Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, **Jordan Boyd-Graber**, and Philip Resnik. **Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence**. *Neural Information Processing Systems*, 2021, 10 pages.

```
    @inproceedings{Hoyle:Goel:Peskov:Hian-Cheong:Boyd-Graber:Resnik-2021,
Title = {Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence},
Author = {Alexander Hoyle and Pranav Goel and Denis Peskov and Andrew Hian-Cheong and Jordan Boyd-Graber a
Booktitle = {Neural Information Processing Systems},
Location = {Online},
Year = {2021},
Url = {http://umiacs.umd.edu/~jbg//docs/2021_neurips_incoherence.pdf},
}
```

Accessible Abstract: Topic models help historians, journalists, and analysts make sense of large text collections. But how do you know if you have a good one? The field has settled on using "Automatic Coherence", but this paper argues that maybe that isn't the right choice if you want to actually make real users happy.

Links:

- ArXiv [https://arxiv.org/abs/2107.02173]
- Research Talk [https://youtu.be/op1DkSB2VdA]
- Code [http://hithub.com/ahoho/topics]

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

Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence

Alexander Hoyle[*] **Pranav Goel**[*] **Denis Peskov**[*] **Andrew Hian-Cheong**[*]
Computer Science

Jordan Boyd-Graber CS, iSchool, UMIACS, LSC
Philip Resnik UMIACS, Lingusitics

University of Maryland
{hoyle,pgoel1,dpeskov,andrewhc,jbg,resnik}@cs.umd.edu

Abstract

Topic model evaluation, like evaluation of other unsupervised methods, can be contentious. However, the field has coalesced around automated estimates of topic coherence, which rely on the frequency of word co-occurrences in a reference corpus. Contemporary neural topic models surpass classical ones according to these metrics. At the same time, topic model evaluation suffers from a *validation gap*: automated coherence, developed for classical models, has not been validated using human experimentation for neural models. In addition, a meta-analysis of topic modeling literature reveals a substantial *standardization gap* in automated topic modeling benchmarks. To address the validation gap, we compare automated coherence with the two most widely accepted human judgment tasks: topic rating and word intrusion. To address the standardization gap, we systematically evaluate a dominant classical model and two state-of-the-art neural models on two commonly used datasets. Automated evaluations declare a winning model when corresponding human evaluations do not, calling into question the validity of fully automatic evaluations independent of human judgments.

1 Revisiting Topic Model Evaluation

Topic models are a machine learning technique widely used outside computer science, including political science (Grimmer and Stewart, 2013; Isoaho et al., 2021), social and cultural studies (Mohr and Bogdanov, 2013), digital humanities (Meeks and Weingart, 2012), and bioinformatics (Liu et al., 2016). Typically, topic model users are domain experts trying to identify global categories or themes present in a document collection (Boyd-Graber et al., 2017). This practice constitutes a computer-assisted form of content analysis (Krippendorff, 2004; Chuang et al., 2014), also related to distant reading in literary studies (Underwood, 2017). In general, topic models help humans understand large corpora.[2]

Evaluation of topic models has vacillated between automated and human-centered. While real-world users of topic models evaluate outputs based on their specific needs, topic model developers have gravitated toward generalized, automated proxies of human judgment to help inform rapid iteration of models (Doogan and Buntine, 2021). Initially, models were evaluated with held-out perplexity, but it disagrees with human interpretability (Chang et al., 2009). Consequently, the field adopted automated coherence metrics like normalized pointwise mutual information (NPMI), a measure of word relatedness that *does* correlate with topic interpretability (Section 2.2; Newman et al., 2010; Aletras and Stevenson, 2013; Lau et al., 2014). The balance shifted towards automated coherence.

Human evaluations have been abandoned by topic model developers in the years since automated coherence metrics were adopted. In a thorough meta-analysis of contemporary topic model methods

[*]Equal contribution

[2]Topic models are also used for other purposes, such as information retrieval or downstream document classification. However, the discovery and application of categories for human interpretation is their dominant use, and other computational applications have been largely eclipsed by modern neural approaches.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

	Classical			Neural		
	station line bridge railway trains	album band music song released	tropical storm hurricane cyclone depression	tropical landfall cyclone utc weakening	spore basidia spores mycologist hyphae	manhattan_project los_alamos_laboratory robert_oppenheimer enrico_fermi physicist
NPMI	0.274	0.285	0.394	0.446	0.456	0.470

