Alexander Hoyle, Lorena Calvo-Bartolomé, Jordan Lee Boyd-Graber, and Philip Resnik. **ProxAnn: Use-Oriented Evaluations of Topic Models and Document Clustering**. *Association for Computational Linguistics*, 2025.


```
    @inproceedings{Hoyle:Calvo-Bartolom\'e:Boyd-Graber:Resnik-2025,
Title = {ProxAnn: Use-Oriented Evaluations of Topic Models and Document Clustering},
Author = {Alexander Hoyle and Lorena Calvo-Bartolom\'e and Jordan Lee Boyd-Graber and Philip Resnik},
Booktitle = {Association for Computational Linguistics},
Location = {Vienna, Austria},
Year = {2025},
Url = {http://cs.umd.edu/~jbg//docs/2025_acl_proxann.pdf},
}
```


Accessible Abstract: Topic models are tools to help people navigate large document collections. However, testing whether a topic model is good or not is notoriously hard, as it's subjective and requires asking real people about whether the outputs make sense. We show that you can ask a language model to recreate the answers of humans, correlating better with ground truth than previous evaluations.

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

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

Alexander Hoyle[*]
ETH Zürich
alexander.hoyle@ai.ethz.ch

Lorena Calvo-Bartolomé[*]
Universidad Carlos III de Madrid
lcalvo@pa.uc3m.es

Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Philip Resnik
University of Maryland
resnik@umd.edu

Abstract

Topic model 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 proxies are statistically indistinguishable from a human annotator and can therefore serve as a reasonable substitute in automated evaluations.[1]

1 Introduction

Suppose a researcher wants to study the impact of donations on politicians' speech (Goel et al., 2023). For two decades, such questions have often been answered with the help of topic models or other text-clustering techniques (Boyd-Graber et al., 2017). 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; Hoyle et al., 2022; 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 follows that topic model (and document clustering) evaluations should help encourage valid categories (Ying

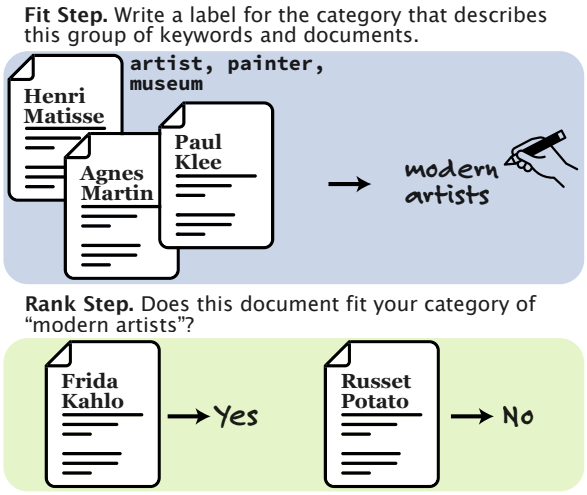


Fig. 1: Our evaluation protocol for topic models and document clustering methods. First, a user reviews documents and keywords related to a topic or cluster and identifies a category. Then, they apply that category to new documents (a third ranking step is not shown). The more human relevance judgments 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.

[*]Equal contribution.

[1]https://github.com/ahoho/proxann contains all human and LLM annotation data, as well as a package (and web interface) to compute metrics on new outputs.

et al., 2022). However, as we discuss in Section 2, the evaluation strategies that are reasonable approximations for this use case are generally dependent on human-derived labels, rendering them hard to scale and reproduce. Conversely, the most common unsupervised automated metrics, while fast to compute, are poor measures of topic quality (Hoyle et al., 2021; 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. 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 analogous of such manual approaches (Bakharia et al., 2016; Baumer et al., 2017; Hoyle et al., 2022). 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

[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: PROXANN-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 standard qualitative analysis (Section 2.1): 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 of prior work in Section 6).

3 PROXANN

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, who rely on experts to create labels); 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 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] Our protocol is also informed by interactive topic modeling for content analysis (Poursabzi-Sangdeh et al., 2016; Li et al., 2024, 2025), where topic model outputs help inform the creation and assignment of categories.

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. L), and model prompts (App. I).

Setup. First, we outline the model outputs required for the evaluation (recall that we are attempting to emulate content analysis, Fig. 1). Through-

[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.

		Krippendorff's α	
		Fit Step	Rank Step
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)
Label-Derived		0.80 (0.13)	0.86 (0.05)

Table 2: Chance-corrected human–human inter-annotator agreement on the two annotation tasks (Krippendorff's α), averaged over eight topics per model (standard deviation in parentheses). *Label-Derived* are six clusters derived from ground-truth `Wiki` labels used in a pilot study, serving as a high-coherence reference. High agreement on the reference indicates the tasks are well-specified. (Topics have ≥ 3 annotators; variation due to filtering out low-quality annotators).

out, we remain as 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 $|\mathcal{D}|$ 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 k^{th} 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 $\mathbf{\Theta} \in \mathbb{R}^{N \times K}$, 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]

[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.

Exemplar documents are a stratified sample over $\boldsymbol{\theta}_k$ (details in App. A).

Label Step: Category identification. After viewing instructions and completing a training exercise (App. L), 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* and the *Rare Diseases Act*.

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

Fit Step: 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, we take the probability-weighted mean over tokens in the scale to obtain relevance judgments, per Wang et al. (2025): $\sum_{s \in \{1 \ldots 5\}} s \cdot p_{LM}(\texttt{s} \mid \texttt{instruction}, \texttt{doc}_i)$.

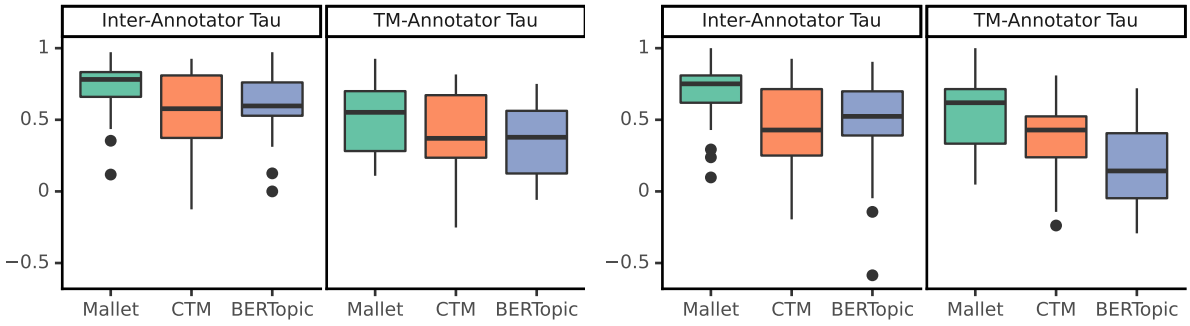
Rank Step: Representativeness ranking. Last, annotators rank the evaluation documents by how representative they are for that category, $\boldsymbol{r}_k^{(i)}$.[10]

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 App. I).

[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.



(a) Relevance Judgements (*Fit Step*)

(b) Representativeness Ranking (*Rank Step*)

Fig. 2: Annotators review the top documents and words from a single topic and infer a category (Label Step), then assign scores to additional documents based on their relationship to the category (Fit and Rank Steps). These scores are correlated with each other (Inter–Annotator Kendall's τ) and with the model's document-topic estimates ($\boldsymbol{\theta}_k$; TM–annotator τ). There are eight topics per model; boxplots report variation in τ over each topic-annotator tuple.

4 Experimental Setup

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 (e.g., Pham et al., 2024; Lam et al., 2024; Li et al., 2024; Zhong et al., 2024): `Wiki` (Merity et al., 2017), a general audience-corpus consisting of 14,000 "good" Wikipedia[11] articles; and `Bills` (Adler and Wilkerson, 2008), a more specialized domain-specific dataset 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 auto-encoders (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 (SBERT, 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 App. F). In a pilot study, we also evaluate a semi-synthetic upper bound model derived from ground-truth `Wiki` labels (App. D).

PROXANN LLMs As LLM annotators, we use OpenAI's GPT-4o (`gpt-4o-mini-2024-07-18`), Llama-3.1-8B, Llama-3.3-70B (Meta LLama Team, 2024), Qwen-2.5-72B, and Qwen 3 in both 8B and 32B variants (Qwen Team et al., 2025a,b). For the label generation step, we set the temperature to 1.0. Then, conditioned on that label, the Fit and Rank steps have temperature 0.[13] We then *re-sample* the chain of steps five times and take the mean responses (just as we average over human annotators). More details can be found in App. I.

4.3 Collecting Human Annotations

A comprehensive human evaluation of all topics would be cost-prohibitive, so we randomly sample eight of the fifty topics for the `Wiki` and `Bills` data on each of the three models. We recruit at least four annotators per topic through Prolific. Low-quality respondents are filtered out with attention checks.[14]

[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] Documents exceeding 100 tokens are truncated, extending to the end of the sentence to avoid incomplete cuts.

[14] `prolific.com`, further recruitment details in App. B. While using more annotators per topic would provide more robust estimates of model performance, Ying et al. (2022) use

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).

Initial Task Validation There are two potential sources of disagreement in the resulting annotations: either the topics can be incoherent, leading to inconsistent category labels between annotators, or the tasks could be poorly defined. As an initial validation step, we run a pilot study with six clusters derived from ground-truth `Wiki` labels to serve as a rough upper bound (comparing them with six CTM and MALLET topics as a reference). Specifically, we create clusters by assigning all documents to their labeled category (e.g., MEDIA AND DRAMA), then rank the documents within a cluster based on their cosine similarity to the centroid (computed with SBERT (Reimers and Gurevych, 2019); additional details and results in App. D).

