PROXANN: Use-Oriented Evaluations of Topic Models and Document Clustering

Anonymous ACL submission

Abstract

Topic models and document-clustering evaluations either use automated metrics that align poorly with human preferences, or require expert labels that are intractable to scale. We design a scalable human evaluation protocol and a corresponding automated approximation that reflect practitioners' real-world usage of models. Annotators—or an LLM-based proxy—review text items assigned to a topic or cluster, infer a category for the group, then apply that category to other documents. Using this protocol, we collect extensive crowdworker annotations of outputs from a diverse set of topic models on two datasets. We then use these annotations to validate automated proxies, finding that the best LLM proxy is statistically indistinguishable from a human annotator and can therefore serve as a reasonable substitute in automated evaluations.

1 Introduction

Suppose a researcher wants to study the impact of donations on politicians' speech. For the past two decades, such questions have often been answered with the help of topic models or other text-clustering techniques (Baden et al., 2022; Ying et al., 2022). Here, the research team might interpret topic model estimates as representing healthcare or taxation categories, and associate each legislator with the topics they discuss. Researchers could then measure the influence of a donation on the change in the legislators' topic mixture—showing that, e.g., money from a pharmaceutical company increases their focus on healthcare.

The crucial supposition of such a "text-as-data" approach is that the interpreted categories are valid measurements of underlying concepts (Grimmer and Stewart, 2013; Ying et al., 2022; Zhang et al., 2024a). Adapting an example from Ying et al., plausible interpretations of model estimates might yield either healthcare or medical research, which would carry "very different substantive implications" for a research area. Facilitating the identification of valid categories is therefore a key concern in real-world settings, which falls under the framework of *qualitative content analysis* (QCA, Mayring, 2000), a primary use case for topic models (Grimmer and Stewart, 2013; Bakharia et al., 2016; Li et al., 2024).

Taking the view that effective evaluations are those that approximate the real-world requirements of the use case (Liao and Xiao, 2023), it then follows that topic model (and document clustering) evaluations should help encourage valid categories (Ying et al., 2022). However, as we discuss in Section 2, the evaluation strategies that are reason-

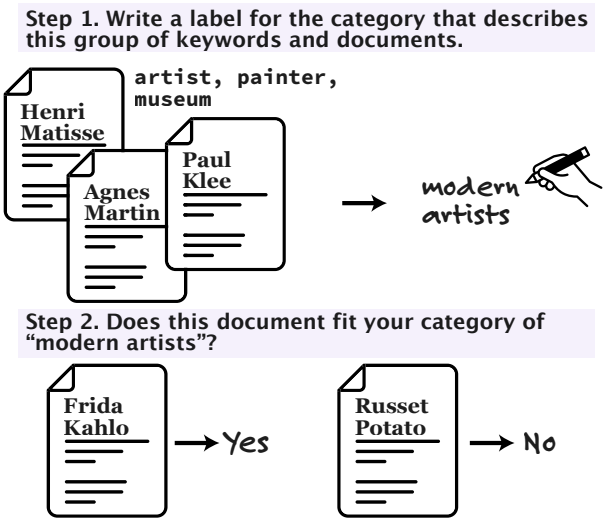


Figure 1: Our evaluation protocol for topic models and document clustering methods. First, a user reviews documents and keywords related to a topic or cluster and identifies a category. Then, they apply that category to new documents (a third ranking step is not shown). The more human relevance judgments align with corresponding model estimates, the better the model. Importantly, the protocol is straightforward to adapt to an LLM prompt, creating a "proxy annotator", PROXANN.

able approximations for this use case are generally dependent on human-derived ground truth, rendering them hard to scale and reproduce. Conversely, the most common unsupervised automated metrics, while fast to compute, tend to be poor measures of topic quality (Doogan and Buntine, 2021).

This paper addresses these shortcomings by introducing both an application-grounded human evaluation protocol and a corresponding automated metric that can substitute for a human evaluator. The protocol approximates the standard qualitative content analysis process, where categories are first derived from text data and subsequently applied to new items, Fig. 1; our human study collects multiple annotations for dozens of topics, making it the largest of its kind.[1] Using both open-source and proprietary large language models (LLMs), we develop "proxy annotators" that complete the tasks comparably to an arbitrary human annotator; we call the method PROXANN. In addition, results from the human evaluation indicate that a classical model (LDA Blei et al., 2003) performs at least as well, if not better, than its modern equivalents.

2 Background and Prior Work

We outline necessary background regarding topic models, clustering methods and their evaluations. We start with the goals of topic modeling, turn to standard automated evaluations, then outline use-oriented measures based on human input.

2.1 Making sense of document collections

The systematic categorization of text datasets is a common activity in many fields, particularly in the social sciences and humanities. A common manual framework to help structure the recognition of categories in texts is *qualitative content analysis* (QCA, Mayring, 2000; Smith, 2000; Elo and Kyngäs, 2008, *inter alia*). Broadly, it consists of an inductive process whereby categories emerge from data, which are then consolidated into a final codeset. These categories are then deductively assigned to new documents, supporting downstream analyses and understanding (e.g., characterizing the changing prevalence of categories over time).[2]

NLP offers techniques that are designed to support this process—and that are often conceived as analogues of manual approaches (Baumer et al., 2017; Bakharia et al., 2016). These methods are typically unsupervised, and among the most prevalent are *topic models* (Blei et al., 2003). A topic model is a generative model of documents, where each document is represented by an admixture of latent topics θ_d, and each topic is in turn a distribution over words types β_k (which a user can interpret as a category). For example, when analyzing a corpus of U.S. legislation, suppose the most probable words for one topic include `doctor`, `medicine`, `health`, `patient` and a document with a high probability for that topic is the text of the *Affordable Care Act*; together, they appear to convey a healthcare category.

More recently, the improved representation capacity of sequence embeddings (e.g., sentence transformers, Reimers and Gurevych, 2019) has led to their use in clustering (see Zhang et al. 2022 for an overview). As with topic modeling, a document is associated with one or more clusters (an equivalent to θ_d); succinct labels (standing in for β_k) for clusters can be obtained with various word-selection methods or language-model summaries.

2.2 Evaluating Categorizations

Topic Coherence. Topic model evaluation has primarily focused on the semantic *coherence* of the most probable words in a topic—the capacity for a set of terms to "enable human recognition of an identifiable category" (Hoyle et al., 2021). Boyd-Graber et al. 2014 consider a topic's coherence to be a precondition for a useful model, and indeed, applied works often validate topics by presenting the top words (Ying et al., 2022)—which, in many cases, is the *only* form of validation. While Ying et al. 2022 attempt to standardize evaluations of topic-word coherence (building on Chang et al. 2009), the reliance on crowdworkers renders them difficult and costly to scale. As a result, methodological contributions—where easily-applied metrics can help guide model development—tend to use *automated* proxies for coherence, like Normalized Pointwise Mutual Information (NPMI, Lau et al., 2014). Despite their ubiquity, automated coherence metrics fail to align closely with human judgments, exaggerating differences between topics (Hoyle et al., 2021).[3] Newer automated metrics

[1] We will release code and data upon acceptance, and commit to retracting the paper if we fail to follow through.

[2] Practitioners in various communities have developed related families of methodologies with similar goals, such as *grounded theory* (Glaser and Strauss, 1967) and *reflexive thematic analysis* (Braun and Clarke, 2006))

[3] Lim and Lauw 2024 have investigated this relationship further, but with artificial topics not generated by a model.

MALLET	CTM	BERTopic
season game games home runs	*career hit games season league*	*yard season team yards league*
Major League Baseball Players and History	Major League Baseball Players and Achievements	Sports and Athletics
Professional baseball players	Former MLB players	Professional Basketball and Baseball Players
American professional baseball players	American baseball league	American sports and their associated famous sportsmen
Baseball knowledge hub	Professional baseball facts and figures	Sports champions
act consumer credit employee card	*fuel credit revenue internal property*	*vehicle recorder motor retrieved retrieval*
Labor and Employment Legislation	Renewable Energy Tax Credits and Incentives	Vehicle Data Privacy and Ownership Rights
Individual Protection Laws	Renewable energy tax and biofuel	Vehicle owner protections
Labor Laws and Protections	Energy tax credits, Alternative fuel credits	Automobile Ownership Legislation
Proposed employee protections	Energy Tax Policy	Vehicle Owner and Safety Legislation

Table 1: GPT-4O and human annotator-provided category labels for a sample of matched topics from each model (*topic model words are in italics*) for `Wiki` (top row) and `Bills` (bottom row) datasets. Labels are consistent across humans and models.

based on LLMs face similar issues, lacking a clear relationship to actual usage and human judgments of quality (details in Section 6).

Beyond Topic Coherence. In contrast, our contribution closely matches the standard qualitative analysis: developing and applying categories to text items. Although coherent topic-words (or category labels) are important for interpretability, they are not sufficient to establish that model outputs are valid. Categories are also assigned to individual text items, and those assignments should be "meaningful, appropriate, and useful" (Boyd-Graber et al., 2014). Furthermore, the coherence of the topic-words may not agree with the perceived quality of the document-topic distribution (Bhatia et al., 2017). For topic models, Doogan and Buntine 2021 therefore argue that measuring the coherence of the top documents for each topic is necessary for a holistic model evaluation.[4] Several prior efforts have situated model evaluation in the context of their use, but these works rely on on manual label assignments (either pre-existing or via interaction), limiting their broader utility (additional discussion in Section 6).

3 Evaluation Methodology

This section proposes a human evaluation protocol for topic models and document clustering methods. The evaluation is oriented toward real-world use, emulating how practitioners develop categories from—and assign them to—text data in applied settings. Alongside the human tasks, we also develop LLM prompts that adapt the human instructions, treating the LLM as a proxy annotator, PROXANN.

