positives suggest challenges around drawing a consistent divide between the implicit and explicit, as well as the passage-level and the book-level, when using LMs for literary text analysis.

We also examine the 20 most severe false negatives, or passages labeled as `discuss` that have the lowest probabilities of our two target topics. Here, the LM's retellings may fail to recap content of interest, or insufficiently relate content to race/culture. For example, one retelling did not acknowledge the passage's description of a racist stereotype. An understudied challenge in text summarization research is determining what content is significant enough to recollect (e.g Ladhak et al., 2020), and these decisions may vary across different downstream tasks. Generally, we observe that classic summarization issues such as faithfulness and coherence (Zhang et al., 2024; Fabbri et al., 2021) do not severely degrade `Retell`. Instead, our observations point towards a need for more research characterizing LMs' *content selection* behavior.

7 Conclusion

We propose a conceptually intuitive "abstractive retelling" approach for topic modeling of literature that pairs language models with classic LDA. Our proposed approach allows LDA to go beyond lexical-level sensory details in narrative text, and offers a competitive alternative to both LDA alone and directly generating topic labels, especially for resource-efficient, "small" LMs. We also examine our method's potential for search and exploration of literary collections in a case study on race in ELA books. Though `Retell` has room for improvement, it holds potential for supporting content analyses of literary collections at scale. Overall, we hope that the simplicity of our method encourages uptake in research around LMs' affordances and caveats in cultural analytics.

Limitations

Our framework proposes the idea that LMs can help us "translate" low-level narrative details (e.g. the clinking of glass at a table) into higher level concepts (e.g. a dinner scene). Though our method performs well against baselines, relying on LMs' "retellings" of passages can be reductive. Reading literature is a subjective, culturally constructed, and interpretive process (Fish, 1982), and LMs' retellings of passages typically only represent one

interpretation. In the main text, we also touch on other shortcomings of LMs' retellings, including their failure to recap content of interest from the original text.

Given our interdisciplinary audience's interests and constraints, we focus mainly on resource-efficient language models. This leaves open the question of how our approach may perform with larger LMs. We include some preliminary results comparing `Retell`-*describe* and TopicGPT-lite with GPT-4o in Appendix C.4. There, we find that this stronger LM is proficient at directly outputting topic labels, so TopicGPT-lite achieves similar crowdsourced ratings of topic relatedness to ground truth topic labels as `Retell`. GPT-4o's TopicGPT-lite's distribution of passages across topics is also less skewed than that of smaller LMs (e.g. the most common top 3 topic, *love*, appears in only 10.5% of passages). These initial findings suggest that the task of outputting summaries/descriptions is easier than outputting topic labels directly, and so when a situation calls for smaller and/or weaker LMs, `Retell` provides a suitable alternative. We leave a more in-depth analysis for future work. Even if 2B to 8B sized LMs like the ones we tested improve, the million parameter models of the future may behave like the small billion parameter models of today. So, having options for topic modeling with different affordances (e.g. easily adjustable k) is still useful.

We use gold labels scraped from online resources to evaluate our proposed method, allowing us to acquire many topics from both readers (Goodreads) and literature experts (LitCharts and SparkNotes). Still, we acknowledge that our use of a few select online resources risks content production biases, where content is affected by the context of a creator population (Olteanu et al., 2019). An issue of coverage also arises in any case of using naturally found data; online resources may not cover all possible tags/themes pertaining to a passage. So, we also include some human evaluation that does not use these gold labels (§4.2, §6).

Ethical Considerations

Our human evaluation in §4 was deemed exempt from review by the institutional review board for human subjects research at our institution.

Our case study in §6 codes passages mentioning race, and thus requires us to define and operationalize this abstract, socially constructed concept. Appendix D.2 describes how we scoped our annotations, including the definitions of race we consulted from prior work (Algee-Hewitt et al., 2020; Bonilla-Silva, 2015; Morning, 2011; Milner, 2017). For example, our annotation of race also includes ethnic and cultural categories. Though these constructs are interrelated, we acknowledge that race, ethnicity, and culture are distinct concepts (e.g. Spencer, 2014; Worrell, 2014).

Our case study also proposes the use of LMs for identifying passages that discuss sensitive topics, e.g. racial/cultural identity. Extant social psychology and education research have shown that educational content and discussions about race, identity, and culture can be enriching and beneficial for all students, promoting empathy and critical thinking and reducing discrimination (Dee and Penner, 2017; Brannon and Lin, 2021; Apfelbaum et al., 2010). Distant reading tools like ours should be used for *supporting*, but not replacing, human researchers whose work intend to inform inclusive pedagogical practices. Our case study is mainly a proof-of-concept to further test our approach. Given the current sociopolitical climate in the U.S. around book banning and the teaching of race (Alter, 2024), computational content analysis tools may be used to target books in harmful ways. LMs' descriptions and summaries of text, in their current form, should not be used for informing policy decisions, and the use of LMs in high stakes settings such as education requires careful collaboration with in-domain human experts.

Acknowledgements

We thank Mackenzie Cramer and Anna Ho for providing human evaluation of passage-level topics (§4.2), and Ariyanna Wesley, Kayla Collins, Kendal Murray, and Shevaun Yip for their efforts in coding racial/cultural identity for our case study (§6). We also thank Julia Proshan for coordinating our coding process. The research reported in this article was supported by funding from the National Science Foundation (IIS-1942591).

References

Annie Abrams. 2023. Shortchanged: How advanced placement cheats students. *Johns Hopkins University Press*.

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2023.

What We Teach About Race and Gender: Representation in Images and Text of Children's Books. *The Quarterly Journal of Economics*, 138(4):2225–2285.

Mark Algee-Hewitt, JD Porter, and Hannah Walser. 2020. Representing race and ethnicity in american fiction, 1789-1920. *Journal of Cultural Analytics*, 5(2).

Alexandra Alter. 2024. Book bans continue to surge in public schools.

Evan P. Apfelbaum, Kristin Pauker, Samuel R. Sommers, and Nalini Ambady. 2010. In blind pursuit of racial equality? *Psychological Science*, 21(11):1587–1592. PMID: 20876878.

David Bamman. 2021. BookNLP. A natural language processing pipeline for books.

Richard Beach, Anthony Johnston, and Amanda Haertling Thein. 2015. *Identity-focused ELA teaching: A curriculum framework for diverse learners and contexts*. Routledge.

Melissa J. Bedford and Shelly Shaffer. 2023. Examining literature through tenets of critical race theory: A pedagogical approach for the ELA classroom. *Multicultural Perspectives*, 25(1):4–20.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2017. An automatic approach for document-level topic model evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 206–215, Vancouver, Canada. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Eduardo Bonilla-Silva. 2015. The structure of racism in color-blind, "post-racial" america. *American Behavioral Scientist*, 59(11):1358–1376.