Table 1: The first three columns are the highest-NPMI topics for a classical topic model (LDA estimated via Gibbs sampling using Mallet, McCallum, 2002; Griffiths and Steyvers, 2004). The next three are counterparts from a neural model (our D-VAE reimplementation, Burkhardt and Kramer, 2019). Models are trained on Wikitext (Merity et al., 2017) with fifty topics, and NPMI is estimated over the top five words in each topic using a 4.6M-document reference Wikipedia corpus. The mean top-five NPMI over all topics is 0.156 for the classical and 0.256 for the neural model.

papers, *none* conduct systematic human evaluations (Section 3). Instead, they rely solely on automated metrics for model comparison.[3] However, current neural topic models are a far cry from the classical models that substantiated the original correlations—manifestly, topics produced by neural models are often qualitatively distinct from those of classical models (e.g., Table 1).[4] This *validation gap* raises the question of whether automated metrics are still consistent with human judgments of topic quality.

Moreover, we should always be cautious when extrapolating outside the range of data that was used to establish a relationship between variables. As an example, a neural model in Hoyle et al. (2020) produces much larger NPMI values than those used to determine human correlations in the original Lau et al. (2014) study; the implicit assumption is that greater NPMI corresponds to more human-interpretable topics. Finally, a myopic focus on a presumed proxy for human preferences can produce low-quality results (Stiennon et al., 2020). Does Goodharts' law—"when a measure becomes a target, it ceases to be a good measure" (Strathern, 1997)—apply to automated metrics of topic models?

Another challenge for automated evaluation, whether of classical or neural topic models, is widespread inconsistency (Section 3). Researchers frequently fail to specify the information needed to calculate automated metrics or diverge from the practices that underpin human correlations. Furthermore, evaluation datasets, preprocessing, and hyperparameter optimization vary dramatically, even within a given paper. This *standardization gap* likely limits the generalizability and reliability of topic model developers' findings.

We address the standardization and validation gaps in topic model evaluation:

1. We present a meta-analysis of neural topic model evaluation (Section 3);
2. we develop standardized, pre-processed versions of two widely-used English-language evaluation datasets, along with a transparent end-to-end code pipeline for reproduction of results (Section 4.1)[5];
3. we optimize three topic models—one classical and two neural—using identical preprocessing, model selection criteria, and hyperparameter tuning (Section 4.2);
4. we evaluate these models using human ratings and word intrusion tasks (Section 5); and
5. we provide new evaluations of the correlation between automated and human evaluations (Section 6).

Our findings challenge the validity of fully-automated evaluations as currently practiced: automated evaluation declares winners between models when the corresponding human evaluations cannot.

[3]Outside of the core method-development literature, human evaluations have been used to develop new metrics and improve understanding of existing model behavior (Bhatia et al., 2017; Morstatter and Liu, 2018; Lund et al., 2019; Alokaili et al., 2019, *inter alia*).

[4]We use "classical" to mean generative models defined by a chain of conjugate exponential family distributions optimized by Gibbs sampling or variational inference.

[5]`github.com/ahoho/topics`

2 Operationalizing Topic Coherence

A topic model is a probabilistic generative model of text that uses latent *topics* to summarize a larger collection of documents. The most influential variant, latent Dirichlet allocation (Blei et al., 2003, LDA), assumes that K latent topics are distributions over word types, $\boldsymbol{\beta}_k$, and that the documents $\mathcal{D}$ are admixtures over the topics, $\boldsymbol{\theta}_d$. Users often evaluate model outputs globally, focusing on the most probable N words of each topic, and locally, considering the most probable topics for each document.

While techniques for topic modeling have progressed from variational inference (Blei et al., 2003) to Gibbs sampling (Griffiths and Steyvers, 2004) to deep generative approaches (Srivastava and Sutton, 2017; Wang et al., 2020b), the core goal discussed in Section 1, obtaining human-understandable categories, remains central. The latest wave of methods, *neural topic models* (NTM), use continuous word representations and gradient optimization to fit parameters. These models claim to produce more interpretable topics than other prior methods, including LDA.

Those claims are supported by improvements on automated measures of topic coherence.