In brief, a sample of documents and keywords for each topic or cluster are shown to an annotator to establish its semantic category (as in the first step in Ying et al. 2022); the annotator then reviews additional documents and labels them based on their relatedness to the category. These *category identification* and *relevance judgment* steps follow that of qualitative content analysis, "a manual process of inductive discovery of codesets via *emergent coding*" (Stemler, 2000). We also include a *representativeness ranking* task as an additional evaluation signal, inspired by "verbatim selection" in qualitative settings (Corden et al., 2006).

As a whole, our proposal builds on the idea that coherence means "calling out a latent concept in the mind of a reader" (Hoyle et al., 2021). By measuring the coherence of the documents within each topic or cluster, it provides a more holistic (and use-oriented) picture of a model's quality than past work. It draws most closely from the tasks in Ying et al. (2022); we adapt and combine their label assignment and validation steps, avoiding the reliance on curated expert labels.[5]

3.1 Evaluation Protocol

We describe the steps for the human evaluation protocol and LLM-proxy, PROXANN, in parallel. Appendices contain instructions, user interface screenshots (app. H), and model prompts (app. I).

Step 0: Setup. First, we outline the model outputs required for the evaluation (recall that we are attempting to emulate content analysis, Fig. 1). Throughout, we remain agnostic as possible to the method that produces these outputs; the evaluation is appropriate for both topic models and other text clustering techniques.

Suppose that there are K topics or clusters and

[4] The same logic holds for document clustering, where the interpretation of a category relies on reading the documents assigned to it.

[5] However, our approach can also use expert labels, and is complementary to their work.

$|\mathcal{D}|$ documents, with each document containing $|W_d|$ word types (total vocabulary size $|W|$). Each document $d \in \mathcal{D}$ has an estimated score indicating its semantic relationship to the kth topic or cluster, θ_{dk}. For topic models, this is the estimated posterior probability for the kth topic. Different clustering methods can produce this value in different ways; e.g., for K-means, a standard estimate is the similarity between the document embedding and the cluster centroid. We place estimates into a matrix $\boldsymbol{\Theta} \in \mathbb{R}^{N \times K}$, and each column of the matrix sorted to produce a ranked list of the most likely documents for each topic or cluster, $\boldsymbol{\theta}_k^{(r)}$.

Topics and clusters are also associated with ranked word types $\boldsymbol{\beta}_k^{(r)}$. For topic models, these are the sorted rows of the topic-word distributions $\mathbf{B} \in \mathbb{R}^{K \times |W|}$; for clustering, it is possible to extract top words for a cluster via tf-idf (Sia et al., 2020).[6]

The final representations shown to users consist of a sample of n_d highly-ranked **exemplar documents** from $\boldsymbol{\theta}_k^{(r)}$ and the most probable n_w **keywords** from $\boldsymbol{\beta}_k^{(r)}$. To balance informativeness with annotator burden, we set the number of documents n_{ex} to seven and the number of words n_w to 15.[7]

When constructing the exemplar documents, Doogan and Buntine (2021) note that only showing the documents at the head of the distribution can lead to an overly-specific view of the topic (e.g., "banning AR-15s" vs. "gun control"). We mitigate this issue by instead sampling documents with a θ_{dk} greater than a threshold t_k. To set t_k, we find the point with maximum curvature using an "elbow"-detection algorithm (Satopaa et al., 2011). Then, we sample from the set $\{d : \theta_{dk} > t_k\}$, where the probability of a sample is proportional to θ_{dk}. Figure 7 (in the appendix) shows the distributions of $\boldsymbol{\theta}_k^{(r)}$ for the 1,000 documents with the largest values over six topics for the two topic models we use (see Section 4).

Step 1: Category identification. After viewing instructions and completing a training exercise (Appendix H), each annotator reviews the exemplar documents and keywords for a single topic. They then construct a free-text **label** that best describes the category they have observed.[8] Continuing the earlier U.S. healthcare example, users might also view the text of the *National Organ Transplant Act of 1984* and the *Rare Diseases Act of 2022*.

The LLM is prompted with condensed instructions and the same exemplars and keywords, also producing a label for the category.

Step 2: Relevance Judgment. An additional sample of seven **evaluation documents**, evenly stratified over $\boldsymbol{\theta}_k^{(r)}$, is shown in random order.[9] For one document at a time, annotators answer the extent to which the document fits their inferred category (on a scale from "1 – No, it doesn't fit" to "5 – Yes, it fits"), producing a set of **fit scores** for annotator i, $\boldsymbol{s}_k^{(i)}$. As a control, one document with near-zero probability for the topic is always shown. Here, an annotator might assign the *Coronavirus Preparedness and Response Act* a "5" and the *Federal Meat Inspection Act* a "3".

For the LLM prompt, the instructions are slightly modified to produce *binary* fit scores, given that models are known to face issues with Likert-like scales (Stureborg et al., 2024).

Step 3: Representativeness ranking. Last, annotators rank the evaluation documents by how representative they are for that category, $\boldsymbol{r}_k^{(i)}$.[10]

Given the complexity of the task, a direct translation to an LLM prompt is not practical. Instead, we modify the question to include two evaluation documents at a time, leading to $\binom{7}{2}$ prompts. The LLM thus produces a set of pairwise ranks per prompt, which we use to infer real-valued "relatedness" scores for each document with a Bradley and Terry model (further details in Appendix I).

		α Fit (Step 2)	α Rank (Step 3)
Wiki	Mallet	0.71 (0.10)	0.74 (0.12)
	CTM	0.55 (0.30)	0.45 (0.11)
	BERTopic	0.57 (0.16)	0.44 (0.20)
Bills	Mallet	0.31 (0.27)	0.49 (0.22)
	CTM	0.37 (0.19)	0.43 (0.26)
	BERTopic	0.32 (0.30)	0.34 (0.17)

Table 2: Chance-corrected human–human inter-annotator agreement (Krippendorff's α), averaged over eight topics per model (standard deviation in parentheses). Each topic has at least 3 annotators.

[6] Ranked word types are not strictly necessary for the evaluation, but their usage as a topic summary is widespread.

[7] See Lau and Baldwin 2016 for a discussion of the relationship between n_w and perceived coherence.

[8] Per Chang et al. (2009), documents are truncated to improve reading times. We limit them to 1000 characters.

[9] Generally, we assume a strict total ordering over evaluation documents; nonstrict orders, as in the case of binary assignments $\theta_{dk} \in \{0, 1\}$, can work but require some alterations to our metrics.

[10] We include a "distractor" document—an Amazon review for kitchen sponges—to filter out poor quality annotations.

4 Experiments

We describe the experimental setup: the choices of datasets, models, annotators, and metrics.

4.1 Datasets

We use two English datasets that are standard in topic modeling evaluations: `Wiki` (Merity et al., 2017), consisting of 14,000 "good" Wikipedia[11] articles; and `Bills` (Adler and Wilkerson, 2008), comprising 32,000 legislative summaries from the 110th–114th U.S. Congresses. We use the preprocessed version of these datasets from Hoyle et al. 2022, in its 15,000-term vocabulary form.

4.2 Models

Topic Models Topic models can be broadly categorized into *classical* Bayesian methods, which use Gibbs sampling or variational inference to infer posteriors over the latent topic-word ($\boldsymbol{B}$) and document-topic ($\boldsymbol{\Theta}$) distributions, and *neural* topic models, often estimated with variational autoencoders (Kingma and Welling, 2013). Clustering techniques can also approximate topic models; in a typical setup (e.g., Zhang et al., 2022), K-means is applied to sentence embeddings (Reimers and Gurevych, 2019) of the documents.[12]

We evaluate one model from each class: LDA (Blei et al., 2003) using the MALLET implementation (hereafter referred to as **MALLET**), **CTM** (Bianchi et al., 2021), and **BERTopic** (Grootendorst, 2022). We reuse the 50-topic MALLET and CTM models from Hoyle et al. 2022 and train BERTopic under the same experimental setup using default hyperparameters (details in Appendix F). In a pilot study, we also evaluate a synthetic upper bound model derived from ground-truth `Wiki` labels (Appendix C).

PROXANN LLMs We employ OpenAI's GPT-4o (`gpt-4o-mini-2024-07-18`) and 8-bit quantized Llama3.1 (`Llama3.1:8B`) as LLM annotators. We set the temperature to 0, `top_p` to 0.1, and `frequency_penalty` to 0. Documents exceeding 100 tokens are truncated, extending to the end of the sentence to avoid incomplete cuts. The Step 1 prompt is consistent across both models, while Steps 2 and 3 prompts are optimized independently per model with DSPy (Khattab et al., 2024) using training samples derived from pilot annotation data on `Wiki` (details in Appendix I).

		Label Sim.	Fit Acc.	Rank τ
`Wiki`	GPT-4o	95%	79%	89%
	Llama3.1:8B	89%	68%	58%
`Bills`	GPT-4o	96%	100%	100%
	Llama3.1:8B	83%	96%	79%

Table 3: Share of topics where the agreement between PROXANN and human annotators is not significantly different than the agreement between humans, across the 3 protocol steps ($p < 0.05$ in a permutation test). For the most part, GPT-4o is a reasonable proxy.

4.3 Collecting Human Annotations

A comprehensive human evaluation of all topics would be cost-prohibitive, so we randomly sample 8 of the 50 topics for the `Wiki` and `Bills` data on each of the three models. We recruit at least 4 annotators per topic through Prolific.[13] Low-quality respondents are filtered out using attention checks.

Model-to-model results on a subset of topics may not be comparable; when sampling, we first pick a random topic from one model, and choose the topics from the remaining models with the smallest word-mover's distance (computed using word embeddings of the topic words, Kusner et al., 2015; Flamary et al., 2021).

4.4 Metrics

We examine four aspects of our approach: the sensibility of the human evaluation protocol; using the protocol to evaluate topic models and clustering; comparing human annotations with the LLM proxy; and using metrics based on the LLM proxies to score topics and clusters.