Jordan Boyd-Graber, Yuening Hu, David Mimno, and 1 others. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Tiffany N Brannon and Andy Lin. 2021. "pride and prejudice" pathways to belonging: Implications for inclusive diversity practices within mainstream institutions. *American Psychologist*, 76(3):488.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27.

Janet Burroway, Elizabeth Stuckey-French, and Ned Stuckey-French. 2019. *Writing fiction: A guide to narrative craft*. University of Chicago Press.

Christy M Byrd. 2016. Does culturally relevant teaching work? An examination from student perspectives. *Sage Open*, 6(3):2158244016660744.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040.

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.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Veronica Margherita Cocco, Elisa Bisagno, Gian Antonio Di Bernardo, Alessia Cadamuro, Sara Debora Riboldi, Eleonora Crapolicchio, Elena Trifiletti, Sofia Stathi, and Loris Vezzali. 2021. Comparing story reading and video watching as two distinct forms of vicarious contact: An experimental intervention among elementary school children. *British Journal of Social Psychology*, 60(1):74–94.

College Board. 2024. [link].

A. Feder Cooper, Aaron Gokaslan, Amy B. Cyphert, Christopher De Sa, Mark A. Lemley, Daniel E. Ho, and Percy Liang. 2025. Extracting memorized pieces of (copyrighted) books from open-weight language models. *Preprint*, arXiv:2505.12546.

Thomas S. Dee and Emily K. Penner. 2017. The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal*, 54(1):127–166.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Matt Erlin. 2017. Topic Modeling, Epistemology, and the English and German Novel. *Journal of Cultural Analytics*, 2(2).

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Stanley Fish. 1982. *Is There a Text in This Class?: The Authority of Interpretive Communities*. Harvard University Press.

Lara Fontanella, Berta Chulvi, Elisa Ignazzi, Annalina Sarra, and Alice Tontodimamma. 2024. How do we study misogyny in the digital age? A systematic literature review using a computational linguistic approach. *Humanities and Social Sciences Communications*, 11(1):1–15.

Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288.

Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. How did we get here? Summarizing conversation dynamics. 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 7452–7477, Mexico City, Mexico. Association for Computational Linguistics.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, Avignon, France. Association for Computational Linguistics.

Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, Gerik Scheuermann, and 1 others. 2015. On close and distant reading in digital humanities: A survey and future challenges. *EuroVis (STARs)*, 2015:83–103.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Matthew L Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054, Online. Association for Computational Linguistics.

Gloria Ladson-Billings. 1995. Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32(3):465–491.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Sarah Levine, Karoline Trepper, Rosalie Hiuyan Chung, and Raquel Coelho. 2021. How feeling supports students' interpretive discussions about literature. *Journal of Literacy Research*, 53(4):491–515.

Percy Lubbock. 1921. *The craft of fiction*, volume 29. Jonathan Cape.

Li Lucy, Camilla Griffiths, Claire Ying, JJ Kim-Ebio, Sabrina Baur, Sarah Levine, Jennifer L. Eberhardt, David Bamman, and Dorottya Demszky. 2025. Racial and Ethnic Representation in Literature Taught in US High Schools. *Journal of Cultural Analytics*, 10(1).

Jan Luhmann and Manuel Burghardt. 2022. Digital humanities—a discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology*, 73(2):148–171.

Marvin Lynn and Laurence Parker. 2006. Critical race studies in education: Examining a decade of research on us schools. *The Urban Review*, 38:257–290.

Andrew McCallum. 2002. Mallet: A machine learning for language toolkit. *http://mallet.cs.umass.edu*.

H. Richard Milner. 2017. Where's the race in culturally relevant pedagogy? *Teachers College Record*, 119(1):1–32.

F. Moretti. 2013. *Distant Reading*. Verso.

Ann Morning. 2011. Introduction: What is race? In *The Nature of Race: How Scientists Think and Teach about Human Difference*. University of California Press.

Borja Navarro-Colorado. 2018. On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. *Frontiers in Digital Humanities*, 5:15.

David J. Newman and Sharon Block. 2006. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Jung-ran Park and Andrew Brenza. 2015. Evaluation of semi-automatic metadata generation tools: A survey of the current state of the art. *Information Technology and Libraries*, 34(3):22–42.

Sylvia Perry, Allison L Skinner, Jamie Abaied, Adilene Osnaya, and Sara Waters. 2020. Initial evidence that parent-child conversations about race reduce racial biases among white us children.

Chau Pham, 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)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Andrew Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press.

Shauli Ravfogel, Valentina Pyatkin, Amir David Nissan Cohen, Avshalom Manevich, and Yoav Goldberg. 2024. Description-based text similarity. In *First Conference on Language Modeling*.

Lisa Rhody. 2012. Topic model data for topic modeling and figurative language. *Journal of Digital Humanities*, 2.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, and 1 others. 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.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Oleg Sobchuk and Artjoms Šeļa. 2024. Computational thematics: comparing algorithms for clustering the genres of literary fiction. *Humanities and Social Sciences Communications*, 11(1):1–12.

Stephen Spencer. 2014. *Race and ethnicity: Culture, identity and representation*. Routledge.

Peiqi Sui, Juan Diego Rodriguez, Philippe Laban, Dean Murphy, Joseph P. Dexter, Richard Jean So, Samuel Baker, and Pramit Chaudhuri. 2025. KRISTEVA: Close reading as a novel task for benchmarking interpretive reasoning. *Preprint*, arXiv:2505.09825.

Timothy R Tangherlini and Peter Leonard. 2013. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, 41(6):725–749.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Ellen M. Voorhees. 2000. The TREC-8 question answering track report. In *Proc. Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246*, pages 77–82.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, and 1 others. 2024. TnT-LLM: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5836–5847.

Frank C. Worrell. 2014. Culture as race/ethnicity. In *The Oxford Handbook of Identity Development*. Oxford University Press.

Tze-I Yang, Andrew Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104, Portland, OR, USA. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. 2022. Describing differences between text distributions with natural language. In *International Conference on Machine Learning*, pages 27099–27116. PMLR.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36.

A Modeling Details

A.1 Retelling Parameters

We run all open language models on 2 L40 GPUs, and download these models from Hugging Face. At the time of our experiments, `GPT-4o mini` in OpenAI's API pointed to `GPT-4o mini-2024-07-18`.

We run all LMs with `max_new_tokens` as 1024. To determine the "default" setups of each LM, we copied proposed settings in LMs' Hugging Face model cards or API documentation. For example, Phi-3.5-mini's Hugging Face model card recommends a temperature of 0. For LMs that allow a system prompt, which excludes Gemma 2, we used "You are a helpful assistant; follow the instructions in the prompt." Our Github repository also includes code concretely showing how we run each model.