This approach reduces noise introduced by low-quality topics: because annotators review documents that belong to the same coherent category, their inferred conceptualizations should be fairly crisp. If the tasks are well-specified, annotations on the tasks should be consistent.

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 the Fit and Rank steps (with ordinal weights).[15] For easier comparison with the topic model–human metrics (next section), we also compute annotator-to-annotator correlations between each annotator's relevance fit scores or ranks and the averaged fits (ranks) of all other annotators.

	Document-Level ρ		Topic-Level ρ	
	Fit	Rank	Fit	Rank
	`Wiki`			
GPT-4o	$0.56^{*\dagger}$	$0.68^{*\dagger}$	$0.66^{\dagger}$	$0.55^{\dagger}$
Llama-3.1-8B	0.22	0.36	0.05	0.11
Llama-3.3-70B	$0.57^{*\dagger}$	$0.67^{*\dagger}$	$0.58^{\dagger}$	$0.50^{\dagger}$
Qwen-3-8B	$0.56^{*\dagger}$	$0.58^{\dagger}$	0.46	0.39
Qwen-3-32B	$0.55^{*\dagger}$	$0.63^{\dagger}$	0.47	0.42
Qwen-2.5-72B	$0.52^{\dagger}$	$0.68^{*\dagger}$	$0.66^{\dagger}$	0.46
	`Bills`			
GPT-4o	$0.65^{*\dagger}$	$0.71^{*\dagger}$	$0.77^{*\dagger}$	$0.75^{*\dagger}$
Llama-3.1-8B	0.30	$0.53^{\dagger}$	0.14	0.44
Llama-3.3-70B	$0.66^{*\dagger}$	$0.67^{*\dagger}$	$0.70^{*\dagger}$	$0.60^{\dagger}$
Qwen-3-8B	$0.66^{*\dagger}$	$0.57^{\dagger}$	$0.80^{*\dagger}$	0.43
Qwen-3-32B	$0.67^{*\dagger}$	$0.68^{*\dagger}$	$0.74^{*\dagger}$	$0.70^{*\dagger}$
Qwen-2.5-72B	$0.61^{*\dagger}$	$0.71^{*\dagger}$	$0.78^{*\dagger}$	$0.65^{\dagger}$

Table 3: Advantage probabilities from the alternative annotator test; the probability that PROXANN is "as good as or better than a randomly chosen human annotator" (Calderon et al., 2025). Document-level scores consider annotations by document; Topic-level over all documents evaluated in the topic. $*$ indicates that win rates over humans are above 0.5, as determined by a one-sided t-test (over 10 resamples of combined annotators). † is the equivalent for Wilcoxon signed-rank.

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 the Fit and Rank steps, 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; see App. A for $\boldsymbol{\theta}^{\text{eval}}$ sampling details).

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 ($\boldsymbol{s}_k^{(i)}$, Fit Step) or their ranks ($\boldsymbol{r}_k^{(i)}$, Rank Step), for annotator i. We contextualize these against the inter-human-annotator τ (see above).

PROXANN–human agreement. For the LLM to serve as a proxy, it should ideally be indistinguishable from a human annotator. This idea is operationalized by the Alternative Annotator Test (alt-test, Calderon et al., 2025). For each annotated instance d_i, the alt-test computes two leave-one-out

two per topic in a similar setup; the statistical test of annotator–LLM substitutability (§4.4) requires only three. Agreement is also high for a synthetic upper-bound (Table 2).

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

similarity metrics: the similarity between between annotator j's responses and the responses of all *other* annotators, $s_{h:h}(d_i)$, and the similarity between the LLM's response and those of all other annotators, $s_{lm:h}(d_i)$. The result is a set of binary outcomes $\mathbb{I}\left[s_{h:h}(d_i) < s_{lm:h}(d_i)\right]$, and a one-sided t-test (Wilcoxon signed-rank for small n) determines whether the LLM wins significantly more often (subject to a slack term ϵ we set to 0.1, per their suggestion for crowdworkers). It computes ω, the (multiple-comparison-adjusted) win rate of LLMs over annotators (over all annotators j), and the *advantage probability* that an LLM is as (or more) reliable than a human, ρ.

We apply the alt-test to annotations on individual documents as well as on the entire topic, using the root mean squared error as the similarity metric. Our annotators independently review only seven documents from a single (model, topic, dataset) tuple, which is insufficient for the test. Following Calderon et al.'s recommendation, we combine annotations to create pseudo-annotators. We combine over all topics within a dataset, such that the "annotator" observes $n_{\text{models}} \cdot n_{\text{topics}} \cdot n_{\text{docs}} = 3 \times 8 \times 7$ items (we bootstrap over ten random permutations, computing ω over the full set; variance of ρ is small and not reported. Results from combining over topics per model in App. G.)

ProxAnn as an automated evaluator. A common use for automated coherence metrics, like NPMI (Lau et al., 2014), 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).

We compare the human evaluations of topics and clusters to metrics based on ProxAnn. Define evaluation metrics per the above descriptions:

$$\text{FIT}\text{–}\tau_{\,h:tm}(k) := \tau\left(\boldsymbol{s}_k^{(h)}, \boldsymbol{\theta}_k^{\text{eval}}\right) \quad (1)$$

$$\text{FIT}\text{–}\tau_{\,lm:tm}(k) := \tau\left(\boldsymbol{s}_k^{(lm)}, \boldsymbol{\theta}_k^{\text{eval}}\right). \quad (2)$$

These are the correlations for topic k between the Fit step responses (from either humans or ProxAnn, $\boldsymbol{s}_k$) and (b) the estimated document scores from the topic model. Define $\text{RANK}\text{–}\tau(k)$ analogously for the Rank-step responses. For each topic and task, there is a "ground-truth" evaluation metric and a "proxy" metric — $\text{FIT}\text{–}\tau_{\,h:tm}(k)$ or $\text{RANK}\text{–}\tau_{\,h:tm}(k)$ —and a "proxy" metric, $\text{FIT}\text{–}\tau_{\,lm:tm}(k)$ or $\text{RANK}\text{–}\tau_{\,lm:tm}(k)$. A second Kendall's τ over these metrics for all k measures the extent to which ProxAnn's rankings over topics agrees with that of the average human.

	$\tau\,(\text{FIT}\text{–}\tau_{\,h:tm}, \cdot)$		$\tau\,(\text{RANK}\text{–}\tau_{\,h:tm}, \cdot)$	
	Wiki	Bills	Wiki	Bills
NPMI	-0.15 (0.14)	0.01 (0.10)	-0.18 (0.10)	-0.02 (0.12)
FIT/RANK–$\tau_{\,lm:tm}$				
Llama-3.1-8B	0.19 (0.18)	0.16 (0.18)	-0.35 (0.14)	0.15 (0.14)
Qwen-3-8B	0.35 (0.16)	0.12 (0.16)	0.33 (0.16)	0.28 (0.13)
Qwen-3-32B	0.20 (0.18)	**0.34 (0.11)**	**0.51 (0.11)**	**0.30 (0.13)**
Llama-3.3-70B	0.41 (0.14)	0.26 (0.15)	0.36 (0.13)	0.19 (0.13)
Qwen-2.5-72B	**0.48 (0.13)**	0.22 (0.17)	0.36 (0.12)	0.21 (0.15)
GPT-4o	0.22 (0.13)	0.31 (0.13)	0.27 (0.14)	0.29 (0.11)
FIT/RANK–$\tau_{\,h:tm}$				
Human	0.41 (0.09)	0.09 (0.14)	0.34 (0.09)	0.18 (0.12)

Table 4: Relationship between automated and human topic rankings. Cells show Kendall's τ between metrics: FIT/RANK–$\tau_{\,h:tm}$ correlates human scores to document–topic probabilities ($\boldsymbol{\theta}_k$); FIT/RANK–$\tau_{\,lm:tm}$ correlates ProxAnn to $\boldsymbol{\theta}_k$. Values are bootstrapped means and standard deviations (resampling over topics). The *Human* row reflects leave-one-out inter-annotator correlations, serving as a reference. While larger Qwen models achieve the strongest correlations, GPT-4o is middling. NPMI is not correlated with human metrics.

5 Results

We discuss results in the same order they were presented above. In tables and figures, **Fit** refers to responses to the relevance judgments of evaluation documents and **Rank** to responses to 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: the agreement metrics for the synthetic *label-derived* clusters 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 be-

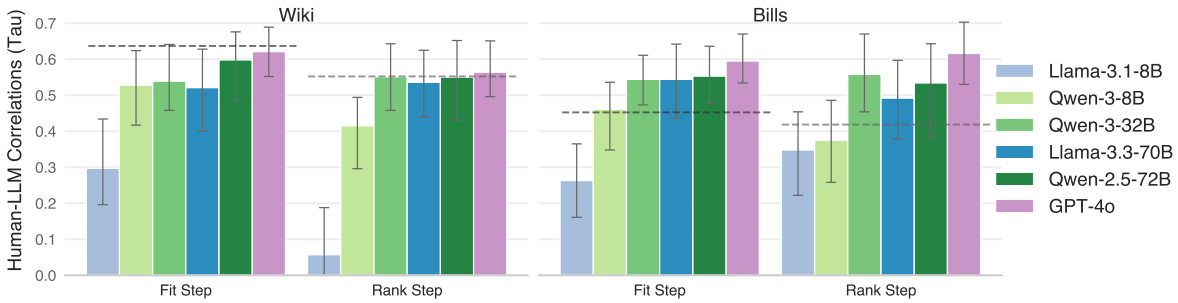


Fig. 3: Correlations between PROXANN and human annotations (Kendall's τ) for the relevance judgment (Fit Step) and representativeness ranking (Rank Step) tasks, averaged over topics (pooled over all three models). While GPT-4o has the best overall correlations and relatively low variance, the Qwen family is a reasonable substitute, even at smaller sizes. Dashed horizontal lines are the average leave-one-out human–human correlation (per Figs. 2 and 6). Error bars represent 95% bootstrapped CIs, resampling over topics.

tween the estimated document-topic probabilities ($\boldsymbol{\theta}_k$) and human judgments on the `Wiki` data (Fig. 2, `Bills` data in Fig. 6 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 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 App. C, we report on two additional metrics, NDCG and binarized agreement.