Human–human agreement on the tasks. Following standards from the content analysis literature, we use Krippendorff 's α to assess the chance-corrected agreement across human annotators for Steps 2 and 3 (with ordinal weights).[14] For easier

[11] `https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria`

[12] Recently, LLM-based topic models (Pham et al., 2024; Lam et al., 2024) offer more "human-readable" topic descriptions, but lack the document-topic and word-topic distributions that other methods provide or approximate. Given these differences, we leave an evaluation to future work.

[13] `prolific.com`, further recruitment details in Appendix A. We also run an initial pilot study on a synthetic upper-bound model based on ground-truth labels (comparing with CTM and MALLET); high agreement on the upper-bound validates that our tasks are reasonable, details and results in Appendix C.

[14] Although it seems natural to use these metrics for topic model comparisons—higher agreement indicating better topics or clusters—there are complications arising from skewed distributions and respondents annotating one topic at a time,

comparison with the model–human metrics (next section), we also compute annotator-to-annotator correlations between each annotator's relevance fit scores (Step 2) or ranks (Step 3) and the averaged fits (ranks) of all other annotators.

Human evaluation of topics and clusters. Per Section 3.1, models estimate real-valued scores θ_{dk} that (should) correspond to the relevance that document d has for category k. In steps 2 and 3, annotators assess the relevance of seven documents over a stratified set of these scores for a topic k, $\boldsymbol{\theta}_k^{\text{eval}}$ (all annotators review the same documents).

As a measure of model quality, we report the correlation coefficients for Kendall's τ (Kendall, 1938) to measure both annotator-model and inter-annotator relationships. The annotator-model correlations are between the estimated probabilities per document $\boldsymbol{\theta}_k^{\text{eval}}$ with either the human relevance scores ($\boldsymbol{s}_k^{(i)}$, Step 2) or their ranks ($\boldsymbol{r}_k^{(i)}$, Step 3), where i is the annotator. We contextualize these against the inter-human-annotator τ (see above).

ProxAnn–human agreement. For the LLM to serve as a proxy, it should be indistinguishable from a human annotator. We operationalize this criterion by computing agreement metrics for the results of the three steps: if the human-to-LLM agreements are significantly worse than human-to-human agreements, then the LLM is not a reasonable proxy. For Step 1, the agreement is the cosine similarity between sentence embeddings of the produced category labels (`all-mpnet-base-v2`). For Step 2, we compute raw agreement (accuracy) over the category relevance judgments for each of the seven evaluation documents (binarizing answers ≥ 4 to match the LLM outputs). Step 3 produces rankings over the evaluation documents, so agreement is Kendall's τ.

The pairwise agreements between humans and ProxAnn are computed for each topic with more than four annotators. Statistical significance is computed via a one-sided permutation test: if the *observed* difference in ProxAnn–human and human–human mean agreement is significantly smaller than the difference when randomly permuting "ProxAnn" and "human" labels, then ProxAnn is an inadequate substitute.

ProxAnn as an automated evaluator. A common use for automated coherence metrics, like NPMI, is the ranking of topics—and the averaging of topics within each model to rank models. Indeed, NPMI is the dominant metric used in the literature to compare proposed models against baselines (Hoyle et al., 2021).

Here, we compare the human evaluations of topics and clusters to metrics based on ProxAnn. Specifically, the evaluation metrics are those described above: the correlations τ between (a) the estimated document scores $\boldsymbol{\theta}_k^{\text{eval}}$) from the topic model (or clustering algorithm) and (b) the responses to the protocol—relevance fits $\boldsymbol{s}_k$ (Step 2) or ranks $\boldsymbol{r}_k$ (Step 3), from either ProxAnn or averaged over human annotators. Hence, for each topic and task, there is a "ground-truth" evaluation metric (the τ between human scores and the topic model scores) and a "proxy" metric (the τ between ProxAnn and the topic model scores). We can then compute an additional Kendall's τ over these metrics to measure the extent to which ProxAnn's rankings over topics agrees with that of the average human.

		Wiki		Bills	
		Human Fit	Human Rank	Human Fit	Human Rank
Cohr.	NPMI	0.029	-0.122	-0.073	-0.058
Model Fit	`Gpt-4o`	0.455	0.182	0.075	0.317
	`Llama3.1`	0.327	0.265	-0.016	-0.008
Model Rank	`Gpt-4o`	0.452	0.448	0.101	0.316
	`Llama3.1`	0.234	0.457	0.167	0.227

Table 4: Relationship between automated and human-based metrics. Each cell shows Kendall's τ correlation between metrics: *Human Fit* and *Human Rank* compare human fit scores and ranks to document-topic probabilities ($\boldsymbol{\theta}_k$); *Model Fit* and *Model Rank* comparesProxAnn fit scores and ranks to $\boldsymbol{\theta}_k$.

5 Results

We discuss results in the same order they were presented above. Note that in tables and figures, **Fit** refers to responses to Step 2 (relevance judgments of evaluation documents) and **Rank** to responses to Step 3 (representative rankings of the documents).

5.1 Human–Human Agreement

Generally, annotators respond consistently, providing qualitatively sensible labels to the topics (Table 1). Average agreement per topic (Krippendorff's α) is reasonably strong overall, particularly for the ranking tasks on the `Wiki` data (Table 2). We emphasize that *low* agreement is likely indicative of a poor model, rather than a misspecified task:

Appendix D.

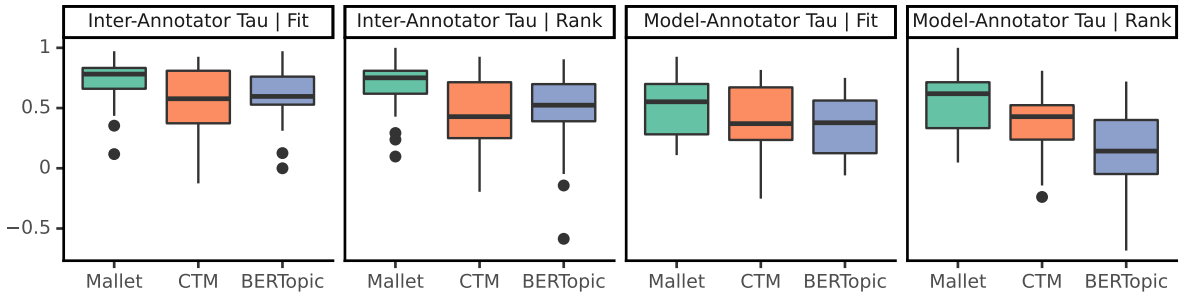


Figure 2: Metrics quantifying the relationship between human relevance judgments and estimated document-topic probabilities ($\boldsymbol{\theta}_k$) for two topic models and a clustering model on the `Wiki` data. From left-to-right, the metrics are inter-annotator Kendall's τ and model-annotator τ for the human relevance judgments (on a 1-5 scale, Step 2); then the same two metrics for their document representativeness rankings (Step 3). Boxplots report variation over topic-annotator pairs.

in Table 5 (appendix), the agreement metrics for a synthetic "upper-bound" model are very strong ($\alpha \geq 0.8$ on both tasks). Overall, MALLET tends to have higher agreement; however, variance over topics is somewhat high, and we caution against using α for model comparisons. Together, these results point to the viability of our evaluation protocol, implying that the demands of the tasks are intelligible and reproducible.

5.2 Human Evaluations of Topics

Our protocol creates consistent and sensible results. There is generally a positive correlation between the estimated document-topic probabilities ($\boldsymbol{\theta}_k$) and human judgments on the `Wiki` data (Fig. 2, `Bills` data in Fig. 8 in the appendix). Comparing the first two plots (human–human) to the second two (human–model), annotator agreement with other annotators is generally higher than than annotator agreement with the model. Both the inter-annotator and model–annotator scores show a consistent ranking over models: MALLET fares better than CTM, and CTM better than BERTopic—in fact, several topics have negative correlations for BERTopic. In Appendix B, we report on two additional metrics, NDCG and binarized agreement.

These results support the idea that MALLET, despite being 20 years old, remains an effective tool for automated content analysis.

5.3 Is PROXANN a good proxy?

Generally, GPT-4o is a reasonable proxy across the three steps and both datasets. Llama-3.1 fares somewhat worse, particularly for the ranking task on the Wiki data—Section 4.3 shows the share of topics, per step, where PROXANN does not have

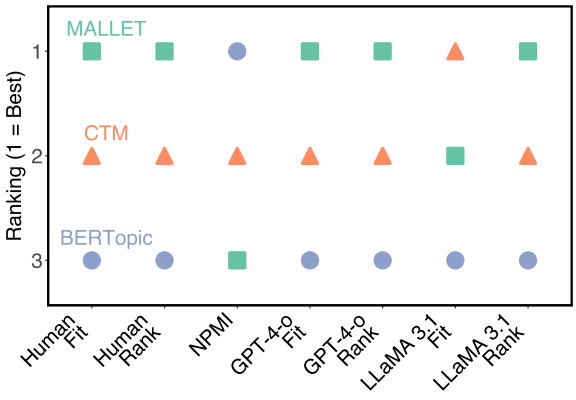


Figure 3: Model rankings based on human-derived and automated metrics. MALLET ranks highest among humans. Rankings based on NPMI deviate from human and LLM-based metrics, with LLM metrics generally aligning better with humans.

significantly poorer agreement with human annotators than humans do with each other. Generally, PROXANN-GPT-4O fails on 1-3 topics in each step. Fig. 4 characterizes its performance on the `Wiki` data for the ranking task. On most topics, the difference between human–human agreement and model–human agreement is indistinguishable.

Lower performance on the Bills data may be attributed to (a) a more specialized dataset requiring additional background knowledge and (b) having tuned prompts on pilot annotations from the `Wiki` data. We plan to explore the effect of expertise and data-specific tuning in follow-up work.