A.2 TopicGPT-lite

TopicGPT originally asked models to generate a topic hierarchy and assign topics within this hierarchy (Pham et al., 2024). We simplify the task to only generate flat topics. During output parsing, we ignore any lines that do not follow our specified format of one topic per line, and during topic assignment, we disregard labels that were not introduced in the topic generation stage.

In TopicGPT, the original topic assignment prompt asks the model to support their assignment with evidence quoted from the passage. We remove this request from our prompts to reduce the difficulty of output formatting for smaller LMs. We also do not include refinement and correction stages in our adaptation of TopicGPT, which we constrain to require similar compute and model calls as `Retell`. That is, Pham et al. (2024) propose reprompting the original LM if its outputs are not distinctive enough after topic generation, or if it hallucinates topics not in the provided list during topic assignment. Some models, such as Phi-3.5-mini, produces outputs that do not match the provided topic pool in 59.9% of examples; reprompting to iteratively correct these issues would require numerous calls to the model, thus increasing the computational resource needs of this topic modeling approach.

We run all TopicGPT-lite experiments using a fork of Pham et al. (2024)'s code, from a version last updated by the authors on March 27, 2024. Figures 5, 6, 7 show our prompts for running TopicGPT-lite. These prompts were minimally edited from the original TopicGPT prompts, to encourage consistent instruction-following and output formatting from resource-efficient LMs.

Literary Passages & Topic Labels	
# of total passages for topic modeling	23,332
# of passages with gold topic labels	11,666
Avg # of white-spaced tokens per passage	214.98
Total # of white-spaced tokens	5.02M
# of books across gold-labeled passages	732
% of passages from Project Gutenberg	25.6
% of passages from contemporary booklists	74.4
Avg # of gold topics per labeled passage	1.67
# of topic-passage pairs in total	21,096
# labeled based on Goodreads tags	9,072
# labeled based on Litcharts themes	11,588
# labeled based on SparkNotes themes	436

Table 6: An overview of the labels and literary passages we use to evaluate topic modeling methods. The percentages from each book source do not total to 100 because they contain some overlapping titles.

B Evaluation Data Collection

B.1 Books

Our evaluation data encompasses both older, out-of-copyright titles and contemporary booklists. We obtain the former from Project Gutenberg, which provides already digitized full texts of books. For the latter, we include newer titles published in 1923-2021, digitizing them by scanning physical books. We use ABBYY Finereader for optical character recognition, and on a sample of excerpts from 20 books, we estimate character error rate to be on average 0.0049 (95% CI [0.0002, 0.0096]).[6] We collect these contemporary booklists from a range of sources: Black-authored texts from the Black Book Interactive Project, non-U.S./U.K. global Anglophone fiction, Pulitzer Prize nominees, bestsellers

[6]Each excerpt is taken from the first page of books' main narratives, excluding paratext, and are at least 100 characters long, respecting paragraph boundaries.

Topic	Goodreads	LitCharts	SparkNotes
love	2269	789	40
literature & stories	572	349	-
evil & crime	429	1344	65
war	320	186	24
death	671	323	22
friendship	244	240	-
childhood, adulthood, & age	353	530	-
religion	742	513	29
family	615	897	23
politics & government	196	545	-
gender	739	1416	65
social class	234	782	52
time	146	145	-
race & racism	767	1002	87
sex & sexuality	83	253	-
forgiveness	57	148	-
education	244	114	-
language	151	383	29
innocence	31	144	-
justice	64	203	-
guilt	46	151	-
hypocrisy	40	99	-
morality	59	255	-
tradition	-	102	-
culture	-	471	-
loyalty	-	94	-
ambition	-	110	-

Table 7: Number of passages within each gold topic label. Topic labels that do not have passages from that label source are indicated with a '-'.

reported by Publisher's Weekly and the New York Times, and Lucy et al. (2025)'s high school ELA dataset.

B.2 Themes, Tags, & Topics

We match books in our collection to their pages on Goodreads, LitCharts, and SparkNotes, and scrape for tags/themes that accompany quoted material. We matched quotes in online resources to passages in books using fuzzy matching, where the normalized indel similarity between the two should exceed 90, on a scale from 0 to 100. Indel similarity considers the minimum number of insertions or deletions between two sequences. Our matching process achieved a match rate of 68.0. Common failure cases include misquotes, such as the quoted content missing a significant portion of the original book excerpt, or quotes that contain "..." signaling a portion of the excerpt was deliberately excluded. Qualitatively, we observed that the preceding passage leading up to a quote often shared similar themes as the quote, with the quote acting like a "punch line" to the passage as whole. Table 6 includes a breakdown of our evaluation dataset, and Table 7 includes our full list of topics per label source and the number of passages with each label.

Our evaluation dataset is a multi-label one, as some quotes have multiple tags/themes, and some tags/themes are recoded into multiple topics. That is, passages originally labeled with *interracial relationships* belong to both `love` and `race`. We create our set of gold topic labels by manually reviewing all Goodreads tags that occur at least 50 times among our books and all Litcharts and Sparknotes themes. To reduce the cognitive load of sifting through a messy pile of labels, we sorted tags/themes by words in them during the recoding process (e.g. the themes *women and gender* and *men and women* would be listed together since they share the word *women*). We exclude tags/themes that did not fall within our final set of 27 gold topic labels, and also exclude Goodreads tags that did not relate to topic, e.g. author-related tags (*john-green*) or tags suggesting meta-level reader commentary (*inspirational*).

C Evaluation

C.1 Automatic Metrics

One way to quantify topic modeling behavior across a range of LMs is to compute the precision and recall of passage pairs linked by each gold label source, e.g. Sparknotes `family` or Goodreads `religion`. We calculate these metrics in a manner that accounts for our multi-label data.

Precision. Given a passage i labeled with an unordered set of n gold topics $y_{i1}, y_{i2}, ..., y_{in} \in Y_i$, a topic modeling method predicts the passage's p topics in order of most to least prominent: $\hat{y}_{i1}$, $\hat{y}_{i2}$, ... $\hat{y}_{ip}$. For LDA-based methods, $p = k$ since each passage has probabilities calculated across all topics, while for TopicGPT-lite, a model may output variable p for each passage, where $p < k$. To calculate precision, we calculate the fraction of predicted positives, which are passage pairs i, j with $\hat{y}_{i1} = \hat{y}_{j1}$, that share at least one gold topic label, or $Y_i \cap Y_j \neq \varnothing$.