These results support the idea that MALLET, at 20 years old, remains an effective tool for automated content analysis (echoing Hoyle et al. 2022).

5.3 Is PROXANN a good proxy?

Generally, PROXANN is a reasonable proxy across both steps and datasets, although there is variation among the underlying LLMs. The correlations between LLM and human responses are generally around $\tau = 0.5$ or greater for the largest ($\geq$ 32B) models (Fig. 3); Qwen-3-8B is competitive in a few cases, but Llama-3.1-8B is generally poor. In many cases, the best models meet or exceed the average leave-one-out human-to-human correlations.

These results are largely corroborated by the alt-test (Table 3). When considering annotations at the document-level, PROXANN (for larger models) is a suitable substitute for human annotators: advantage probabilities ρ are generally over 0.5 for the stronger models, and have significantly higher agreement rates with humans than humans do with each other. The picture is a little less rosy for the topic-level annotations, where agreement is computed as the τ. Here, the tests are lower-powered (as the number of instances has been reduced by a factor of $n_{\text{docs}} = 7$), making statistical wins less probable. Second, the high human agreements for the relevance judgment (fit) tasks on the `Wiki` data make it harder for an LLM to perform as well. In addition, it may be that the LLM provides overly-specific topics on Wikipedia due to greater domain "knowledge" (further discussion in Section 5.5).

5.4 Ranking Topics and Models

We now measure whether metrics derived from PROXANN rank topics and models similarly to humans. Generally, no model is dominant, with almost all correlations less than $\tau = 0.5$. Qwen-2.5-72B performs reasonably well on the `Wiki` data, but performance on `Bills` is generally low. There, low correlations 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.

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` τ for the rank task is 0.33). We also report results with *binarized* $\boldsymbol{\theta}$, corresponding to hard assignments, which tend to show better correlations (App. K). Meanwhile, the standard automated metric, NPMI, fails to capture the human judgments.

Together, these results indicate that there is some capacity for PROXANN to accurately rank topics at least as well as an arbitrary annotator. Last, when aggregating at the *model*-level (i.e., over CTM, BERTOPIC, MALLET) model rankings align for `Wiki` and are generally close for `Bills` (App. K).

5.5 A Qualitative View of Agreement

Last, we examine how annotator agreement—and agreement between annotators and the topic model or LLM—reflects topic quality (Appendix E).

Topics with low human-to-human agreement tend to be too broad or multi-themed, leading to disagreement in both categorical labels and document fits across human annotators (Table 7 in appendix). In contrast, low human-to-topic model agreement (conditioned on high human–human agreement) often reveals model-specific limitations: BERTopic may underassign relevant documents due to its hard clustering approximation, while MALLET may fully assign a document to a single topic ($\theta_d = 1$) simply because no better alternative is available. This pattern highlights that under high human–human agreement, low human–topic model agreement is more likely to indicate model failure than annotation ambiguity.

We also analyze two case studies of high-human agreement topics where LLM judgments nonetheless diverge (Tables 8 and 9 in appendix). While the human-to-LLM agreement yields moderate Kendall's τ values (0.48 and 0.58), qualitative inspection shows that actual differences in fit judgments are often small. Table 8 shows near-perfect alignment between human and LLM fit scores, despite the lower τ. In Table 9, where τ is slightly higher, fit scores also agree closely, though minor differences emerge from how the topic is interpreted. Together, these examples raise the question of what constitutes a "low" or "bad" τ in this context; values below 0.5 may still reflect reasonable alignment. Interpreting τ in isolation may be misleading, and thresholds for "good" agreement should be grounded in qualitative examples.

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 closer to our protocol, Ying et al. 2022 propose crowdworker "label validation" tasks, designed to assess the quality of individual document-topic distributions using already-identified expert labels. Furthermore, the tasks require a curated set of labels covering *all* relevant topics, whereas our setup can assess topics independently. Although the above evaluations are better aligned with real-world use than topic coherence, they rely on some form of manual labeling, and are difficult to scale (Ying et al. 2022 only evaluate on one dataset).

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.[16] 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.

There are several promising directions in these areas: the development of specialized models for automated topic and cluster annotation, rather than generalized LLMs; extending our approach to non-English languages; incorporating annotations from experts who have specific information-seeking needs. To support adoption and further experimentation, we provide both a demo interface[17] and a local deployment option for computing ProxAnn metrics on new model outputs.

[16] 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.

[17] See link at https://github.com/ahoho/proxann

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).[18] Exploring the role of expertise in topic model evaluation is an important direction for future research—in particular, studying the relationship between expert annotations and those from language models (as well as how they differ from crowdsourced annotations).

A final limitation is our exclusive use of English-language datasets. While we do not have access to the exact pretraining mixture for the LLMs, it is reasonable to assume that English data is a dominant component, in addition to being heavily favored in evaluation tasks. We therefore do not expect our findings to generalize directly to non-English settings.

Acknowledgments

Many thanks are owed to Rupak Sarkar for discussion and insight during initial planning stages of this work. Kris Miler, Aaron Schein, and Michelle Mazurek provided comments on the human evaluation protocol, which appeared in Alexander's dissertation. We also gratefully acknowledge Saúl Blanco Fortes for his valuable help with system configuration and the setup of the ProxAnn test frontend. The work of Lorena Calvo-Bartolomé has been partially supported by Grant PID2023-146684NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF/UE. This work was also supported in part by the U.S. National Science Foundation awards 2124270 (Resnik) and 2229885 (Boyd-Graber). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

[18] A reviewer also suggested combining LLM and human annotations, per He et al. (2024).

References

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

Ron Artstein. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.

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 Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Found. Trends Inf. Retr.*, 11(2–3):143–296.

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.

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

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *Preprint*, arXiv:2501.10970.

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)*, Vancouver, British Columbia, Canada. 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, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo 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.

Pranav Goel, Nikolay Malkin, SoRelle W. Gaynor, Nebojsa Jojic, Kristina Miler, and Philip Resnik. 2023. Donor activity is associated with us legislators' attention to political issues. *PLOS ONE*, 18(9):1–24.

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. *Preprint*, arXiv:2203.05794.

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

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, Online.

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.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

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, Lorena Calvo-Bartolomé, Alexander Hoyle, Paiheng Xu, Alden Dima, Juan Francisco Fung, and Jordan Boyd-Graber. 2025. Large language models struggle to describe the haystack without human help: Human-in-the-loop evaluation of llms. *Preprint*, arXiv:2502.14748.

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. *Preprint*, arXiv:2306.03100.

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

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for llm compression and acceleration. *Preprint*, arXiv:2306.00978.

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)*.

Meta LLama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Qwen Team, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025a. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team, : An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

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.

Victor Wang, Michael J. Q. Zhang, and Eunsol Choi. 2025. Improving llm-as-a-judge inference with the judgment distribution. *Preprint*, arXiv:2503.03064.

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. *Preprint*, arXiv:2406.09008.

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.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. 2024. Explaining datasets in words: Statistical models with natural language parameters. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 79350–79380, Vancouver, British Columbia, Canada. Curran Associates, Inc.

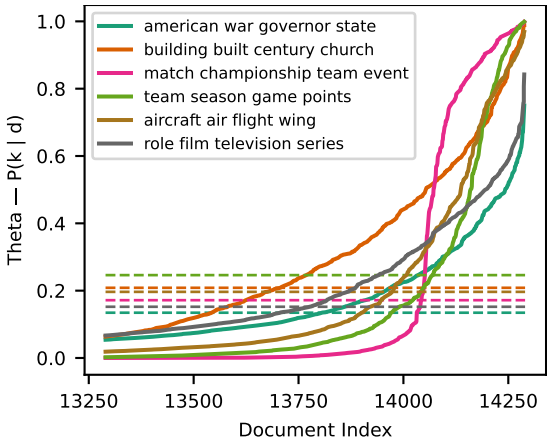
A Exemplar Document Selection

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 4 (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).

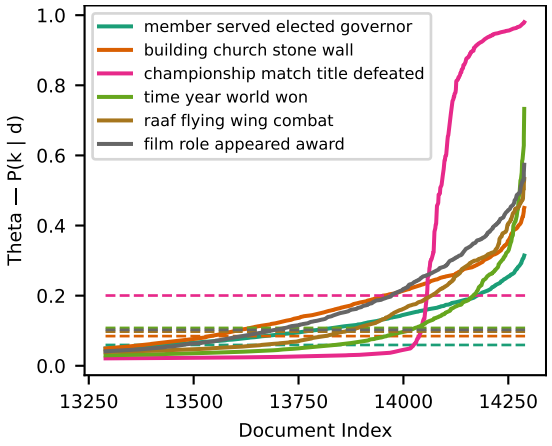
In Fig. 4, we visualize these distributions for CTM and MALLET for the pilot topics alongside the detected threshold. Documents above this threshold are sampled (proportional to θ_{dk} to produce the exemplar documents .