5.4 Ranking Topics and Models

Last, we measure the ability of metrics derived from PROXANN to rank topics and models similarly to humans. Generally, GPT-4o Step 3-based

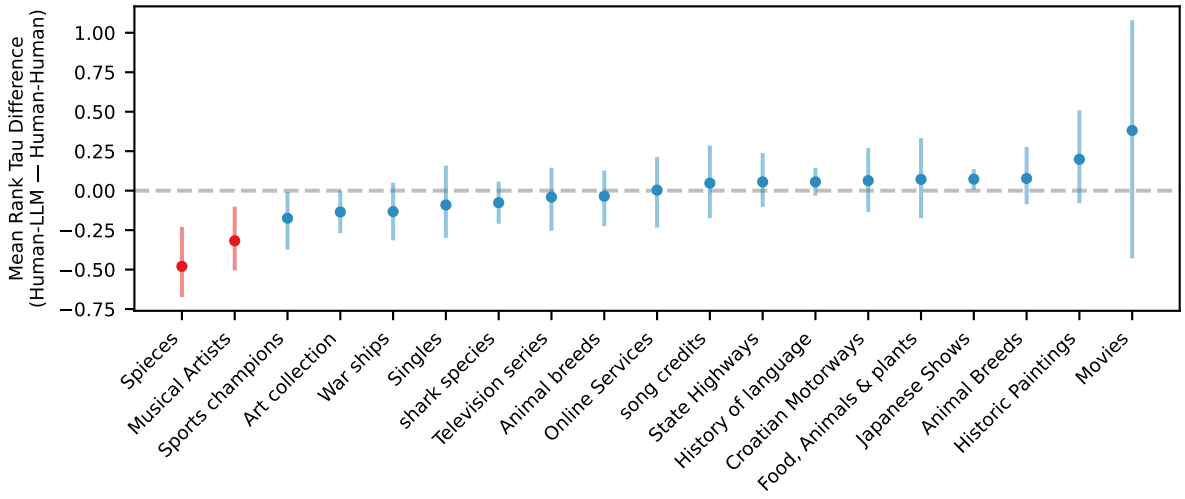


Figure 4: Mean difference in ProxAnn–human and human–human ranking agreement (Step 3) for the `Wiki` data across topics with $>= 4$ annotators (bootstrapped 95% CIs). On most topics, the LLM annotator (ProxAnn-GPT-4o) is not distinguishable from a random human. Red labels have sig. lower ProxAnn–human agreement; topic labels are the shortest available.

metrics tend to be best, with a 0.46 correlation for `Wiki` and 0.32 for `Bills`. While not very high, these values are comparable to leaving out one *human* annotator and computing their agreement with the average of the other humans (e.g., the mean `Wiki` Rank τ is 0.34). Meanwhile, the standard automated metric, NPMI, fails to capture the human judgments. Aggregating the scores over models shows that the ProxAnn metrics produce a more reliable ranking than they do for individual topics, mostly matching human-derived ranks (Fig. 10).

6 Prior Work

Use-oriented evaluations Poursabzi-Sangdeh et al. 2016 and Li et al. 2024 invoke topic models' usage in content analysis settings to inform new interactive methods, which are evaluated by measuring the alignment between method outputs and ground-truth labels. In a different use-inspired approach, Ying et al. 2022 propose crowdworker "label validation" tasks, designed to assess the quality of individual document-topic distributions using already-identified expert labels. Although these evaluations are better aligned with real-world use than topic coherence, they rely on some form of manual labeling, and are therefore difficult to scale.

LLM-based evaluations. Metrics based on LLMs have become increasingly common in the NLP literature, notably in machine translation and human preference modeling (Zheng et al., 2023). Within topic modeling, past efforts construct prompts designed to replicate human annotation tasks. Both Stammbach et al. 2023 and Rahimi et al. 2024 prompt LLMs to emulate the word intrusion and rating tasks from Chang et al. 2009, but these tasks assess only the top topic-words, an incomplete view of model outputs. In addition, the correlations with human judgments are also mixed, with standard automated coherence metrics performing better in some cases.[15] In Yang et al. 2024, a topic model and an LLM separately produce keywords to label documents: if the keywords tend to align, then this indicates a good model. Although LLM keywords align well with human-generated ones for one of two datasets, the metric does not assess the overall cohesiveness of topics, and so the connection between this task and real-world use is unclear.

7 Conclusion

The quality of models is determined their ability to meet real-world needs (Liao and Xiao, 2023). This work aims to meet those needs by designing a human evaluation protocol and corresponding automated approximation, ProxAnn that together reflect practitioners' real-world usage of topic models and clustering methods. We anticipate that both the collected human evaluation data and automated approach will inspire future work in improving models, metrics, and downstream usage.

[15] Stammbach et al. 2023 also propose an alternative document-labeling metric, but it is used for selecting an optimal number of topics, rather than measuring overall quality.

8 Limitations

A primary limitation of our LLM-proxy is that it is a substitute for a *single* human annotator. However, a strong indicator of a poor cluster or topic is disagreement among *multiple* annotators. In future work, we intend to model disagreement directly, e.g., following recent approaches for finetuning reward models in the presence of human disagreement (Zhang et al., 2024b), or earlier work on Bayesian models of annotation (Paun et al., 2018). Addressing this issue could also help solve another limitation: LLMs are more costly to deploy than previous automated metrics, but a model finetuned for this task could be smaller.

Another shortcoming of our approach is the use of crowdworkers. Although we use several mechanisms to ensure high-quality annotators (training questions, multiple comprehension and attention checks, requiring a bachelor's degree or higher, bonuses for good responses), the annotators are not experts pursuing a research question. That said, we believe our use of multiple annotators per topic, along with the filtering described, ensures annotations of reasonably high quality (as seen by the consistent labels and annotations). In future work, we hope to explore the role of expertise in the annotation process, and to measure expert agreement with language models.

References

E. Scott Adler and John Wilkerson. 2008. Congressional Bills Project. http://www.congressionalbills.org. Accessed: insert access date here.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18.

Aneesha Bakharia, Peter Bruza, Jim Watters, Bhuva Narayan, and Laurianne Sitbon. 2016. Interactive topic modeling for aiding qualitative content analysis. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, page 213–222, New York, NY, USA. Association for Computing Machinery.

Eric Ps Baumer, David Mimno, Shion Guha, Emily Quan, and Geri Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68.

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.

Jordan L. Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models. In *Handbook of Mixed Membership Models and Their Applications*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345. Publisher: [Oxford University Press, Biometrika Trust].

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

Anne Corden, Roy Sainsbury, et al. 2006. *Using verbatim quotations in reporting qualitative social research: researchers' views*. University of York York.

Barbara Di Eugenio and Michael Glass. 2004. Squibs and discussions: The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, AurÃ©lie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, LÃ©o Gautheron,

Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Barney G. Glaser and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, 1st edition. Aldine Publishing, Chicago.

Justin Grimmer and Brandon Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267 – 297.

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

K.L. Gwet. 2012. *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic evaluation broken? the incoherence of coherence. In *NeurIPS (Spotlight Presentation)*.

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.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, chapter 14. SAGE Publications, Inc.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

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 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–487, San Diego, California. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Zongxia Li, Andrew Mao, Daniel Stephens, Pranav Goel, Emily Walpole, Alden Dima, Juan Fung, and Jordan Boyd-Graber. 2024. Improving the TENOR of labeling: Re-evaluating topic models for content analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 840–859, St. Julian's, Malta. Association for Computational Linguistics.

Q. Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv (2306.03100)*.

Jia Peng Lim and Hady W. Lauw. 2024. Aligning Human and Computational Coherence Evaluations. *Computational Linguistics*, pages 1–58.

Philipp Mayring. 2000. Qualitative inhaltsanalyse. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 1(2).

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

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.

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. ALTO: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1158–1169, Berlin, Germany. Association for Computational Linguistics.

Hamed Rahimi, David Mimno, Jacob Hoover, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2024. Contextualized topic coherence metrics. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1760–1773, St. Julian's, Malta. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ville A. Satopaa, Jeannie R. Albrecht, David E. Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171.

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.

Charles P Smith. 2000. Content analysis and narrative analysis. *Handbook of research methods in social and personality psychology*, 2000:313–335.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.

Steve Stemler. 2000. An overview of content analysis. *Practical assessment, research, and evaluation*, 7(1):17.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Shu Xu and Michael F. Lorber. 2014. Interrater agreement statistics with skewed data: evaluation of alternatives to cohen's kappa. *Journal of consulting and clinical psychology*, 82 6:1219–27.

Xiaohao Yang, He Zhao, Dinh Phung, Wray Buntine, and Lan Du. 2024. Llm reading tea leaves: Automatically evaluating topic models with large language models.

Luwei Ying, Jacob M. Montgomery, and Brandon M. Stewart. 2022. Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. *Political Analysis*, 30(4):570–589.

Bolun Zhang, Yimang Zhou, and Dai Li. 2024a. Can human reading validate a topic model? *Sociological Methodology*.

Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2024b. Diverging preferences: When do annotators disagree and do models know? *Preprint*, arXiv:2410.14632.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Annotator Recruitment

Annotators must be fluent in English and have a college degree or higher. Given the western-centrism of the English `Wiki` data respondents must be located in the U.S., Canada, Ireland, or U.K.; for the U.S.-centric `Bills` data, we exclude those outside North America. We recruit at least 4 annotators per topic using Prolific. Demographic information is not made available to us, and we retain no identifying information. Annotators were presented with information about the nature of the task and asked to provide consent before participation. We set pay at a 15 USD per hour equivalent (`Wiki` completion time was estimated at 15 minutes, paying 3.75 USD

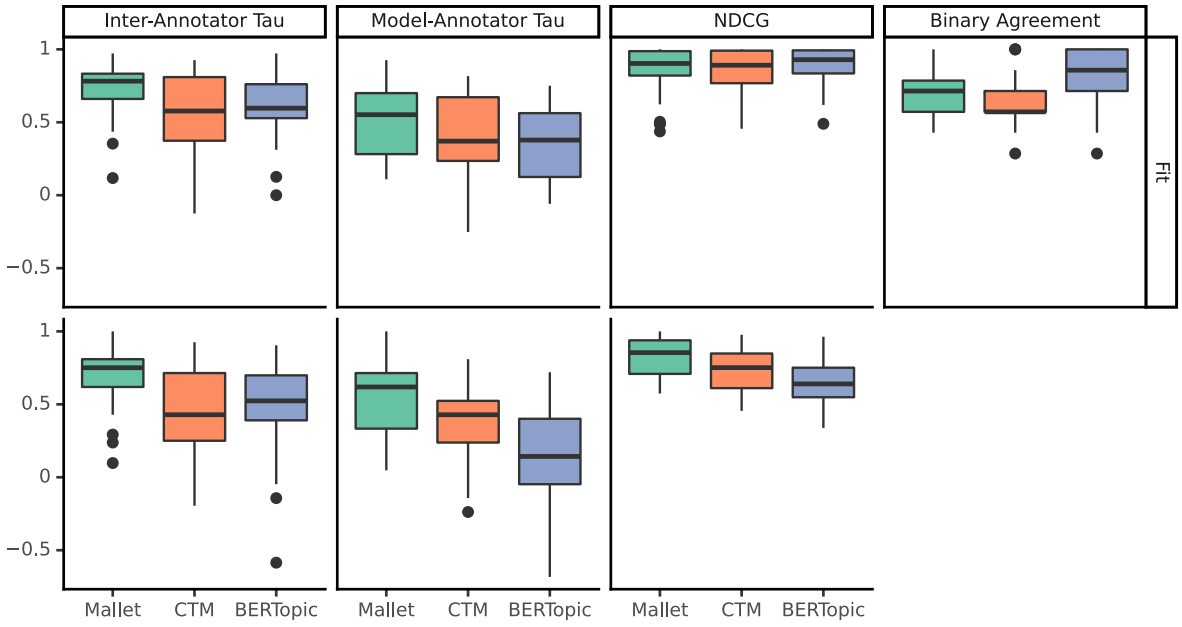


Figure 5: Correlation metrics between human relevance judgments and estimated document-topic probabilities ($\boldsymbol{\theta}_k$) for the three models on all eight `Wiki` topics. From left-to-right, the metrics are inter-annotator Kendall's τ, model-annotator τ, relevance agreement, and NDCG. The top row of figures reports relationships with human relevance judgments (on a 1-5 scale), and the bottom row relationships with their document rankings. Boxplots report variation over topic-annotator pairs. We emphasize that the "Labeled" model is not a true topic model, but a synthetic supervised benchmark with access to ground-truth categories.

per survey; `Bills` was updated to 4.25 for 17 minutes). To encourage careful responses, we instruct annotators to "give the answers you think most other people would agree with", awarding a 1.50 USD bonus to those who have over 0.75 correlation with the average ranking of the other annotators for that topic. Annotators who fail attention checks are not awarded a bonus and are excluded from the data. An ethics review board deemed this study to not be human subjects research, and therefore exempt from review.

B Additional Metrics

In this section, we report on additional measures for the **human evaluations of topics** (Section 4.4).

We use Normalized Discounted Cumulative Gain (NDCG, Järvelin and Kekäläinen, 2002), a well-established IR metric that places more importance on items with higher ranks. NDCG is designed to average over multiple user annotations and queries (here corresponding to topics).

Last, we also report the raw agreement over binarized relevance. For the human scores, we consider any documents where the fit to the category is 4 or 5 to be relevant. For the models, a document is considered to be relevant to a topic k if its most probable topic is k. The agreement is then the proportion of relevance judgments in common.

Results are in Fig. 5 and Fig. 8—of note is that BERTopic cluster assignments tend to have higher agreement with human relevance jugments (binarized responses to Step 2), likely due to it being a clustering model.

	Fit α (Step 2)	Rank α (Step 3)
Mallet	0.59 (0.16)	0.71 (0.09)
CTM	0.64 (0.15)	0.67 (0.13)
Labeled	0.80 (0.13)	0.86 (0.05)

Table 5: Chance-corrected human–human agreement (Krippendorff's α), averaged over the six pilot topics per model (standard deviation in parentheses) on the `Wiki` data. Each topic has between 3 and 5 annotators (the variance is due to filtering). High agreement on the synthetic labeled dataset indicates that the task is sensible.

C Pilot Study

We first run a pilot annotation study on using the `Wiki` data on six topics from CTM and MALLET.

To help validate the sensibility of the human evaluation protocol, we also introduce an informal upper-bound, we evaluate a synthetic model

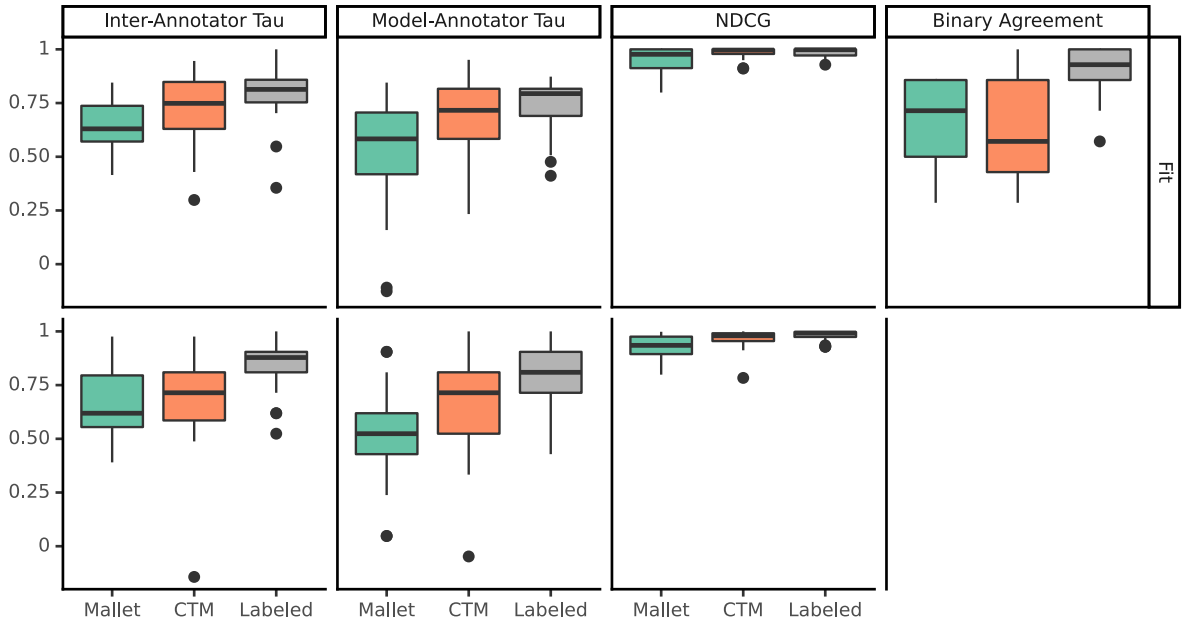


Figure 6: Metrics quantifying the relationship between human relevance judgments and estimated document-topic probabilities ($\boldsymbol{\theta}_k$) for two models and a synthetic upper-bound, using six topics from the pilot data. From left-to-right, the metrics are inter-annotator Kendall's τ, model-annotator τ, relevance agreement, and NDCG. The top row of figures reports relationships with human relevance judgments (on a 1-5 scale), and the bottom row relationships with their document rankings. Boxplots report variation over topic-annotator pairs. We emphasize that the "Labeled" model is not a true topic model, but a synthetic supervised benchmark with access to ground-truth categories.

(termed LABELED) using ground-truth category labels for the `Wiki` data. For each label in data, take the documents assigned to the label k and embed them (using the same embedding model as CTM). To construct a pseudo-ranking over documents for the topic, $\tilde{\boldsymbol{\theta}}_k$, we calculate the cosine similarity between the document embeddings (for all documents) and the centroid of all k-labeled documents. We further correct the similarities for the kth label by adding 1 to all the k-labeled documents, ensuring that they are ranked above those that are not labeled for the document. Synthetic top words for the topic are found by concatenating all k-labeled documents and computing the tf-idf for this pooled "document". The result is that all *exemplar documents* are known to relate to a single ground-truth label (e.g., video games).

Results show that both inter-annotator and model-annotator agreement metrics are substantially higher for the synthetic model, Table 5. Of particular note are the binary agreement scores (Fig. 6, implying that human annotators agree with a ground-truth assignment at very high rates.

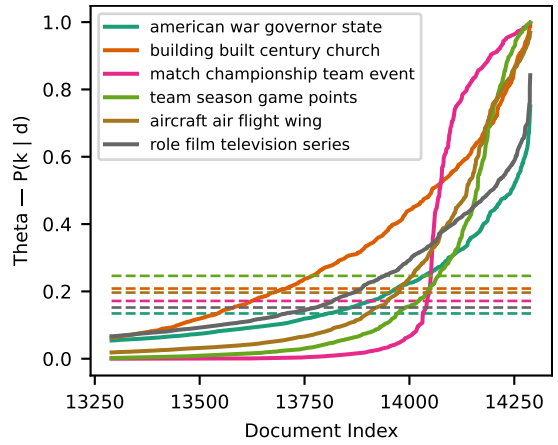
The resulting annotation data is used to help tune the LLM prompts in Appendix I.