Recall. We calculate recall based on a passage i's m unordered gold tag/theme labels $z_{i1}, ..., z_{im} \in Z_i$. These tags/themes are unique to each source, e.g. *unrequited-love* in Goodreads, and they are finer-grained than gold topic labels, e.g. `love`. Using tags/themes allows us to avoid over-penalizing the failure to recall broad topics with higher k. Considering our multi-label setting, we compute recall with passages' top three $\hat{y}$. That is, we calculate the fraction of passage pairs i, j with $Z_i \cap Z_j \neq \varnothing$ that have $\{\hat{y}_{i1}, \hat{y}_{i2}, \hat{y}_{i3}\} \cap \{\hat{y}_{j1}, \hat{y}_{j2}, \hat{y}_{j3}\} \neq \varnothing$.

Results. Linking passage pairs with shared themes is a difficult task; abstractive retelling can

	Default LDA						TopicGPT-lite			Retell-*describe*										Retell-*summarize*								
	no language model						O		L	O					L					O					L			
Rating	14	89	206	50	100	200	14	89	206	14	89	50	100	200	206	50	100	200	14	89	50	100	200	206	50	100	200
✓	0.07	0.22	0.20	0.03	0.25	0.20	0.30	0.28	0.48	0.38	0.67	0.68	0.68	0.67	0.55	0.53	0.48	0.60	0.20	0.65	0.60	0.60	0.67	0.67	0.58	0.60	0.63
?	0.18	0.20	0.12	0.12	0.13	0.20	0.28	0.30	0.42	0.37	0.22	0.20	0.23	0.27	0.35	0.40	0.37	0.22	0.60	0.22	0.30	0.30	0.27	0.20	0.33	0.25	0.25
✗	0.75	0.58	0.68	0.85	0.62	0.60	0.42	0.42	0.10	0.25	0.12	0.12	0.08	0.07	0.10	0.07	0.15	0.18	0.20	0.13	0.10	0.10	0.07	0.13	0.08	0.15	0.12

Table 8: An version of Table 2 containing crowdsourcing results for more values of k and one open LM (Llama 3.1 8B) and one closed one (GPT-4o mini).

be more precise than baseline methods, but has similar or lower recall (Table 9). As discussed in the main text (§4), high recall for methods like TopicGPT-lite arise from this approach labeling many passages with the same label, e.g. *human nature*. Though these common labels may be generally relevant to passages, they are not always informative or distinctive. Our automatic metric results do support our *tell* vs. *show* motivation; Retell-*summarize* and Retell-*describe* tend to perform better than Retell-*paraphrase* across all four LMs.

Figure 3 breaks down precision and recall results for Retell by label source, providing additional confirmation of the aggregated trends of Table 9. That is, Retell-*describe* and Retell-*summarize* perform better than Retell-*paraphrase*. Generally, descriptions and summaries are more similar to each other than they are to paraphrases. The Jensen-Shannon divergence between the word distributions of each LM's descriptions and summaries ($M = 0.012$, $SD = 0.004$) is lower than that of paraphrases vs. summaries ($M = 0.029$, $SD = 0.021$) and paraphrases vs. descriptions ($M = 0.048$, $SD = 0.034$).

Past work on topic modeling has highlighted stability issues in some methods, where small changes in parameters or rerunning the same model may yield consequentially different results (Hoyle et al., 2022). Two of our LMs, GPT-4o mini and Llama 3.1 8B, use default parameters where they are run with non-zero temperatures and non-greedy decoding, and so they produce different retellings when reprompted. Figure 4 shows that there is some variation between two different runs of Retell. Still, Retell-*summarize* and Retell-*describe* still typically have better performance than Retell-*paraphrase*, and for precision, the former two consistently outperform default LDA.

C.2 Topic Relatedness Crowdsourcing

In §4.1, we compare passage sets' most prominently predicted labels with ground truth ones. We recruited Prolific workers based in the U.S. whose first language is English, with an approval rate between 95 and 100 and at least 10 prior submissions. Across all method and parameter combinations shown in Table 2, 8, and 10, crowdworkers annotated 2940 pairs of predicted and ground truth labels, as each approach involved annotating 60 passage sets. We calculate inter-annotator agreement on 304 doubly annotated examples. Workers annotated examples in batches of up to 12 task examples at a time, where one example may be doubly annotated with another batch, and one example was a quality check example. We recollected ratings in cases where workers failed this check.

These quality check examples were sampled from the following 20 possibilities, with the correct rating in parentheses and an ampersand delineating the faux gold and predicted topic labels, respectively: sports & sports (Very Related), animals & literature (Not Related), sports & pitcher, sports, inning, player, bat (Very Related), animals & table, couch, lamp, rug, coffee (Not Related), nature & the natural world (Very Related), fruits and vegetable & spaceship (Not Related), farm & cow, chicken, farmer, farm, barn (Very Related), vehicle, & asleep, nap, dreamy, pillow, sleep (Not Related), hatred & hate (Very Related), kindness & lawn, grass, weeds, bee, garden (Not Related), hygiene & toothbrush, dental, shower, wellness, wash (Very Related), technology & fruit (Not Related), sports & athletics (Very Related), animals & crisis (Not Related), vehicle & driving, motor, car, vehicle, transport (Very Related), kindness & bakery (Not Related), technology & technology (Very Related), hygiene & be, the, can, if, will (Not Related), farm & farming and rural life (Very Related), hatred & cat, dog, pet, fish, hamster (Not Related).

Our full task instructions are the following:

Rate how much a word, phrase, or list of words is related to a given theme. There are three options: "Very related", "Somewhat related", and "Not related".

Note: "very related" includes cases where the provided word/s contains the theme.

Examples:

Theme: nature
Word/s or Phrase: environment
Answer: Very related

Theme: nature
Word/s or Phrase: biotechnology
Answer: Somewhat related

Theme: nature
Word/s or Phrase: table
Answer: Not related

Theme: nature
Word/s or Phrase: tree, lake, mountain, animals, nature
Answer: Very related

Theme: nature
Word/s or Phrase: biology, cells, organism, microscope, experiment
Answer: Somewhat related

Theme: nature
Word/s or Phrase: computer, keyboard, mouse, wifi, hardware
Answer: Not related

Then, for each predicted and gold topic label pair, we asked:

How related are the given word/s or phrase to the theme?
Theme: `gold label`
Word/s or Phrase: `predicted label`

C.3 Passage-level Topic Annotation

Our two annotators for this passage-level evaluation were finishing bachelor's degrees in rhetoric, English, data science, and cognitive science when conducting their annotations. Each had approximately two years of experience working with cultural analytics datasets, including literature and film. They annotated 50 passage and topic sets in text files, with one file per passage. Within each passage's annotation file, the passage itself was repeated above each of six sets of four topics to reduce scrolling. Annotators were not told how the topics were generated or that only one topic was an "intruder".