B 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



(a) MALLET



(b) CTM

Fig. 4: 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.

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 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.

	Fit Step α	Rank Step α
MALLET	0.59 (0.16)	0.71 (0.09)
CTM	0.64 (0.15)	0.67 (0.13)
Label-Derived	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 Additional Results from the Human Study

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. 6—of note is that BERTOPIC cluster assignments tend to have higher agreement with human relevance judgments (binarized responses to the Fit Step), likely due to it being a clustering model.

We also report the distributions of inter-annotator correlations *per topic* in Fig. 7, showing that certain individual topics can have relatively high variance in human annotations.

D 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"

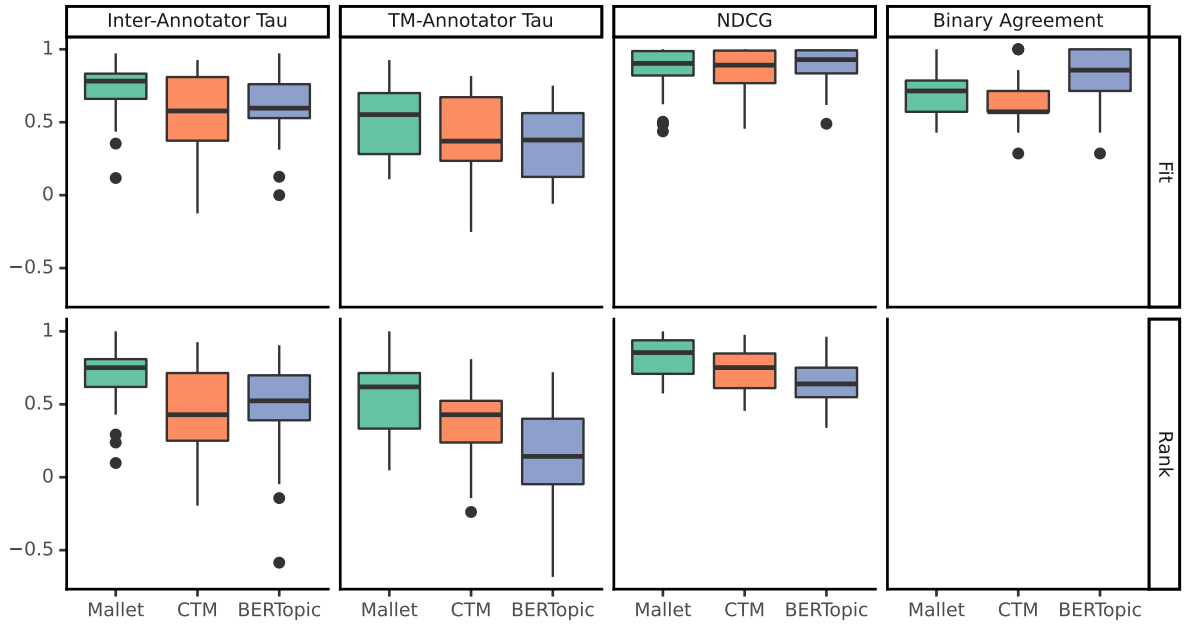


Fig. 5: Metrics quantifying the relationship between human annotations and estimated document-topic probabilities ($\boldsymbol{\theta}_k$) for the three topic models on all eight `Wiki` topics. From left-to-right, the metrics are inter-annotator Kendall's τ, model-annotator τ, relevance agreement, and NDCG. The top row of figures reports relationships with human relevance judgments (on a 1-5 scale, Fit Step in the protocol), and the bottom row relationships with their document rankings (Rank Step). Boxplots report variation over topic-annotator pairs (binary agreement does not apply to the rank task).

(termed *Label-Derived*) 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. 8, 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 App. I.

E 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.[19] While it is possible to average these values over topics, their occasionally counter-intuitive nature makes them less desirable for model comparison. For a more in-depth overview of inter-annotator agreement in linguistic annotation, we refer the reader

[19] 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.

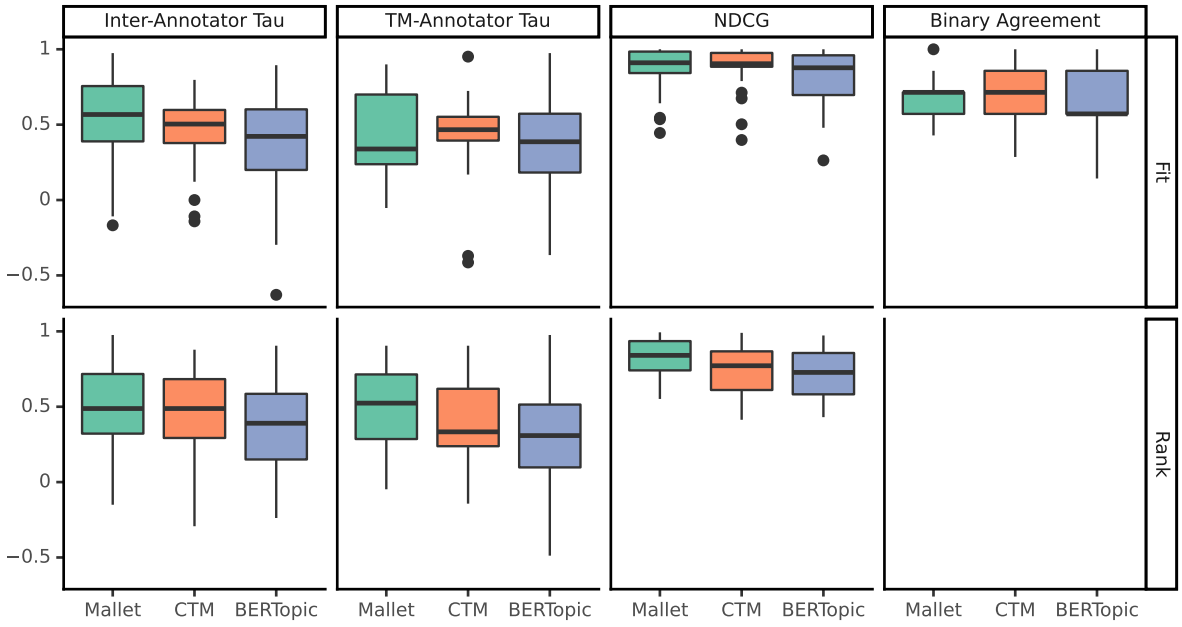


Fig. 6: Metrics quantifying the relationship between human relevance judgments and estimated document-topic probabilities ($\boldsymbol{\theta}_k$) for three models on all eight `Bills` topics. See Fig. 5 for additional details.

to Artstein (2017).

F BERTopic training details

Although the BERTopic author advises against data preprocessing[20], 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 use the probabilities derived from BERTopic's `approximate_distribution` function as document-topic distribution to obtain the evaluation documents.

G Alternative Annotator Combination Strategy for the Alt-Test

Given that each crowdworker only annotates one topic (hence seven documents), the standard application of the alt-test would be insufficiently powered. As discussed in Section 4.4, must therefore combine annotators to create "pseudo-annotators" who appear to have annotated multiple topics, per Calderon et al. (2025).

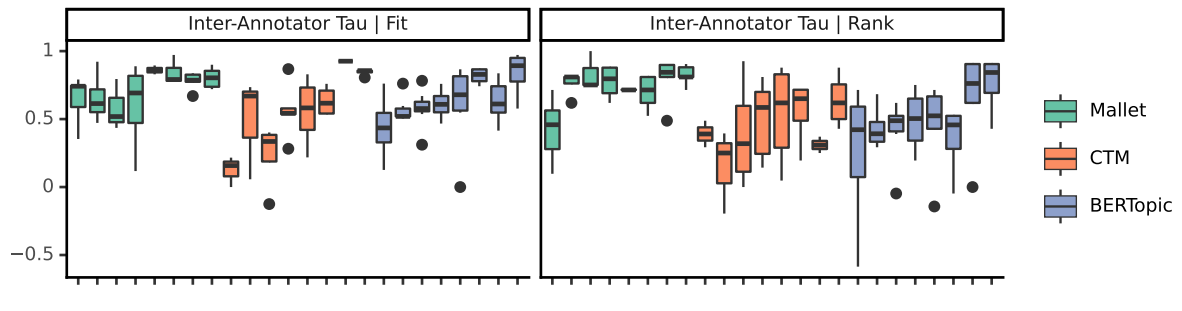
However, there are multiple ways of combining annotators. For the results in Table 3, we combine annotators over all 24 topics per dataset; given that each topic has at least three annotators (after filtering), this produces three pseudo-annotators who each observe a set of 168 unique documents Recall that the statistical tests are run over the computed human-human (or human-LLM) similarities across the annotated instances (i.e. documents), so higher numbers imply more power.

As an alternative that generates more pseudo-annotators (but fewer documents per annotator), we randomly combine topics per *model*, rather than per dataset. This introduces more noise, as the distribution of topics viewed by each "annotator" is variable (e.g., some could observe all low-quality topics with high disagreement), and makes the statistical tests harder to "pass." Results are in **??**: the ρ are roughly the same as before, but statistical win rates above 0.5 (as indicated by the $*$ and †) are less frequent, presumably due to the higher variance and lower power.