D Notes on Agreement Metrics

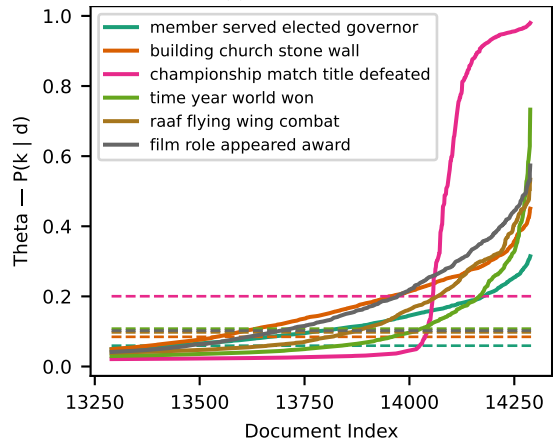
The most straightforward way to assess relative model performance using the human annotations is to compute the chance-corrected inter-annotator agreement—indeed, this corresponds most closely to the way a manual qualitative content analysis is assessed. A topic with high agreement across annotators is likely to be better than one with low agreement. However, the idea is complicated by annotators only viewing one topic each. Measures like Krippendorff's α (Krippendorff, 2019) use the empirical distributions to estimate expected agreement when correcting for chance, so a topic with relatively high raw agreement (i.e., a very skewed distribution) may have a low value relative to what is qualitatively considered a "good" topic.[16] While it is possible to average these values over topics, their occasionally counter-intuitive nature makes them less desirable for model comparison.

[16] There is extensive literature on this issue (Di Eugenio and Glass, 2004; Gwet, 2012; Xu and Lorber, 2014). Nonetheless, in the political science community, Krippendorff's α and Cohen's κ remain essentially universal. As far as we can tell, this is also true more broadly in the social sciences.

E Visualizing the Document-Topic Distribution



(a) MALLET



(b) CTM

Figure 7: Distribution of the top 1,000 theta values across six topics for two models. Topics have been aligned between models based on the word-mover's distance (Kusner et al., 2015). Dashed lines correspond to automatically determined "elbows" that threshold the θ_k to produce representative documents. Some topics, like the championship topic (in pink), have a sparser distribution and steep dropoff in values; others, like the building topic (orange), have a more gradual decline in value.

In Section 3.1, we outlined a method for selecting the **exemplar documents** based on finding a knee point in the document-topic distributions $\mathbf{\Theta}$. In Fig. 7, we visualize these distributions for CTM and MALLET for the pilot topics alongside the detected threshold. Documents above this threshold are sampled (proportional to θ_{dk} to produce the exemplar documents .

F BERTopic training details

Although the BERTopic author advises against data preprocessing[17], we apply the same minimal preprocessing used for training MALLET and CTM models (tokenization and entity identification) to ensure comparable conditions (we also find that, qualitatively, topics are better after preprocessing). Contextualized embeddings are generated separately using the raw (i.e., unprocessed) text and BERTopic's default embedding model (`all-MiniLM-L6-v2`). The preprocessed data and pre-calculated embeddings are then passed to the model. We train BERTopic with `calculate_probabilities=True` to compute topic probabilities for each document during the HDBSCAN clustering step. Due to the hard–clustering nature of HDBSCAN, the resulting approximation of document-topic distribution ($\mathbf{\Theta}_{\text{t}}^*$) often assigns a value of 1 to documents confidently associated with a cluster, while other probabilities remain close to zero. To generate a smoother document-topic distribution to obtain the evaluation documents, we combine $\mathbf{\Theta}_{\text{a}}^*$ with probabilities derived from BERTopic's `approximate_distribution` function ($\mathbf{\Theta}_{\text{a}}^*$), which uses c-TF-IDF representations to estimate topic probabilities for new documents. The final distribution is computed as $\mathbf{\Theta}^* = \text{round}(\mathbf{\Theta}_{\text{t}}^*, 2) + \mathbf{\Theta}_{\text{a}}^*/100$.

G Additional Bills Results

Figure 8, Fig. 9, and Fig. 10 depict evaluations on the `Bills` data, corresponding to Fig. 5, Fig. 4, and Fig. 3 in the main text.

H User Interface

Figures 11 to 14 are screenshots of the annotation interface presented to users. Figure 15 is the consent page shown at the start.

I Prompting details

Here, we outline our prompt engineering process used to configure the LLM-based proxy for the evaluation protocol. For the three steps of the evaluation protocol, we use a concise system prompt (I.1–I.3) to summarize the tasks and instruct the LLM to simulate human-like behavior, combined with an instruction prompt (I.4–I.6) that provides

[17] https://maartengr.github.io/BERTopic/faq.html#should-i-preprocess-the-data

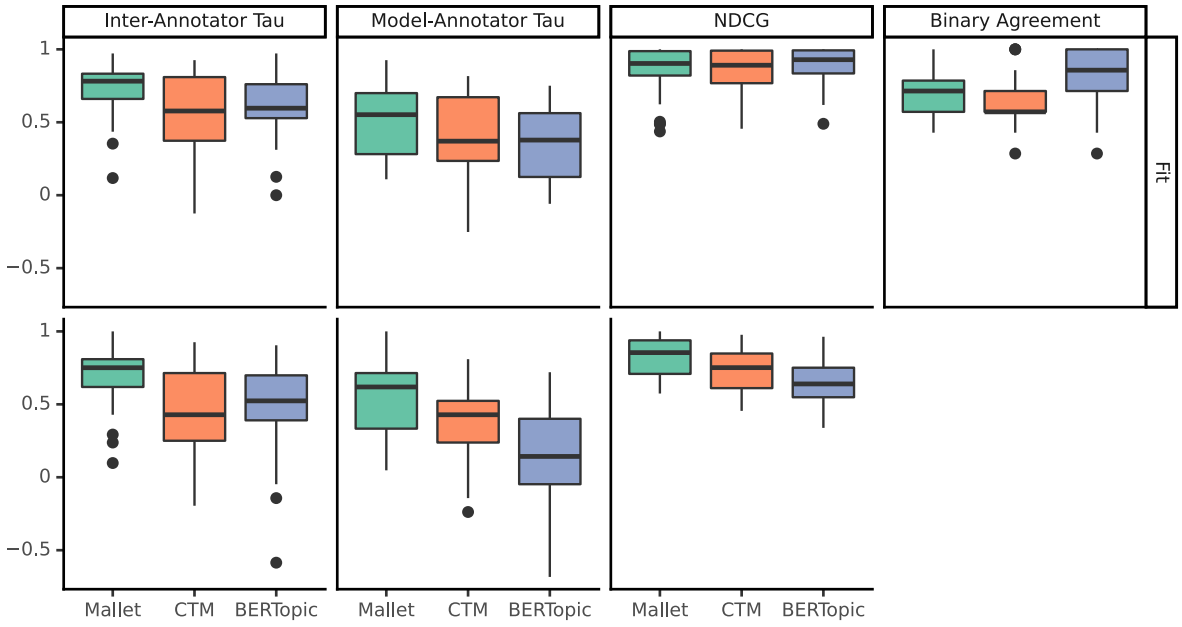


Figure 8: Metrics quantifying the relationship between human relevance judgments and estimated document-topic probabilities ($\boldsymbol{\theta}_k$) for three models on all eight `Bills` topics. See Fig. 5 for additional details.

detailed guidance for completing each task. In all cases, we use few-shot prompting in the instruction prompt. LLama models were run on an NVIDIA 4090 (24 GB RAM); prompting the models for all topics takes under two hours.

I.1 Prompt optimization

The *Step 1* prompt was manually optimized by the authors. In its final version, the LLM is provided with documents and keywords related to a topic cluster and tasked with identifying their shared category. While we experimented with variations, such as asking the LLM to generate a brief description of the label and its key characteristics, we retained the original version as it demonstrated better performance. This prompt includes a single few-shot example, sourced from the pilot study, and its the same for both LLM types (GPT-4o and Llama3.1).

Step 2 and *Step 3* prompts were optimized independently for each LLM type using `DSPy`(Khattab et al., 2024), employing a $50/25/25$ train-test-validation split on the pilot data, with the optimizer bootstrapping up to 4-shot examples and 16 candidate programs during random search. Let $\mathbf{f}_i$ represent the fit scores assigned to document i by all users. Let $\mathbf{f}_i$ and Let $\mathbf{r}_i$ be the fit and rank scores assigned to document i by all users. The datasets for each step were generated as follows:

- *Step 2:* For each topic, we create one sample per evaluation document and user category label. The fit score for each document was determined by averaging across users and binarizing as 1 if $\bar{\mathbf{f}}_i \geq 4$ and 0 otherwise. This process resulted in a total of 566 samples.
- *Step 3:* For each topic, we generate all possible evaluation document combinations, creating one sample per pair and user category. For each document pair (i, j), we computed differences in ranks ($\Delta \boldsymbol{r} = \mathbf{r}_i - \mathbf{r}_j$) and fits ($\Delta \boldsymbol{f} = \mathbf{f}_i - \mathbf{f}_j$). Rank / fit agreement holds if Δ is consistent across users. If agreement exists, the pairwise winner is the document with the higher rank score. The resulting dataset contains a total of 1701 samples.

The final prompts (I.4–I.6) include a placeholder that consists of a User message–Assistant message combination, where two examples are provided as input to guide the model's response generation.

We also experimented with combining *Step 1* with *Step 2* (*Step 1 & 2*) and *Step 1* with *Step 3* (*Step 1 & 3*) to evaluate whether placing the LLM in conditions more similar to those faced by humans could improve performance. While the results for *Step 2 & 3* were particularly promising, we ultimately retained the independent version due to its slightly better overall performance.