We included the following instructions:
In this task, you will read a literary passage. Try your best to understand the themes and concepts in the passage thoroughly.
Then, you will rate the relevance of topics, or themes, to a passage. For each rating, you have three options: 3 - Very Relevant, 2 - Somewhat Relevant, and 1 - Not Relevant. Type your judgement within 1 to 3 next to each topic (see example below).
A relevant topic should represent prominent themes or concepts in the passage. Each topic is either represented by a list of words (e.g. "school, students, education, high, student"), or by a single word or phrase (e.g. "Relationships"). You will be presented with four topics at a time, and there are six sets of them in total.
You may ignore errors caused by the book's digitization process, which may create misspellings, extraneous letters or numbers, or unusual line/paragraph breaks.

At the end of these instructions, we include one example passage from *Always Running: La Vida Loca* by Luis J. Rodriguez, along with four labeled `Retell` topics, to show annotators how to format their responses.

Annotators doubly annotated ten passages (ratings for 240 topics in total) when calculating inter-annotator agreement (IAA). The two annotators and the first author discussed the topic options for one passage, to calibrate our expectations of what ratings of "Very Relevant", "Somewhat Relevant", and "Not Relevant" may entail. During this discussion, we acknowledged that this task involves subjective or inferential interpretation of literary text. That is, the two annotators may have *valid* disagreements when relating topics to passages, and thus should follow their own intuition while annotating, rather than rate what they expect the other annotator to rate solely for increasing IAA.

When annotating the first 10 passages, annotators also asked the first author clarifying questions around the instructions. For example, one annotator initially stated that a military topic was *not* relevant to a passage from *The Three Musketeers* by Alexandre Dumas, because the musketeer characters are *private* soldiers rather than formally part of the military. Initial discussions over cases of uncertainty like these were helpful for calibrating the scale of what is or is not relevant, and that topics may cover broad concepts (e.g. "military" may not differentiate different types of soldiers).

Our annotators also made other observations during this task. For example, some topics (which in some cases were intruders) included in `Retell` topic sets were uninformative, e.g. "*can, who, than, how, rather*" do not convey much actual content. Default LDA, which we did not annotate at the passage-level, is especially liable to producing stopword-esque topics like these (e.g. Table 1), and though `Retell` is less prone to this issue, it still occurs. In addition, some `Retell` topics may be difficult to interpret if one is unfamiliar with LDA-style outputs. For example, not all would agree that *black, racial, white, community, individuals* should relate to passages that discuss race, but do not focus on Black or White identity. That is, sometimes only a few, but not all, words in an LDA-style topic label may be relevant to a passage. Our annotators tended to rate relevance based off of the most relevant word in a `Retell` topic's top five words.

C.4 GPT-4o Performance and Behavior

Our main text primarily focuses on resource-efficient members of various model families. Here, we conduct a preliminary analysis of GPT-4o, primarily comparing `Retell`-*describe* and TopicGPT-lite. Since GPT-4o-mini tends to perform better with *describe* than *summarize* (Table 8), we choose the former verb for our GPT-4o experiments. For TopicGPT-lite, since GPT-4o does not devolve in the same manner as smaller LMs with multiple-topic generation, we use the prompt in Figure 6 to generate topics (much like the $k = 89$ version of GPT-4o-mini's TopicGPT-lite). For topic assignment, we use the prompt in Figure 7, matching our setup with all other LMs. This process produced $k = 152$ topics, and so for "fair" comparison, we run GPT-4o's `Retell`-*describe* with $k = 152$ as well. Evaluation results for these two approaches can be found in Table 10. We find that TopicGPT-lite improves with a stronger LM, with higher precision and recall than `Retell` and similar crowdsourced topic relatedness ratings.

D Case Study: ELA Books

***Content warning**: the following section includes depictions of racial/ethnic violence and racist language.*

D.1 Passage Sampling

As discussed in the main text, preliminary annotations revealed that identifying passages of interest that mention and/or discuss race required reading many passages before hitting a positive example, much like searching for needles in a haystack. Thus, to yield a sufficient number of positive examples within the months we spent with our annotators, we annotated passages sampled from all texts that contain at least one keyword. Possible keywords include the following:

- 109 common words or phrases corresponding to racial, ethnic, cultural, or nationality identifiers found in our ELA dataset of 396 books. We obtained these by running the named entity recognition pipeline from BookNLP (Bamman, 2021) and manually filtering through the top 500 most common nominal PERSONs. Unlike externally scraped dictionaries and wordlists, his process yields terms that are actually used within our dataset, such as "Oriental", "Scotchman", and "Creole".

- 484 names of countries and capitals from GeoNames, a geographical database.

- 1,948 demonyms,[7] ethnonyms,[8], Native American tribe names,[9] and nationalities[10] from Wiktionary.

- 78 words that tend to co-occur within the same 250 token window with the following seed words: "race", "racial", "racist", "racists", "racism", "immigrants", "immigrant", "ethnic", "minority", "minorities", "diversity", "discrimination", "identity", "blackness", "whiteness". Co-occurrence is measured by calculating the normalized pointwise mutual information between a potential keyword and one of these seed words in ELA books, and we use a threshold of 0.15 for determining significant co-occurrence. From this resulting list, we manually filter for words that may be suggestive of race-related discussion, e.g. "segregation", "migrant", and "supremacy."

[7] https://en.wiktionary.org/wiki/Category:en:Demonyms

[8] https://en.wiktionary.org/wiki/Category:en:Ethnonyms

[9] https://en.wiktionary.org/wiki/Category:en:Native_American_tribes

[10] https://en.wiktionary.org/wiki/Category:en:Nationalities

D.2 Annotation Codebook

D.2.1 Defining "race"

After sampling passages to annotate, the second author, a social psychologist, constructed an annotation codebook for determining what constitutes explicit mentions and discussions of race. Race is a socially constructed, contested, and multidimensional concept (Field et al., 2021; Hanna et al., 2020). So, our instructions for annotators included several definitions of race, presenting it as a phenomenon that overlaps with ethnicity, religion, geography, biology, and culture. We generally follow precedent set by Algee-Hewitt et al. (2020), who conducted a cultural analytics study of racial and ethnic representation in American fiction:

> The categories include broad racial/ethnic groups (e.g., black, white, Native American, Pacific Islander, Latinx), ethno-religious distinctions (e.g., Catholic, Jewish, Middle Eastern and Muslim), geographical origins (e.g., South Asian, Filipino, East Asian, Eastern European, German/Dutch, Irish, Italian, Latin American, Scandinavian), and one catch-all field (Immigrants).

To Algee-Hewitt et al. (2020)'s definition above, we added racial/ethnic groups of Pacific Islanders and Hispanic/Latinx people. We also theoretically grounded ourselves in definitions of race from sociologists and educators, including:

- "Race is a biological or cultural category easy to read through marks in the body (phenotype) or the cultural practices of groups" (Bonilla-Silva, 2015).
- "Race is a system for classifying human beings that is grounded in the belief that they embody inherited and fixed biological characteristics that identify them as members of racial groups" (Morning, 2011).
- "Culture can be defined as deep-rooted values, beliefs, languages, customs, and norms shared among a group of people" (Milner, 2017).