2.1 Human Metrics of Topic Coherence

Like the concept of *interpretability*, that of real-world *coherence* is "simultaneously important and slippery" (Lipton, 2018). We will not attempt to formalize it here—though see discussion in Section 7. For present purposes, the term has its roots in Latin *cohaerere*, "to stick together," and we will think of coherence as an intangible sense, available to human readers, that a set of terms, when viewed together, enable human recognition of an identifiable category.[6] We review two human ratings of topic quality: direct ratings and intrusion.

Rating Raters see a topic and then give the topic a quality score, conventionally on a three-point ordinal scale (Newman et al., 2010; Mimno et al., 2011; Aletras and Stevenson, 2013, *inter alia*).

Intrusion Chang et al. (2009) devise the *word intrusion* task as a behavioral way to assess topic coherence. The core idea is that when the top words in a topic identify a coherent latent category, it is easier to identify words that do not belong to that category. Operationally, each topic is represented as its top words plus one "intruder" word which has a low probability of belonging to that topic, but a high probability of belonging to a different topic. Topic coherence is then judged by how well human annotators detect the "intruder" word.

2.2 NPMI: The Standard Automated Topic Model Coherence Evaluation

Using the word intrusion task, Chang et al. (2009) showed that perplexity—the original topic model evaluation metric—*negatively* correlates with human evaluations of topic quality. This finding revealed a need for an automated measurement of topic coherence: an automated metric can measure model quality without expensive, time-consuming, and difficult-to-reproduce human experiments.

Lau et al. (2014) find some metrics that *positively* correlate with human intrusion and rating scores, particularly when aggregating scores over all topics from a given model. Because of that validation, the prevailing evaluation for model comparison is pairwise normalized pointwise mutual information. NPMI scores topics highly if the top N words—summed over all pairs w_i and w_j—have high joint probability $P(w_j, w_i)$ compared to their marginal probability:[7]

$$\sum_{j=2}^{N}\sum_{i=1}^{j-1} \frac{log \frac{P(w_j,w_i)}{P(w_i)P(w_j)}}{-logP(w_i, w_j)}. \tag{1}$$

The probabilities are estimated using word co-occurrence counts from a *reference corpus* for a specific context window (which can range from ten words to the entire document). As a result, the choice of reference corpus determines the strength of human correlation (Lau et al., 2014; Röder et al., 2015).

[6] This perspective aligns with Propositions 2 and 3 of Doogan and Buntine (2021): "an interpretable topic is one that can be easily labeled," and "has high agreement on labels."

[7] Alternative metrics exist, but they typically also rely on either joint probability estimates or NPMI directly (e.g., C_v Röder et al., 2015).

Evaluation	Count	
Number of human evaluations	0	(0%)
Automated Coherence		
Metric		
NPMI	26	(72%)
Other	22	(61%)
Explicit implementation	22	(61%)
Explicit ref. corpus	10	(28%)
Perplexity w/o coherence	3	(8%)

Experimentation	Count	
Preprocessing		
Inconsistent over datasets	12	(30%)
Ambiguous preprocessing	9	(23%)
Model comparisons		
All models tuned	5	(13%)
Unclear h.param search	16	(40%)
Unclear LDA baseline, if used	7	(24%)
Recent baseline (w/in 2 yrs)	31	(78%)
Multiple runs / sig. testing	11	(28%)

Table 2: Meta-analysis of forty neural topic modeling papers (denominator may change, as not all conditions are applicable). No recent neural topic modeling papers use human evaluations of coherence, and the metrics and models are difficult to replicate.

A measurement is *valid* to the extent that it measures what it is intended to measure in the real world. Historically, automated coherence has been validated using *human* judgements from either crowdworkers (Newman et al., 2010; Aletras and Stevenson, 2013) or experts (Mimno et al., 2011). However, correlations based on classical models may not be applicable for NTMs. Our skepticism is motivated by theory, as neural word representations are intimately connected to NPMI, as explicitly used by Aletras and Stevenson (2013) and which produce similar NPMI scores as Lau et al. (2014). Levy and Goldberg (2014) show that multiple representations create factorizations of PMI matrices. Topic models that have access to these rich representations (e.g. Dieng et al., 2020, and others) could thus create topics with good NPMI scores without explaining the corpus well to a user. In contrast to classical topic models, no one has investigated the validity of NPMI evaluation for NTMs.