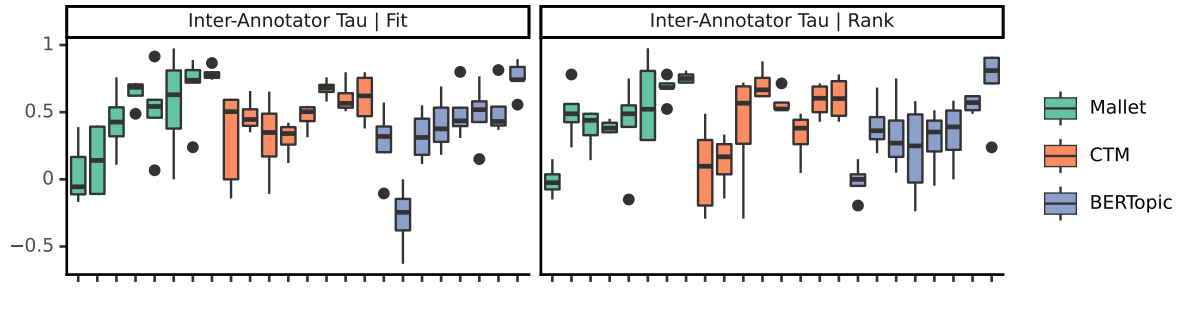
H Additional Bills Results

Figure 6 depict evaluations on the `Bills` data, corresponding to Fig. 5 in the main text.

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



(a) Wiki



(b) Bills

Fig. 7: Distributions of leave-one-out inter-annotator correlations (Kendall's τ) over all topics. Boxplots report variation over the annotators.

	Document-Level ρ		Topic-Level ρ	
	Fit	Rank	Fit	Rank
	`wiki`			
GPT-4o	$0.58^{\dagger}$	$0.68^{\dagger}$	0.67	0.55
Llama-3.1-8B	0.22	0.37	0.04	0.12
Llama-3.3-70B	$0.58^{\dagger}$	$0.67^{\dagger}$	0.58	0.48
Qwen-3-8B	$0.58^{\dagger}$	0.58	0.46	0.38
Qwen-3-32B	$0.57^{\dagger}$	0.63	0.51	0.42
Qwen-2.5-72B	$0.53^{\dagger}$	$0.68^{\dagger}$	0.67	0.46
	`bills`			
GPT-4o	$0.65^{*\dagger}$	$0.71^{*\dagger}$	0.77	0.76
Llama-3.1-8B	0.30	$0.53^{\dagger}$	0.14	0.44
Llama-3.3-70B	$0.66^{*\dagger}$	$0.67^{\dagger}$	0.69	0.60
Qwen-3-8B	$0.66^{*\dagger}$	$0.57^{\dagger}$	0.80^{*}	0.43
Qwen-3-32B	$0.67^{*\dagger}$	$0.68^{\dagger}$	0.75	0.71
Qwen-2.5-72B	$0.61^{*\dagger}$	$0.70^{\dagger}$	0.79	0.65

Table 6: Advantage probabilities ρ from the alternative annotator test using an alternative combination method (details in App. G). ρ is the probability that ProxAnn is "as good as or better than a randomly chosen human annotator" (Calderon et al., 2025). Document-level scores consider annotations by document; Topic-level over all documents evaluated in the topic. $*$ indicates that win rates over humans are above 0.5, as determined by a one-sided t-test (over 10 resamples of combined annotators). † is the equivalent for Wilcoxon signed-rank.

I Prompting details

I.1 Evaluation protocol configuration

Here, we outline our prompt engineering process used to configure the LLM-based proxy for the evaluation protocol.

Label Step We use a concise system prompt (M.1) to summarize the tasks and instruct the LLM to simulate human-like behavior. This is paired with an instruction prompt (M.2) that provides task-specific details, augmented with few-shot exemplars.

Fit and Rank Steps Following the findings of Wang et al. (2025), we prompt the LLM using a single instruction prompt per task, without Chain-of-Thought reasoning or few-shot exemplars (M.3 and M.4). For the **Fit Step**, we adopt a pointwise scoring approach to compute the Fit Score. Rather than relying on the LLM's most-probable token alone, we extract the log-probabilities of the top-20 tokens and interpret them as soft judgments over the Likert scale 1–5. We then compute a weighted average across the Likert candidates, using the LLM-assigned probabilities as weights.

The **Rank Step** involves pairwise ranking, where the LLM is presented with pairs of texts and asked to choose which one better fits a given cat-

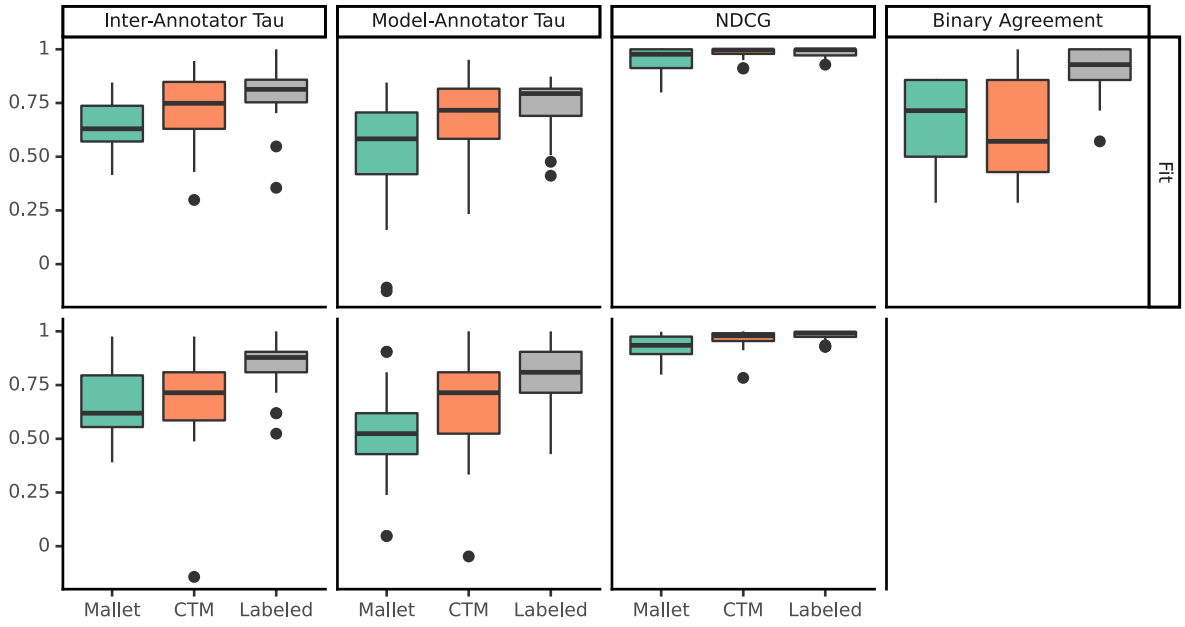


Fig. 8: Metrics quantifying the relationship between human relevance judgments and estimated document-topic probabilities ($\boldsymbol{\theta}_k$) for two models and a synthetic upper-bound ("Labeled", or *Label-Derived* in the text), using six topics from the pilot data. From left-to-right, the metrics 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.

egory. 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, 2025). As in Wang et al. (2025), we take the probability-weighted average over the tokens in both directions before taking the final rank.

I.2 Bradley-Terry

After applying the *Rank Step* prompt to each topic on all $\binom{7}{2}$ combinations of evaluation document 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`[21]library, applying the `ilsr_pairwise` method, setting the regularization term α to 0.001 to ensure numerical stability.

I.3 Deployment Infrastructure

LLama-8B models was run on an NVIDIA 4090 (24 GB RAM); all other models were run on an NVIDIA A100 (80GB RAM). 70B Models were quantized with AWQ (Lin et al., 2024). vLLM was used for inference (Kwon et al., 2023);[22] prompting across all 24 topics takes under an hour.

J Agreement evaluation

Tables 7, 8 and 9 contain qualitative examples to illustrate the analysis carried out in Section 5.5.

K Additional Topic Ranking Results

Fig. 9 reports topic model rankings using our automated metrics, FIT–$\tau_{lm:tm}$, RANK–$\tau_{lm:tm}$.

We also report how well topic rankings correlate when using binary assignments from the topic model—that is, whether the topic is the most-probable for that document, rather than the original real-valued $\boldsymbol{\theta}_k$. Generally, the correlations are higher (Appendix K), suggesting the use of assignments may be preferable for topic rankings.