We want to be clear that we do not conflate the categories of race, ethnicity, and culture, and we understand them as distinct socially constructed identifiers (e.g. Spencer, 2014; Worrell, 2014). Our coding effort required us to operationalize these dimensions and aggregate across them, so as not to produce too many themes.

D.2.2 Codebook development and codes

We used a directed content analysis approach (Hsieh and Shannon, 2005) in developing our coding scheme. We start with face-valid and theory-driven themes and using the training phases of the annotation process to fine-tune, iterate, and further define these initial themes, and to add new themes to capture instances that could not be coded by the initial themes. We only code for explicit mentions or discussions of race, ethnicity, and culture, due to the fact that more implicit instances would sometimes require the coder to use information outside of the passage itself to infer that there was a reference to race, ethnicity, or culture. In this section, we include a few examples of what types of content would or would not fall within each code; we provided more extensive examples to our annotators during the coding process. Passages can be labeled with multiple codes.

We started with three themes that captured explicit and face-valid instances of race, ethnicity, or culture being described or discussed. In each case, we wanted to capture language where there was an explicit engagement with identity:

A. Descriptions of racial/ethnic/cultural identity groups. This code includes descriptions of a racial/ethnic group, or a community as such. It does not include descriptions of an individual person, unless that person is explicitly racialized and framed as representative of their group. These descriptions can be in a first person or third person perspective, and include descriptions of a trait, experience, practice, tradition, or condition shared by a group.

- Affirmative example: ... *he didn't know that all Indian families are unhappy for the same exact reasons: the fricking booze* in *The Absolutely True Diary of a Part-Time Indian* by Sherman Alexie. Note that this example would also be coded as **C**, as defined below.
- Non-example: *Nikki was lightskinned and from Texas, about five-foot-nine* in *No Disrespect* by Sister Souljah.

B. Instances of racially-based violence. This code includes the discussion or mention of racially-based violence.

- Affirmative example: ... *let those [n-word] stand there with guns, and we don't accommodate them? They want war, let's give them*

war in *A Gathering of Old Men* by Ernest J. Gaines.

- Non-example (too implicit): ... *true that some of the elders of the Somali community remained angry ...* in *The Burgess Boys* by Elizabeth Strout.

C. Negative racist language or dialogue. This code includes racist or prejudiced language spoken by, to, or about a character. It can involve non-dialogical language indicating racist beliefs or behaviors of characters. The language may be directed towards a group or towards a person.

- Affirmative example: *We're not such paupers as to sell to Japs, are we?* in *Snow Falling on Cedars* by David Guterson.

During the early training phase of coding, we encouraged coders to suggest additional themes for instances where a passage did not fit within an existing theme. From this process, we added two additional themes:

D. Mentions of racial, ethnic, or cultural identity. This code includes cases where race or racial identity appears in a passage, but may not reach the level of engagement as initial codes. The group should be explicitly mentioned as group identity - i.e., talked about as a racial, ethnic, or cultural group and not simply as a descriptor or mention of a country. Character names or language can be used as cues to group identity only if they corroborate other cues to group identity. Can include mentions of racial/ethnic groups in fictional or inanimate contexts (e.g., on TV).

- Affirmative example: ... *who grew up in such poverty that other poor Indians called her family poor?* in *Reservation Blues* by Sherman Alexie.

- Non-example: *I sat dumbfounded among the Chinese antiquities and Greek vases...* in *The Goldfinch* by Donna Tartt.

E. Mentions of racism or inequality. Explicit or blatant mention or labeling of something as racism, anti-racist, inequality, racial injustice, racial segregation, desegregation, or racial privilege, without the characters necessarily engaging in racist language or acts.

- ... *we couldn't have racism in America without racists* in *All American Boys* by Jason Reynolds and Brendan Kiely.

In the main text (§6), passages labeled as codes **A**, **B**, **C**, and **E** constitute passage set `discuss`, or deeper engagement with topics around racial/cultural identity, while passages labeled as **D** but *not* **A**, **B**, **C**, nor **E** constitute passage set `mention`. Passages that do not fall under any of the codes are considered negative examples in `neither`.

D.3 Annotation Process

We annotated passages in two stages, as our annotation process spanned two academic terms, and different student annotators were available in each term.

Stage 1. Two undergraduate students with academic backgrounds in psychology and cognitive science annotated 711 passages in the first term. This time period was focused on refining and finalizing the coding scheme. We collectively coded some passages, refined descriptions and definitions for existing themes, and added new themes when necessary (Appendix D.2.2). Our process matched how directed content analysis is generally conducted (Hsieh and Shannon, 2005).

Annotators were assigned batches of passages to annotate and discuss with the first and second authors in a weekly meeting, with the number of passages increasing with every week. We started with 10 passages the first week and 150 passages by the 6th week. Both annotators coded the same 711 passages. In each team meeting, the second author compiled each instance where annotators disagreed about the presence or absence of a given code for the same passage. For each disagreement, the annotators discussed their reasoning for why they did or did not include a code, came to an understanding about the 'correct' code, and recorded the agreed-upon code. In the instance where there wasn't an existing theme for a given passage, the team considered additions to the coding scheme. When new themes were added, annotators went back to previously coded passages to determine whether any passages should be assigned the newly defined themes. Table 11 includes levels of inter-annotator agreement we obtained at the end of five weeks.

Stage 2. We recruited two new undergraduate student annotators, whose academic backgrounds also included cognitive science and psychology. During this stage, we focused solely on achieving

agreement between new annotators and annotating a larger sample of passages. Our collective annotation process mirrored that of Stage 1, and during this phase, we annotated 435 passages. Table 11 details our inter-annotator agreement after six weeks of collective annotation. We note that percent agreement calculations do not take into account the number of instances, while Cohen's κ does, and so our agreement for **A** has a lower κ score than other categories. In addition, there were not enough instances of **B** and **C** to properly calculate Cohen's κ, hence those values are listed as 'NA'. We concluded our second annotation stage with the students independently coding 500 passages, or 250 per student.