I.2 Bradley-Terry

After applying the *Step 3* prompt to each topic on all $\binom{7}{2}$ combinations of evaluation document

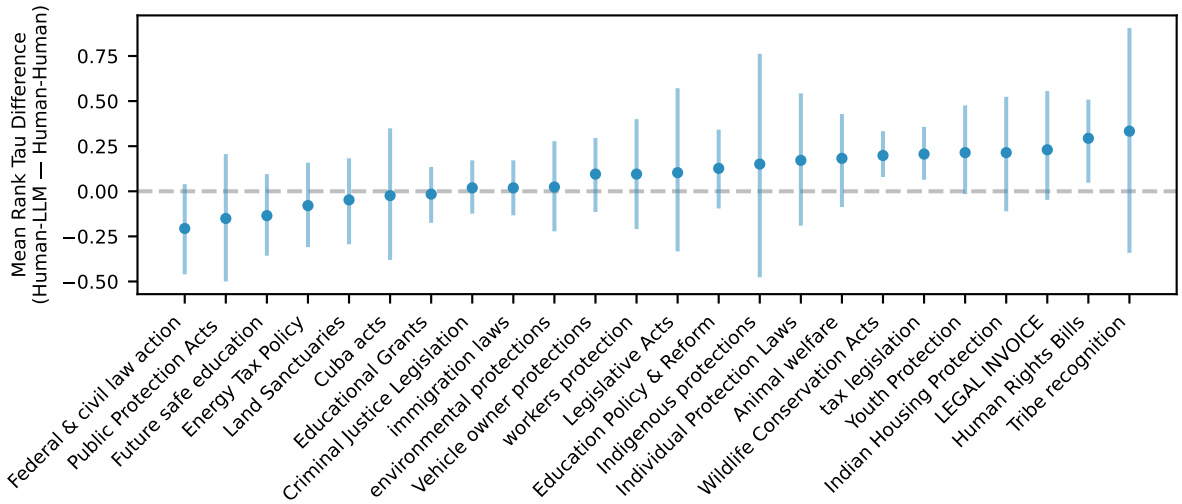


Figure 9: Mean difference in LLM–human and human–human ranking agreement (Step 3) for the `Bills` data across topics with $>= 4$ annotators (bootstrapped 95% CIs). Red labels have sig. lower LLM–human agreement; topic labels are the shortest available.

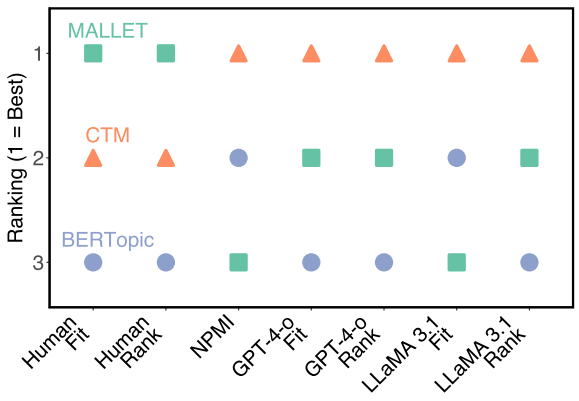


Figure 10: Model rankings based on mean Human Fit/Rank metrics, NPMI coherence, and Fit/Rank LLM-metrics (GPT-4O, LLAMA3.1) across topics for the `Bills` data.

pairs, we infer the real-valued "relatedness" for the topic by aggregating pairwise comparisons using the Iterative Luce Spectral Ranking (ILSR) algorithm. To compute the rankings, we use the implementation from the `choix`[18]library, applying the `ilsr_pairwise` method, setting the regularization term α to 0.001 to ensure numerical stability and prevents overfitting in cases where the comparison graph is not fully connected.

To ensure a fair comparison in the prompt, evaluation documents are referred to as A and B to avoid biasing the model (e.g., implying significance based on numerical identifiers). However, this approach may still introduce a preference for one letter over the other. To mitigate this, we implemented a "both-ways" approach, running the prompt twice for each document pair: once with the first document as A and the second as B, and vice versa (following Wang et al. 2024). Evaluation of the results showed that this bidirectional method did not improve the overall rankings, as the models demonstrated no systematic letter-based bias. Consequently, we adopted the simpler one-way approach to compute the final rankings.

[18] https://choix.lum.li/en/latest/

Introduction

When people work with large document collections, they often want to organize those documents into different categories or themes. For instance, someone analyzing patients' comments about their experiences in hospitals might discover that there are categories like "Long Emergency Room Wait Times" or "Caring Nurses".

In this survey, you will be answering questions about a small group of a few documents. We want to know when a group brings to mind a descriptive category. This will help us in developing better ways to help researchers study large quantities of text.

For this study, all the documents you will look at are summaries of legislation in the United States Congress.

Instructions

First, you will read a group of documents and keywords. For that group, you will **form an idea for a category** that the group seems to be about, then **write a label that describes that category**. Think of a category that fits both the keywords and documents as closely as possible, and that could help someone identify whether a document is in that category or not. Sometimes, it may not be easy to identify a good category or figure out a good label at all—just try your best.

Next, you will answer several questions about additional documents. For each question, you will read a document, and **answer whether the category applies to that document or not.**

Finally, you will **rank documents based on how well they fit the category**.

If you have trouble answering, try to think about how others would respond. **We will award a bonus** if your answers are close to those of other respondents (it can take a few days for us to process results first).

Expectations and Payment Policy

We expect you to put in a reasonably thorough effort. In earlier studies, most people take between 5 and 15 minutes and are approved automatically. **Use of generative AI tools (e.g., ChatGPT) is not allowed**. Please note that **there are attention checks** and some straightforward questions to test your comprehension. If you fail these checks, you will not receive a bonus, and you **may be asked to return your submission without payment after manual review**. Extremely low effort responses risk rejection, although this is exceptionally rare (less than 1% of cases in our experience).

Figure 11: Instructions for the human annotation protocol.

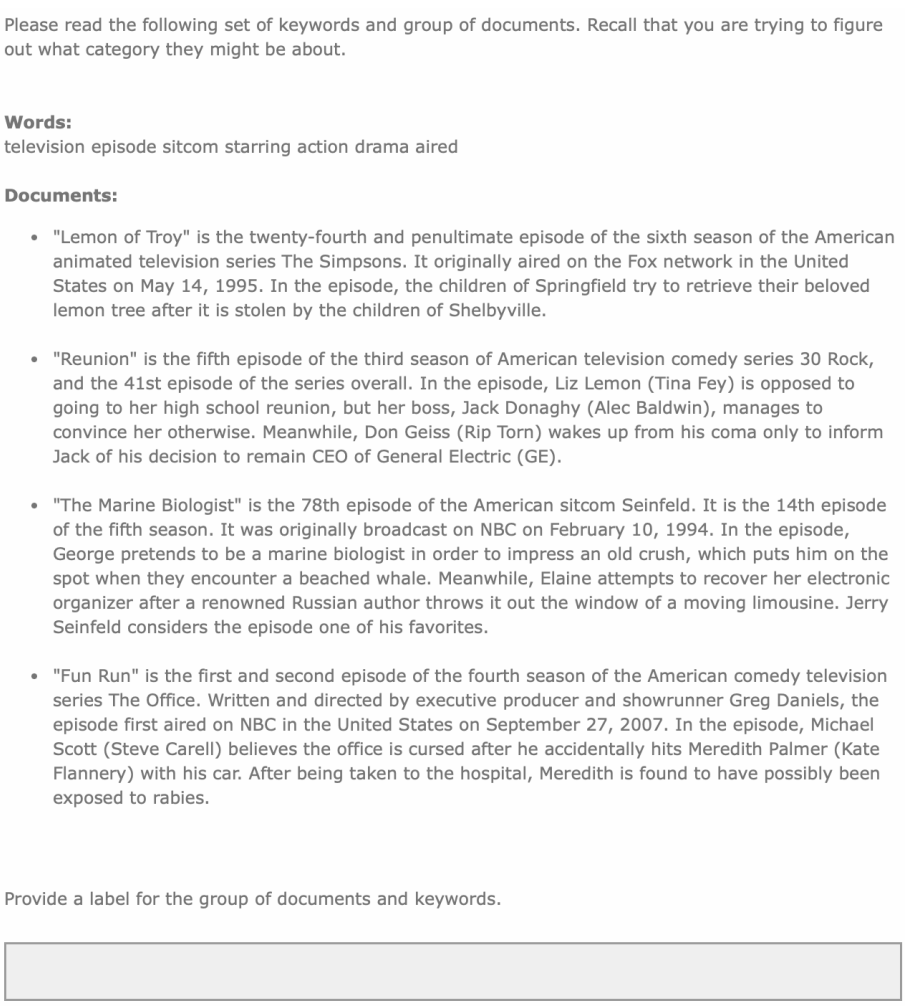


Figure 12: Step 1. Category identification in the human annotation protocol for the practice question.

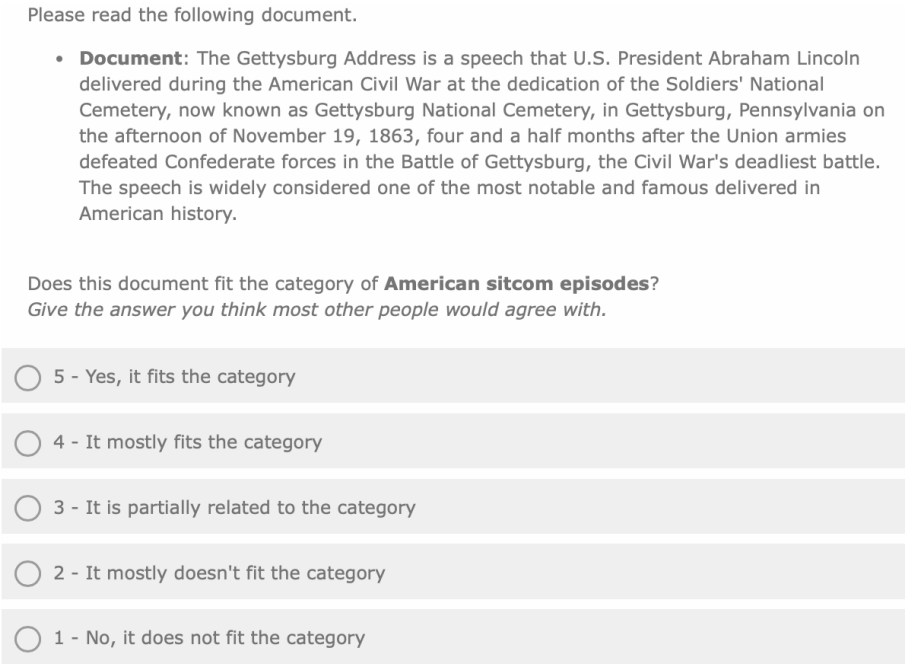


Figure 13: Step 2. Relevance judgment in the human annotation protocol for the practice question.