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

[22] https://docs.vllm.ai/

Topic Words	Categories	Exemplar Document 1	Exemplar document 2	Evaluation Document 1	Evaluation Document 2
Lowest H:H ($\alpha < 0.7$)					
		BERTopic on `Bills` ($\alpha = -0.16$)			
student, school, students, education, schools, leas, higher, ihes	• Aiding children in school to receive a proper, well informed, education in school and post secondary. • Educational reform and students' welfare • K-12 federal education legislation • High School Student Initiative	***Doc ID** 1947:* $\theta_d = 1.00$ Text: English Language Instruction Improvement Act of 2007 - Amends title III (Language Instruction for Limited English Proficient and Immigrant Students) [...]	***Doc ID** 1375:* $\theta_d = 1.00$ Text: Scholarships for Opportunity and Results Reauthorization Act or the SOAR Reauthorization Act This bill amends the District of Columbia Code to [...]	***Doc ID** 197: Human Fit* $= 2.50 \pm 1.66$, $\theta_d = 1.00$ Text: Community College Partnership Act of 2007 - Amends the Higher Education Act of 1965 to establish a Community College Opportunity program to help [...]	***Doc ID** 1387: Human Fit* $= 2.50 \pm 1.66$, $\theta_d = 0.33$ Text: Dynamic Repayment Act of 2016 This bill amends the Higher Education Act of 1965 to replace several existing federal student loan programs with a [...]
		MALLET *on* `Bills`($\alpha = -0.15$)			
act, funds, year, federal, fiscal, fund, amounts, state	• Guidelines for approving Congressional Budget Expenditures • Specific government spending limitation `Bills` • Legislative Acts • Federal American Fiscal Funding Acts	***Doc ID** 2946:* $\theta_d = 0.99$ Text: Realize America's Maritime Promise Act or the RAMP Act - Requires the total budget resources for expenditures from the Harbor Maintenance Trust Fund [...]	***Doc ID** 9315:* $\theta_d = 0.82$ Text: Midshipmen Education Certainty Act Makes appropriations available each fiscal year for operations of the U.S.	***Doc ID** 11330: Human Fit* $= 3.50 \pm 1.12$, $\theta_d = 0.00$ Text: (This measure has not been amended since the Conference Report was filed in the House on June 29, 2010.	***Doc ID** 14524: Human Fit* $= 3.50 \pm 1.12$, $\theta_d = 0.17$ Text: Target Practice and Marksmanship Training Support Act This bill amends the Pittman-Robertson Wildlife Restoration Act to facilitate the [...]
Low H:TM (High H:H $\alpha > 0.7$)					
		BERTopic on `Wiki`($\alpha = 0.72$)			
breed, horses, horse, breeds, arabian, dogs, dog, bred	• Animal Breeds • animal breeds, horse and dog breeds • Animal breeds • Domestic animals • Breeds horses dogs historical working • Native four legged animals	***Doc ID** 5378:* $\theta_d = 1.00$ Text: Carolina Marsh Tacky = The Carolina Marsh Tacky or Marsh Tacky is a rare breed of horse , native to South Carolina .	***Doc ID** 5362:* $\theta_d = 1.00$ Text: Clumber Spaniel = The Clumber Spaniel is a breed of dog of the spaniel type , developed in the United Kingdom .	***Doc ID** 9874: Human Fit* $= 4.57 \pm 0.73$, $\theta_d = 0.17$ Text: Paynter (horse) = Paynter (foaled March 4 , 2009) is an American-bred Thoroughbred racehorse notable for a promising three-year-old racing [...]	***Doc ID** 5298: Human Fit* $= 4.57 \pm 0.73$, $\theta_d = 0.34$ Text: Field Spaniel = The Field Spaniel is a medium-sized breed dog of the spaniel type .
		`Wiki` *on* MALLET ($\alpha = 0.81$)			
species, shark, long, sharks, females, fish, found, birds	• Sharks in the pacific ocean • Description of marine predators • Shark species descriptions and habitats • shark species • Sea wildlife species	***Doc ID** 170:* $\theta_d = 0.99$ Text: Banded houndshark = The banded houndshark (Triakis scyllium) is a species of houndshark , in the family Triakidae , common in the northwestern [...]	***Doc ID** 190:* $\theta_d = 0.98$ Text: Coral catshark = The coral catshark (Atelomycterus marmoratus) is a species of catshark , and part of the family Scyliorhinidae .	***Doc ID** 224: Human Fit* $= 3.00 \pm 0.63$, $\theta_d = 0.33$ Text: Fish = A fish is any member of a paraphyletic group of organisms that consist of all gill-bearing aquatic craniate animals that lack limbs with [...]	***Doc ID** 204: Human Fit* $= 3.00 \pm 0.63$, $\theta_d = 1.00$ Text: Pyjama shark = The pyjama shark or striped catshark (Poroderma africanum) is a species of catshark , and part of the family Scyliorhinidae , [...]
High H:TM (High H:H $\alpha > 0.7$)					
		CTM on `Wiki` ($\alpha = 0.95$, $\tau = 0.85$)			
daily, trunkline, roadway, national_highway_system, travels, designated, surveys, entire	• highway routes traffic traveling • State Highways • State Highway Rules • American highway routes	***Doc ID** 716:* $\theta_d = 0.37$ Text: Ohio State Route 85 = State Route 85 (SR 85 , OH 85) is an east – west state highway in the northeastern Ohio .	***Doc ID** 5768:* $\theta_d = 0.37$ Text: Delaware Route 42 = Delaware Route 42 (DE 42) is a state highway in Kent County , Delaware .	***Doc ID** 598: Human Fit* $= 5.00 \pm 0.00$, $\theta_d = 0.35$ Text: K-22 (Kansas highway) = K-22 is a 3.087-mile-long (4.968 km) highway in the U.S.	***Doc ID** 5696: Human Fit* $= 5.00 \pm 0.00$, $\theta_d = 0.42$ Text: Delaware Route 44 = Delaware Route 44 (DE 44) is a state highway in Kent County , Delaware .
		CTM on `Wiki` ($\alpha = 0.98$, $\tau = 0.82$)			
career, hit, games, season, league, baseball, major_league_baseball, signed	• Former MLB players • American baseball league • Professional baseball facts and figures	***Doc ID** 2943:* $\theta_d = 0.61$ Text: Brian Wilson (baseball) = Brian Patrick Wilson (born March 16 , 1982) is a former American professional baseball relief pitcher .	***Doc ID** 2928:* $\theta_d = 0.59$ Text: Byron McLaughlin = Byron Scott McLaughlin (born September 29 , 1955) is an American retired professional baseball player , alleged counterfeit [...]	***Doc ID** 3001: Human Fit* $= 5.00 \pm 0.00$, $\theta_d = 0.61$ Text: Johnny Evers = John Joseph Evers (July 21 , 1881 – March 28 , 1947) was an American professional baseball second baseman and manager .	***Doc ID** 2870: Human Fit* $= 5.00 \pm 0.00$, $\theta_d = 0.74$ Text: Jon Lieber = Jonathan Ray Lieber (born April 2 , 1970) is a former Major League Baseball (MLB) pitcher .

Table 7: Examples of topics with the lowest human-to-human agreement (H:H), as well as topics with low and high human-to-topic model agreement (H:TM) conditioned on high H:H (defined as Krippendorff's $\alpha > 0.7$). Human-to-topic model agreement is measured using Kendall's τ on the fit scores. For each topic, we show: (1) the top eight words from the topic model, (2) the annotators' categories associated with the topic, (3) the top two exemplar documents, and (4) two evaluation documents selected from the topics with the lowest or highest H:H or H:TM agreement, depending on the condition. Both exemplar and evaluation documents display the model's θ_d, and the evaluation documents additionally include the human-assigned fit rating (mean and standard deviation). Topics with low H:H tend to be broad or multi-themed, leading to disagreement in both category framing and document fit. Disagreement in low H:TM cases appears to stem from model limitations—e.g., BERTOPIC's hard clustering approximation or MALLET defaulting to the best available topic despite poor fit.

BERTopic on `Bills` ($\alpha = 0.76,\ \tau = 0.48$)		
Topic Words	**Human Categories**	**LLM Category**
spirits, distilled, beer, wine, excise, brewers, cider, wines	• Distilled Goods Legislation • Alcohol Internal Revenue Code • LEGAL INVOICE • tax reform on alcohol products	Alcoholic Beverage Taxation and Regulation

Exemplar Documents:

- ***Doc ID** 7046: $\theta_d = 1.00$*
  Text: Amends the Internal Revenue Code to exclude from determination of the production period for distilled spirits any period allocated to the natural [...]
- ***Doc ID** 7886: $\theta_d = 1.00$*
  Text: Aged Distilled Spirits Competitiveness Act - Amends the Internal Revenue Code to exclude the aging period from the production period for distilled [...]
- ***Doc ID** 7487: $\theta_d = 1.00$*
  Text: Reinvesting in U.S.

Evaluation Documents:

- ***Doc ID** 6797: Human Fit $= 5.00 \pm 0.00$, LLM Fit $= 4.77$, $\theta_d = 1.00$*
  Text: Amends the Internal Revenue Code to reduce from $18 to 9$ (its pre-1991 level) the per-barrel tax on beer.
- ***Doc ID** 7335: Human Fit $= 4.75 \pm 0.43$, LLM Fit $= 4.13$, $\theta_d = 1.00$*
  Text: Amends the Internal Revenue Code to exclude from determination of the production period for distilled spirits any period allocated to the natural [...]
- ***Doc ID** 7916: Human Fit $= 4.75 \pm 0.43$, LLM Fit $= 5.00$, $\theta_d = 0.38$*
  Text: Cider Industry Deserves Equal Regulation Act or the CIDER Act
  Amends the Internal Revenue to revise the definition of "hard cider," for purposes [...]
- ***Doc ID** 7576: Human Fit $= 5.00 \pm 0.00$, LLM Fit $= 4.98$, $\theta_d = 0.31$*
  Text: Brewers Excise and Economic Relief Act of 2011 - Amends the Internal Revenue Code to: (1) reduce from $18 to 9$ (the pre-1991 level) the per-barrel [...]
- ***Doc ID** 7073: Human Fit $= 2.00 \pm 1.73$, LLM Fit $= 1.94$, $\theta_d = 0.28$*
  Text: Amends the Internal Revenue Code to increase the excise tax rate on small cigars to $19.50 per thousand (the same rate as for small cigarettes).
- ***Doc ID** 7927: Human Fit $= 4.75 \pm 0.43$, LLM Fit $= 4.99$, $\theta_d = 0.17$*
  Text: Distillery Excise Tax Reform Act of 2015
  Amends the Internal Revenue Code to allow a reduction (from $13.50 to 2.70$ on each proof gallon produced [...]
- ***Doc ID** 16: Human Fit $= 1.00 \pm 0.00$, LLM Fit $= 1.01$, $\theta_d = 0.00$*
  Text: Medicare Part D Drug Class Protection Act of 2007 - Amends part D (Voluntary Prescription Drug Benefit Program) of title XVIII (Medicare) of the [...]