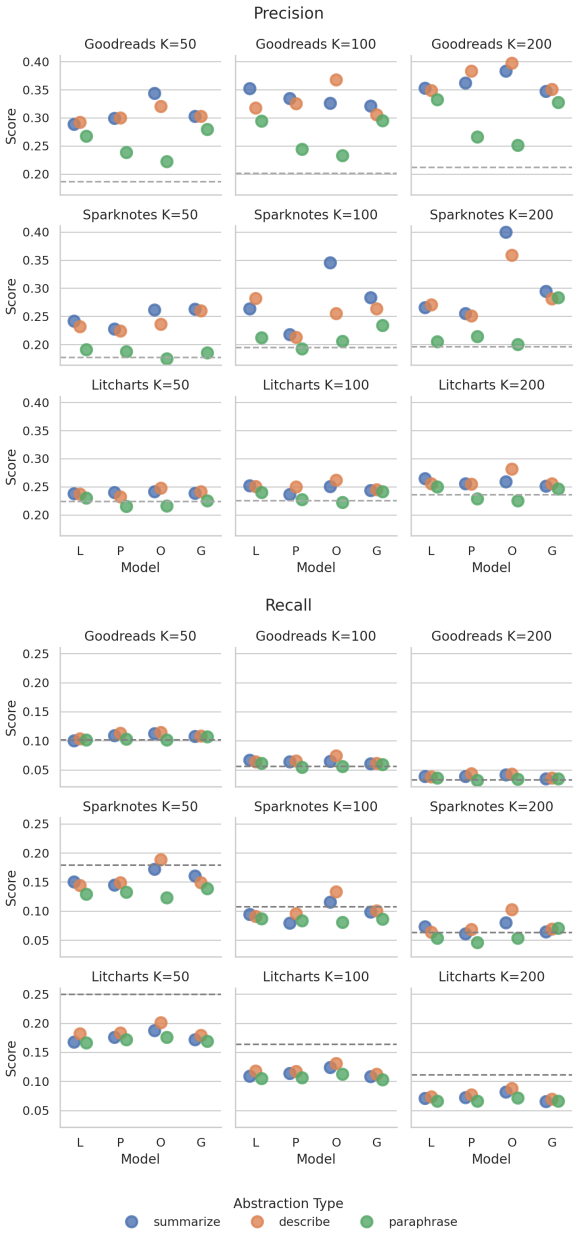

Figure 3: Additional evaluation results for $k = 50$, 100, and 200, disaggregated by the source of ground truth labels. Dashed gray lines indicate default LDA's performance, and the models on the x-axis include Llama 3.1 8B (L), Phi-3.5-mini (P), Gemma 2 2B (G) and GPT-4o mini (O). Since TopicGPT-lite is not designed for use with predetermined k, it is not included in these plots.

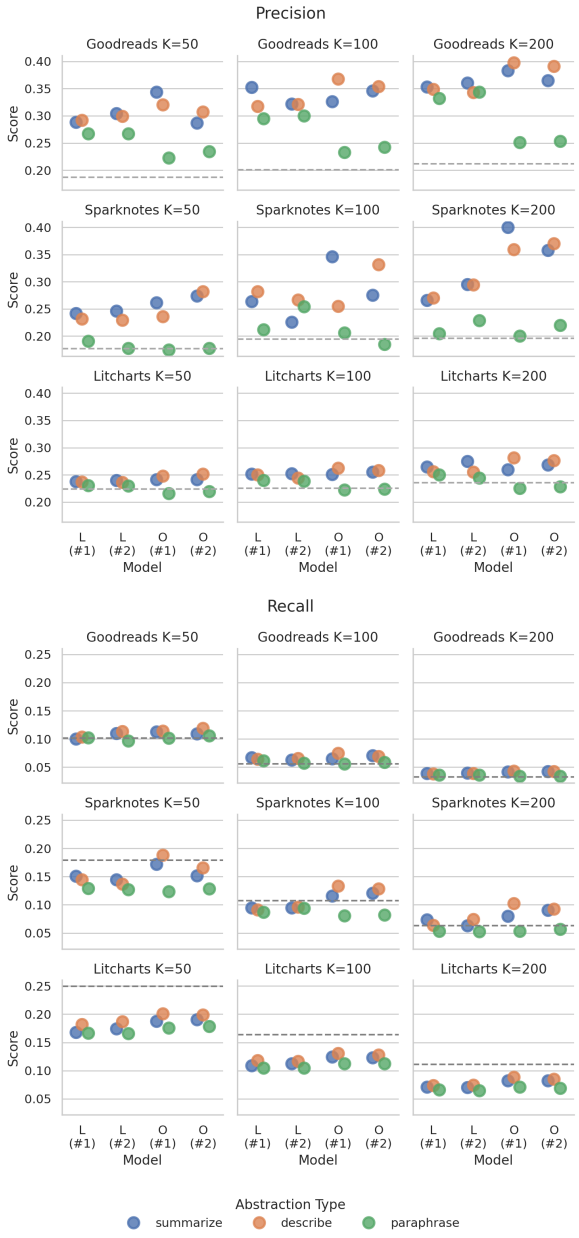

Figure 4: Results for two different runs of Llama 3.1 8B (L) and GPT-4o mini (O). Dashed gray lines indicate default LDA's performance.

Model	k	Method	Precision	Recall
Llama 3.1 8B	50	`Retell`-*summarize*	0.128	0.103
	50	`Retell`-*describe*	**0.128**	0.107
	50	`Retell`-*paraphrase*	0.121	0.105
	50	default LDA (no LM)	0.100	**0.108**
	206	`Retell`-*summarize*	**0.155**	0.040
	206	`Retell`-*describe*	0.149	0.039
	206	`Retell`-*paraphrase*	0.135	0.036
	206	default LDA (no LM)	0.107	0.036
	206	TopicGPT-lite	0.127	**0.211**
Phi-3.5-mini	50	`Retell`-*summarize*	**0.131**	0.112
	50	`Retell`-*describe*	0.131	**0.116**
	50	`Retell`-*paraphrase*	0.110	0.106
	50	default LDA (no LM)	0.100	0.108
	538	`Retell`-*summarize*	**0.170**	0.020
	538	`Retell`-*describe*	0.159	0.021
	538	`Retell`-*paraphrase*	0.128	0.017
	538	default LDA (no LM)	0.114	0.018
	538	TopicGPT-lite	0.139	**0.048**
Gemma 2 2B	50	`Retell`-*summarize*	0.131	0.111
	50	`Retell`-*describe*	**0.133**	**0.111**
	50	`Retell`-*paraphrase*	0.123	0.110
	50	default LDA (no LM)	0.100	0.108
	157	`Retell`-*summarize*	**0.147**	0.045
	157	`Retell`-*describe*	0.145	0.046
	157	`Retell`-*paraphrase*	0.138	0.044
	157	default LDA (no LM)	0.104	0.042
	157	TopicGPT-lite	0.139	**0.201**
GPT-4o mini	50	`Retell`-*summarize*	**0.141**	0.116
	50	`Retell`-*describe*	0.137	**0.118**
	50	`Retell`-*paraphrase*	0.106	0.105
	50	default LDA (no LM)	0.100	0.108
	14	`Retell`-*summarize*	0.110	0.244
	14	`Retell`-*describe*	0.118	0.243
	14	`Retell`-*paraphrase*	0.100	0.226
	14	default LDA (no LM)	0.092	0.222
	14	TopicGPT-lite	**0.127**	**0.248**
	89	`Retell`-*summarize*	0.141	0.079
	89	`Retell`-*describe*	**0.154**	0.082
	89	`Retell`-*paraphrase*	0.112	0.068
	89	default LDA (no LM)	0.105	0.069
	89	TopicGPT-lite	0.148	**0.186**