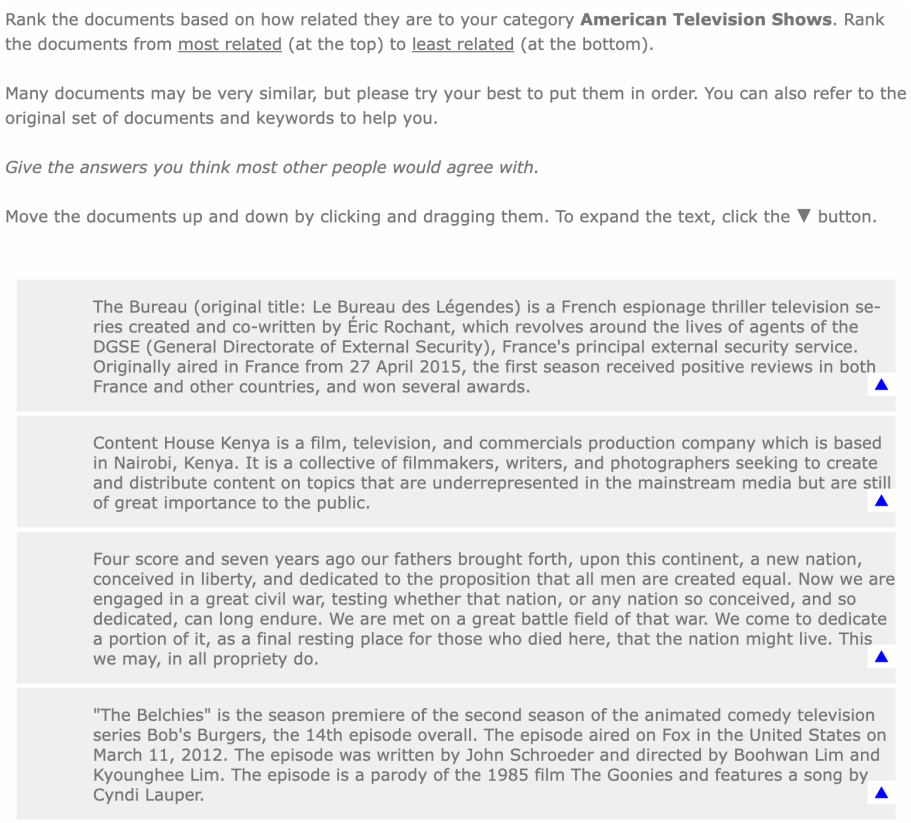


Figure 14: Step 3. Representativeness ranking in the human annotation protocol for the practice question.

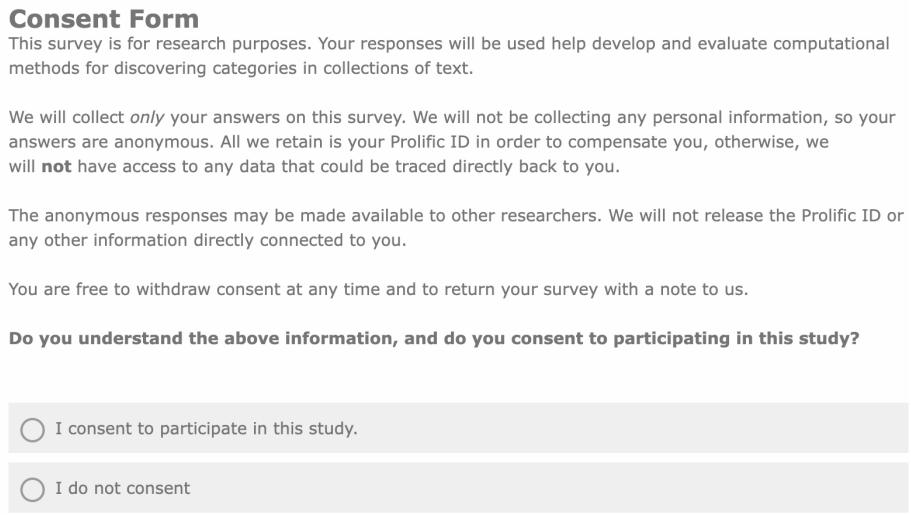


Figure 15: Consent page (shown at beginning)

I.3 Prompt templates

System Prompt I.1: Category Identification (Step 1)

```
You are a helpful AI assistant tasked with creating descriptive labels for a set of keywords and a
    group of documents, each focused on a common topic, as similar as possible to how a human
    would do. The goal is to provide meaningful, concise labels that capture the central theme or
    key concepts represented by the keywords and documents.
```

System Prompt I.2: Relevance Judgment (Step 2)

```
You are a helpful AI assistant tasked with determining if a document fits a given category, aiming
    to make judgments closely aligned with human reasoning.capture the central theme or key
    concepts represented by the keywords and documents.
```

System Prompt I.3: Representativeness Pairwise Ranking (Step 3)

```
You are a helpful AI assistant tasked with determining if a document fits a given category, aiming
    to make judgments closely aligned with human reasoning.capture the central theme or key
    concepts represented by the keywords and documents.
```

Instruction Prompt I.4: Category Identification (Step 1)

```
You will be provided with a set of keywords and a group of documents, each centered around a common
     topic. Your task is to analyze both the keywords and the content of the documents to create a
     clear, concise label that accurately reflects the overall theme they share.

Task Breakdown:
1. Examine the Keywords: Use the keywords as clues to identify the general subject area or themes
    present in the documents.
2. Review the Documents: Skim the summaries provided to understand their main ideas and any
    recurring elements.
3. Generate a Label: Based on the keywords and document content, come up with a single label that
    best describes the topic connecting all the documents.

Examples:
--------
{}

#########

KEYWORDS: {}
DOCUMENTS: {}
Based on the keywords and document content, come up with a single category that best describes the
    topic connecting all the documents. Return just the category.
CATEGORY:
```

Instruction Prompt I.5: Relevance Judgment (Step 2)

```
System message:

Your input fields are:
1. `CATEGORY` (str)
2. `DOCUMENT` (str)

Your output fields are:
1. `reasoning` (str)
2. `FIT` (str): Whether the DOCUMENT fits with the given CATEGORY or not (YES or NO)

All interactions will be structured in the following way, with the appropriate values filled in.

[[ ## CATEGORY ## ]]
{{CATEGORY}}

[[ ## DOCUMENT ## ]]
{{DOCUMENT}}

[[ ## reasoning ## ]]
{{reasoning}}

[[ ## FIT ## ]]
{{FIT}}

[[ ## completed ## ]]

In adhering to this structure, your objective is:
       Determine whether the DOCUMENT fits with the given CATEGORY or not


User message:

[[ ## CATEGORY ## ]]
{{FEW SHOT CATEGORY}}

[[ ## DOCUMENT ## ]]
{{FEW SHOT DOCUMENT}}

Respond with the corresponding output fields, starting with the field `[[ ## reasoning ## ]]`, then
      `[[ ## FIT ## ]]`, and then ending with the marker for `[[ ## completed ## ]]`.


Assistant message:

[[ ## reasoning ## ]]
{{FEW SHOT reasoning}}

[[ ## FIT ## ]]
{{FEW SHOT FIT}}

[[ ## completed ## ]]

User message:

[[ ## CATEGORY ## ]]
{category}

[[ ## DOCUMENT ## ]]
{document}

Respond with the corresponding output fields, starting with the field `[[ ## reasoning ## ]]`, then
      `[[ ## FIT ## ]]`, and then ending with the marker for `[[ ## completed ## ]]`.
```

Instruction Prompt I.6: Representativeness Pairwise Ranking (Step 3)

```
System message:
Your input fields are:
1. `CATEGORY` (str)
2. `DOCUMENT_A` (str)
3. `DOCUMENT_B` (str)

Your output fields are:
1. `reasoning` (str)
2. `CLOSEST` (str): Document that is more closely related to the category (A or B)
All interactions will be structured in the following way, with the appropriate values filled in.

[[ ## CATEGORY ## ]]
{{CATEGORY}}

[[ ## DOCUMENT_A ## ]]
{{DOCUMENT_A}}

[[ ## DOCUMENT_B ## ]]
{{DOCUMENT_B}}

[[ ## reasoning ## ]]
{{reasoning}}

[[ ## CLOSEST ## ]]
{{CLOSEST}}

[[ ## completed ## ]]

In adhering to this structure, your objective is:
       Determine which document is more closely related to the given category

User message:
[[ ## CATEGORY ## ]]
{{FEW SHOT CATEGORY}}

[[ ## DOCUMENT_A ## ]]
{{FEW SHOT DOCUMENT_A}}

[[ ## DOCUMENT_B ## ]]
{{FEW SHOT DOCUMENT_B}}

Respond with the corresponding output fields, starting with the field `[[ ## reasoning ## ]]`, then
      `[[ ## CLOSEST ## ]]`, and then ending with the marker for `[[ ## completed ## ]]`.

Assistant message:
[[ ## reasoning ## ]]
{{FEW SHOT reasoning}}

[[ ## CLOSEST ## ]]
{{FEW SHOT CLOSEST}}

[[ ## completed ## ]]

User message:
[[ ## CATEGORY ## ]]
{category}

[[ ## DOCUMENT_A ## ]]
{doc_a}

[[ ## DOCUMENT_B ## ]]
{doc_b}

Respond with the corresponding output fields, starting with the field `[[ ## reasoning ## ]]`, then
      `[[ ## CLOSEST ## ]]`, and then ending with the marker for `[[ ## completed ## ]]`.
Response:
```