Table 8: Case study of a topic with high human-to-human agreement (Krippendorff's $\alpha = 0.76$) but moderate-to-low human-to-LLM (GPT-4o) agreement (Kendall's $\tau = 0.48$), based on BERTOPIC applied to `Bills`. We show the top eight topic words, human- and LLM-generated categories, two exemplar documents, and all evaluation documents. Evaluation documents include human fit ratings (mean and standard deviation), model-assigned topic probabilities (θ_d), and LLM-assigned fit scores. While the τ value may suggest poor alignment, qualitative inspection shows strong agreement between humans and the LLM on individual document fit, indicating that even moderate τ values can correspond to good topics.

CTM on `Wiki` ($\alpha = 0.98,\ \tau = 0.58$)		
Topic Words	**Human Categories**	**LLM Category**
career, hit, games, season, league, baseball, major_league_baseball, signed	• Former MLB players • Professional baseball facts and figures • American baseball league	Major League Baseball Players and Achievements

Exemplar Documents:

- ***Doc ID** 2943:* $\theta_d = 0.61$
  Text: Brian Wilson (baseball) = Brian Patrick Wilson (born March 16 , 1982) is a former American professional baseball relief pitcher .
- ***Doc ID** 2928:* $\theta_d = 0.59$
  Text: Byron McLaughlin = Byron Scott McLaughlin (born September 29 , 1955) is an American retired professional baseball player , alleged counterfeit [...]
- ***Doc ID** 2989:* $\theta_d = 0.58$
  Text: Cy Seymour = James Bentley " Cy " Seymour (December 9 , 1872 – September 20 , 1919) was an American center fielder and pitcher in Major League [...]

Evaluation Documents:

- ***Doc ID** 2870: Human Fit* $= 5.00 \pm 0.00$*, LLM Fit* $= 3.97$*,* $\theta_d = 0.74$
  Text: Jon Lieber = Jonathan Ray Lieber (born April 2 , 1970) is a former Major League Baseball (MLB) pitcher .
- ***Doc ID** 3001: Human Fit* $= 5.00 \pm 0.00$*, LLM Fit* $= 4.37$*,* $\theta_d = 0.61$
  Text: Johnny Evers = John Joseph Evers (July 21 , 1881 – March 28 , 1947) was an American professional baseball second baseman and manager .
- ***Doc ID** 2932: Human Fit* $= 5.00 \pm 0.00$*, LLM Fit* $= 3.22$*,* $\theta_d = 0.49$
  Text: Bobo Holloman = Alva Lee " Bobo " Holloman (March 7 , 1923 – May 1 , 1987) was an American right-handed pitcher in Major League Baseball who [...]
- ***Doc ID** 2865: Human Fit* $= 5.00 \pm 0.00$*, LLM Fit* $= 5.00$*,* $\theta_d = 0.37$
  Text: Barry Bonds = Barry Lamar Bonds (born July 24 , 1964) is an American former professional baseball left fielder who played 22 seasons in Major [...]
- ***Doc ID** 4300: Human Fit* $= 1.33 \pm 0.47$*, LLM Fit* $= 1.00$*,* $\theta_d = 0.25$
  Text: Anthony Davis (basketball) = Anthony Marshon Davis , Jr .
- ***Doc ID** 5905: Human Fit* $= 1.00 \pm 0.00$*, LLM Fit* $= 1.00$*,* $\theta_d = 0.12$
  Text: 1879 Navy Midshipmen football team = The 1879 Navy Midshipmen football team represented the United States Naval Academy in the 1879 college football [...]
- ***Doc ID** 10610: Human Fit* $= 1.00 \pm 0.00$*, LLM Fit* $= 1.00$*,* $\theta_d = 0.00$
  Text: Edge (wrestler) = Adam Joseph Copeland (born October 30 , 1973) is a Canadian actor and retired professional wrestler .

Table 9: Same as Table 8, but for a topic from CTM trained on `Wiki`. This topic has very high human-to-human agreement ($\alpha = 0.98$) and moderate human-to-LLM agreement ($\tau = 0.58$). While both humans and the LLM associate the topic with baseball players, the LLM appears to emphasize notable achievements, leading to slightly lower fit scores on documents like *Doc ID 2932*, which reads as a biography of a less prominent player. The disagreement here is minimal but highlights how subtle differences in topic scope can influence agreement.

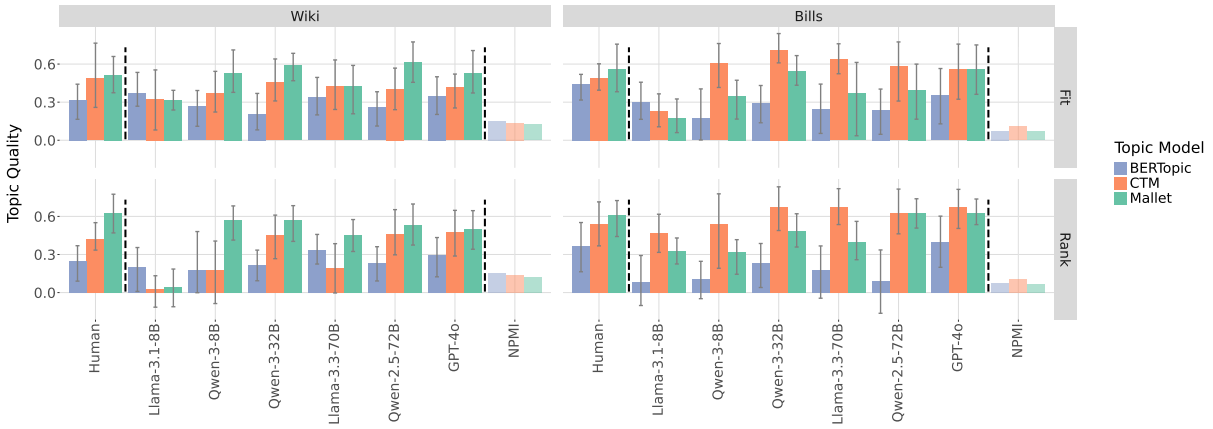


Fig. 9: Model rankings based on FIT–$\tau_{h:tm}$/ RANK–$\tau_{h:tm}$ (correlation of human scores with document-topic probabilities), FIT–$\tau_{lm:tm}$/RANK–$\tau_{lm:tm}$ (correlation of LLM scores with document-topic probabilities), and NPMI coherence. Correlations are computed using Kendall's τ. Error bars for Human and LLM metrics are 95% bootstrapped confidence intervals, resampled over topics. CTM performs best on `Bills`, while MALLET leads on `Wiki`. LLM-based metrics align well with human judgments, unlike NPMI. Rankings are consistent across LLMs.

	τ (FIT–$\tau_{h:tm}$, ·)		τ (RANK–$\tau_{h:tm}$, ·)	
	`Wiki`	`Bills`	`Wiki`	`Bills`
NPMI	-0.04	-0.03	0.04	0.03
FIT/RANK–$\tau_{lm:tm}$				
Llama-3.1-8B	0.25 (0.17)	0.04 (0.16)	-0.01 (0.21)	0.09 (0.14)
Qwen-3-8B	0.47 (0.17)	0.22 (0.19)	0.44 (0.15)	0.22 (0.14)
Qwen-3-32B	0.33(0.20)	0.45 (0.17)	0.41 (0.15)	0.48 (0.12)
Llama-3.3-70B	0.38 (0.19)	0.37 (0.15)	0.36 (0.20)	0.32 (0.16)
Qwen-2.5-72B	0.59 (0.13)	0.56 (0.12)	0.40 (0.19)	0.25 (0.16)
GPT-4o	0.60 (0.15)	0.59 (0.15)	0.26 (0.21)	0.26 (0.16)
FIT/RANK–$\tau_{h:tm}$				
Human	0.40 (0.08)	0.33 (0.08)	0.14 (0.12)	0.20 (0.11)

Table 10: Relationship between automated and human topic rankings given *binary* topic assignments. Cells show Kendall's τ between metrics: FIT/RANK–$\tau_{h:tm}$ correlates human scores to binary assignments: $\mathbb{I}[\text{argmax}(\boldsymbol{\theta}_k) = k]$. FIT/RANK–$\tau_{lm:tm}$ correlates PROXANN to binary assignments. Values are bootstrapped means and standard deviations (resampling over topics). The *Human* row reflects leave-one-out inter-annotator correlations, serving as a reference.

L User Interface

Figures 10 to 13 are screenshots of the annotation interface presented to users. Figure 14 is the consent page shown at the start.

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).

Fig. 10: Instructions for the human annotation protocol.