Table 9: Precision and recall of `Retell` against baselines for $k = 50$ or various k set by TopicGPT-lite. `Retell` can have higher precision than baseline approaches, and the high recall of TopicGPT-lite can lead to lower interpretive utility (§5, Table 2). The largest 95% CI over 1k bootstrapped passage pairs for any value in this table is $\pm 8.0 \times 10^{-5}$, and since our results are rounded to the thousandth, we do not include CI ranges in this table. Highest values for each LM and choice of k are bolded.

	TopicGPT-lite	`Retell`-*describe*
✓	0.60	0.63
?	0.25	0.23
✗	0.15	0.13
Precision	0.24	0.15
Recall	0.12	0.05

Table 10: When used with a stronger LM (GPT-4o), TopicGPT-lite has higher pairwise precision and recall than `Retell`-*describe*. Its topic relatedness ratings by Prolific crowdworkers (first three rows) also improve to become similar to `Retell`-*describe*.

	Code	% agreement	Cohen's κ
Stage 1	A. Race description	97.1	0.603
	B. Race-based violence	98.9	0.794
	C. Racism language	99.4	0.854
	D. Race mention	94.3	0.865
	E. Racism mention	97.7	0.870
Stage 2	A. Race description	93.0	0.500
	B. Race-based violence	99.0	NA
	C. Racism language	97.0	NA
	D. Race mention	93.0	0.855
	E. Racism mention	97.7	0.740

Table 11: Inter-annotator agreement during the two annotation stages for our case study of race in ELA literary passages. We report the % of agreed-upon passages and unweighted Cohen's κ.

TopicGPT-lite: generating one topic per passage

```
You will receive a document and a set of topics.  Your task is to identify a generalizable
topic within the document.  If this topic is missing from the provided set, please add it.
Otherwise, output an existing topic as identified in the document.

[Topics]
Topics

[Examples]
Example 1:  Adding "Religion"
Document:
But a religion true to its nature must also be concerned about man's social conditions...
Such a religion is the kind the Marxists like to see-an opiate of the people.

Your response:
Religion:  Describes purposes and roles of religion for people.

Example 2:  Respond with an existing topic, "Love"
Document:
''I have reason to think,'' he replied, ''that Harriet Smith ...  is he sure that Harriet means
to marry him?''

Your response:
Love:  Discusses romantic relationships, such as marriage.

[Instructions]
Step 1:  Determine the topic mentioned in the document.
- The topic label must be as GENERALIZABLE as possible.  It must not be document-specific.
- The topic must reflect a SINGLE topic instead of a combination of topics.
- A new topic must have a short general label and a topic description.
Step 2:  Perform ONE of the following operations:
1.  If there are already a duplicate or relevant topic in the provided set of topics, output
that topic and stop here.
2.  If the document contains no topic, return "None".
3.  Otherwise, add a new topic to the set of topics.

[Document]
Document

Your response should be one line and ONLY contain a topic, written in the format "[Topic
label]:  [your reasoning]".
Your response:
```

Figure 5: Prompt for generating a possible topic pool over 1k passages. We truncate the few-shot examples here for brevity (with "..."), but show the beginnings and ends to support their reconstruction. The first example is from Martin Luther King Jr.'s *Stride Toward Freedom*, while the second is from Jane Austen's *Emma*.

TopicGPT-lite: generating multiple topics per passage

```
You will receive a document and a set of topics.  Your task is to identify generalizable
topics within the document.  If any relevant topics are missing from the provided set, please
add them.  Otherwise, output any existing topics identified in the document.

[Topics]
Topics

[Examples]
Example 1:  Adding "Religion"
Document:
But a religion true to its nature must also be concerned about man's social conditions...
Such a religion is the kind the Marxists like to see-an opiate of the people.

Your response:
Religion:  Describes purposes and roles of religion for people.

Example 2:  Respond with an existing topic, "Love"
Document:
‘‘I have reason to think,’’ he replied, ‘‘that Harriet Smith ...  is he sure that Harriet means
to marry him?’’

Your response:
Love:  Discusses romantic relationships, such as marriage.

[Instructions]
Step 1:  Determine topics mentioned in the document.
- The topic labels must be as GENERALIZABLE as possible.  They must not be document-specific.
- The topics must reflect a SINGLE topic instead of a combination of topics.
- Each new topic must have a short general label and a topic description.
Step 2:  Perform ONE of the following operations:
1.  If there are already duplicates or relevant topics in the provided set of topics, output
those topics and stop here.
2.  If the document contains no topic, return "None".
3.  Otherwise, output new, additional topic(s).

[Document]
Document

Your response should ONLY contain topics, written in the format "[Topic label]:  [your
reasoning]".
Your response:
```

Figure 6: A modified version of Figure 5 for generating a possible topic pool over 1k passages, where instead of asking an LM to generate one topic per passage, it can generate multiple.

TopicGPT-lite: assigning topics to passages

```
You will receive a document and a set of topics.  Assign the document to the most relevant
topics.  Then, output the topic labels and assignment reasoning.  DO NOT make up new topics.

[Topics]
tree

[Examples]
Example 1:  Assign "Religion" to the document
Document:
The second step for me to morph from animal to human was taking me to the church youth group
...  pray in separate rooms so religions wouldn't collide.

Assignment:
Religion:  Describes people's religious practices.

Example 2:  Assign "Love" to the document
Document:
"Good-bye, my love," answered the countess ...  smile fluttering between her lips and her
eyes, she gave her hand to Vronsky.

Assignment:
Love:  Describes a scene where one person declares and shows affection for another.

[Instructions]
1.  Topic labels must be present in the provided set of topics.  You MUST NOT make up new
topics.
2.  Output topic(s) you assign in order of their prominence in the document, with the most
prominent topic first.
3.  Each line of your response should contain a topic written in the format "[Topic label]:
[your reasoning]".

[Document]
Document

Your response should ONLY contain topics.  Double check that your assignment exists in the
provided set of topics!
Your response:
```

Figure 7: Prompt for assigning topics to passages. Few-shot examples are shortened with "..." for brevity purposes, and the first is from *Gabi, a Girl in Pieces* by Isabel Quintero, while the second is from *Anna Karenina* by Leo Tolstoy.