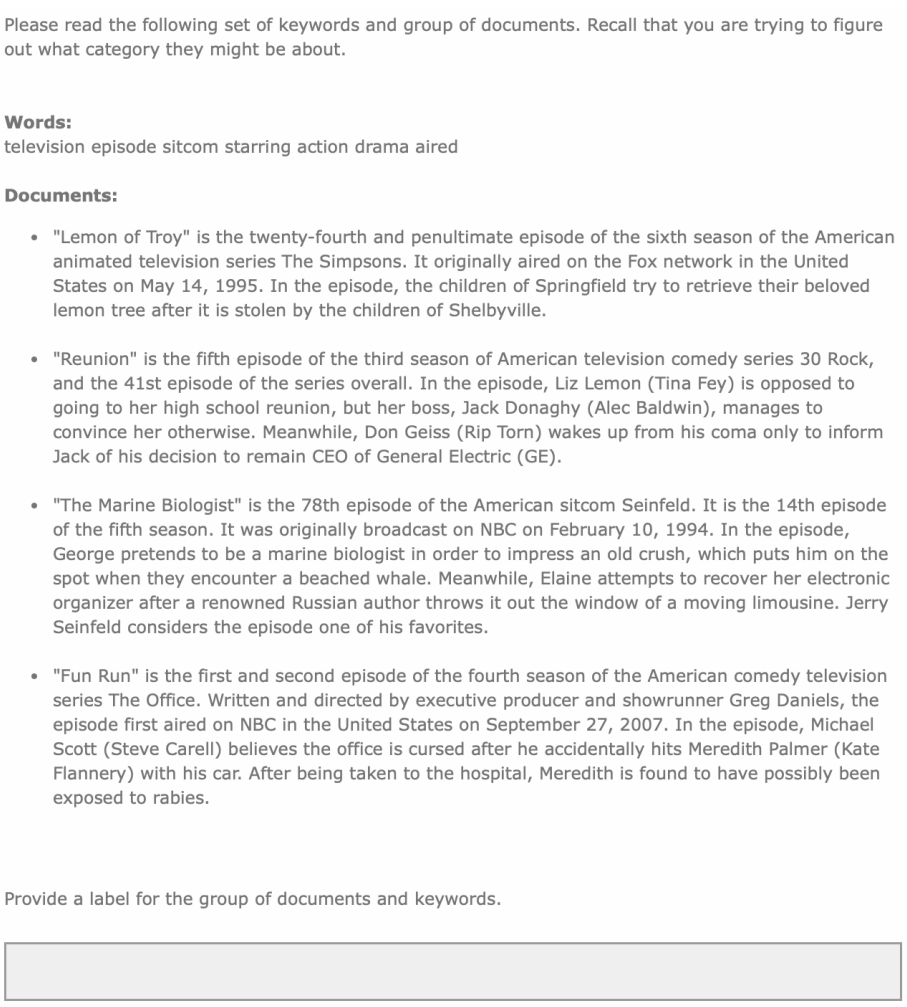


Fig. 11: Label Step. Category identification in the human annotation protocol for the practice question.

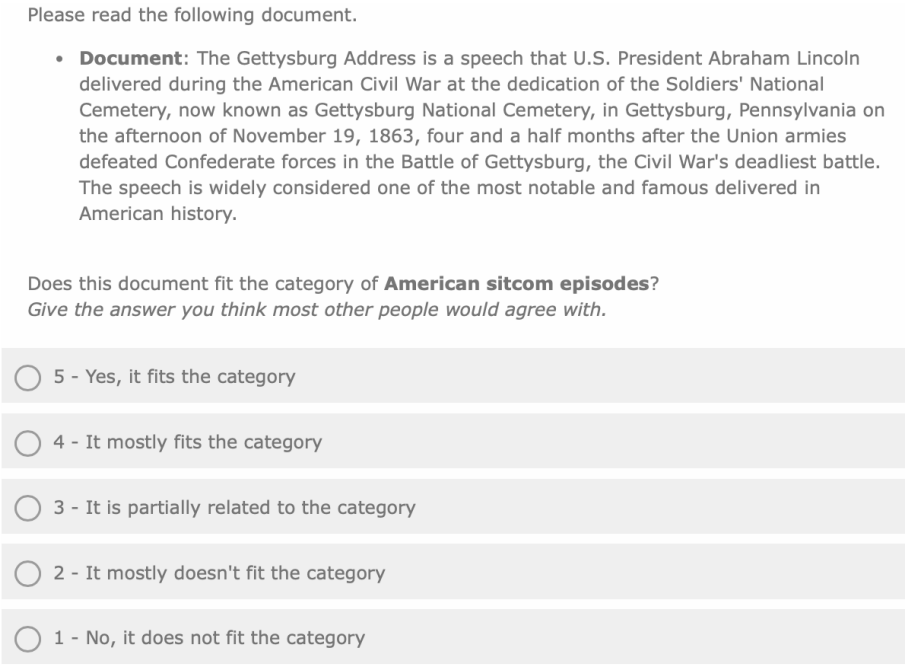


Fig. 12: Fit Step. Relevance judgment in the human annotation protocol for the practice question.

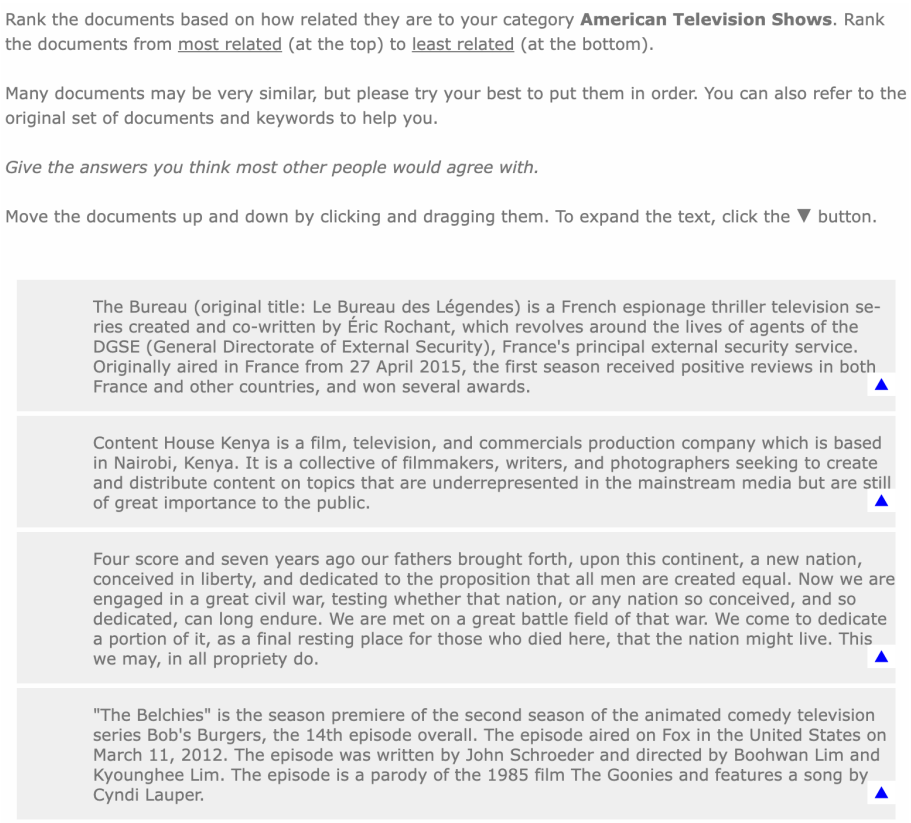


Fig. 13: Rank Step. Representativeness ranking in the human annotation protocol for the practice question.

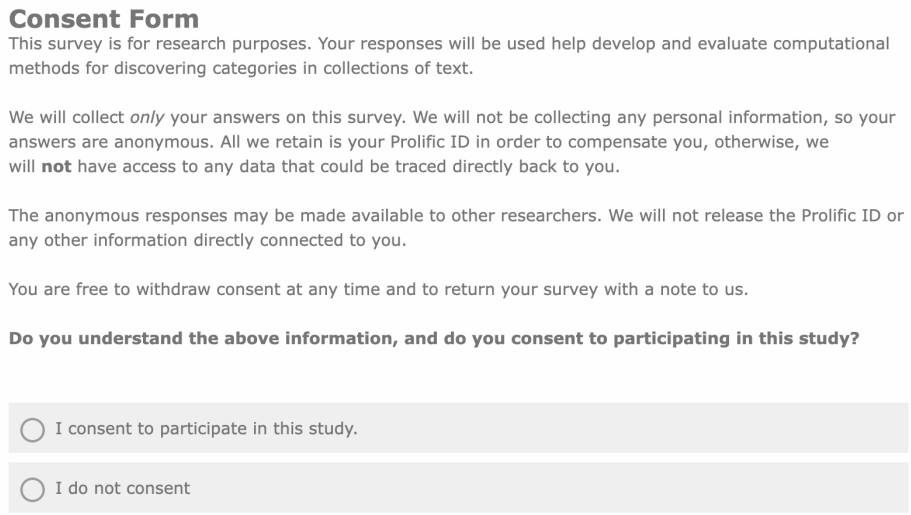


Fig. 14: Consent page (shown at beginning)

M Prompt templates

M.1: Category Identification (Label Step, System Prompt)

```
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.
```

M.2: Category Identification (Label Step, Instruction Prompt)

```
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:
```

M.3: Relevance Judgment (Fit Step)

```
Please act as an impartial judge and assign an integer score from 1 to 5 indicating how well the
    DOCUMENT fits the given CATEGORY. Do not provide any reasoning or explanation

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

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

M.4: Representativeness Pairwise Ranking (Rank Step)

```
Please act as an impartial judge and determine which of the two documents (A or B) is more closely
    related to the given CATEGORY. Avoid any positional bias, and ensure that the order in which
    the documents are presented does not influence your decision. Output your verdict strictly
    using this format: 'A' if DOCUMENT_A is more closely related to the CATEGORY, or 'B' if
    DOCUMENT_B is more closely related.

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

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

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