Given this lacuna, we conduct experiments aimed at validating that automated topic evaluations still correlate with human judgments of neural topic model quality. We compare against two common human evaluations of individual topic quality: direct rating and intrusion. Human evaluations, like automated topic modeling, lack standardization, which we address in Section 5.

3 A Meta-Analysis of Neural Topic Modeling

We survey the neural topic modeling (NTM) literature to assess the state of evaluation in contemporary topic model development. First, we take all references made by an existing, comprehensive survey of NTMs (Zhao et al., 2021b), from which we select (a) modeling papers which (b) mention topic interpretability and (c) compare models' topics with an existing baseline. This yields forty models, which all claim superior topic coherence. We examine data processing steps, hyperparameter tuning, baseline selection, and automated coherence calculations. Table 2 summarizes our results and Appendix A.1 enumerates the papers.

Our analysis reveals variance in all areas. Preprocessing, which can significantly affect model quality and automated metrics, is often (30%) inconsistent across datasets within the same paper. When preprocessing *is* consistent, authors omit details necessary to fully replicate the pipeline. These issues imply that automated metrics for the same baselines and source datasets vary across papers. Compounding the problem, researchers often train their models on different datasets from those used to establish the relationships between human annotations and automated metrics; Doogan and Buntine (2021) find that the same metrics may not predict interpretability in new domains. Mirroring findings from Dodge et al. (2019), 40% of papers fail to clearly specify their model tuning procedure, often even the metric used for model selection.

Calculation of automated coherence metrics is equally fraught. As discussed in Section 2.2, a complete specification for NPMI involves several pieces of information, including the reference corpus used to estimate joint word probabilities, the co-occurrence window size, and the number of words selected from the head of the topic distribution. Three out of four papers fail to explicitly indicate the reference corpus; even when we can assume the input corpus is used (13 cases), it remains uncertain whether authors use, e.g., a held-out set or the training documents themselves. For the 61% that specify the implementation of their coherence metric (by pointing to a code repository or writing out the formula), some of these factors may still be in question. For instance, six authors reference Lau et al. (2014) and the supporting code,[8] but the implications are ambiguous: the original paper suggests a large corpus from the same source as the training data, but the repository script defaults to Wikipedia. In other cases, authors use bespoke implementations, which creates room for errors, or deviate from the settings used in human experiments. For example, several papers use a document-wide context window with NPMI, which has not been correlated with human judgments.

Last, *even if* automated evaluations are consistent, all claims of coherence improvement depend on the validity results in Lau et al. (2014) generalizing to neural topic models.

4 Closing the Standardization Gap for Topic Models

Our human evaluation of topic model outputs serves multiple purposes: (a) establishing whether NTMs show improved coherence over a classical baseline and (b) re-evaluating the efficacy and reliability of automated coherence metrics. In addition, a key goal is (c) to provide a standardized preprocessing pipeline to support head-to-head comparisons as new methods are developed.[9]

We identify two commonly-used datasets, which we in turn process using a standard pipeline. We then estimate topic models on each dataset following a computationally fair hyperparameter search. Our standardization efforts are similar to concurrent work by Terragni et al. (2021); the main differences are that we (a) mandate consistent preprocessing between training and reference corpora, (b) support multi-word expressions during vocabulary creation (see below), and (c) support distributed hyperparameter searches.

4.1 Datasets and Preprocessing

Following Chang et al. (2009), we use English articles from Wikipedia and the New York *Times* (Table 7). For Wikipedia, we use Wikitext-103 (WIKI, Merity et al., 2017), and for the *Times*, we subsample roughly 15% of documents from LDC2008T19 (NYT, Sandhaus, 2008), making it an order of magnitude larger than WIKI. To compute reference counts, we use a 4.6M document Wikipedia dump from September 2017 and the full 1.8M document LDC2008T19 set, processed identically to the training data.

We use SpaCy (Honnibal et al., 2020) to tokenize and identify entities in the text. We create new tokens for detected entities of the form `New_York_City`, per Krasnashchok and Jouili (2018). Schofield and Mimno (2016) find that lemmatization and word-stemming can hurt English topic interpretability, so we do not lemmatize. To maintain a roughly equal vocabulary size over datasets, we use a power-law relationship of corpus size (c.f. Zipf, 1949) to rule out tokens occurring in fewer than a given number of documents.[10] In addition to a standard stopword list, we define corpus-specific stopwords as tokens appearing in more than 90% of documents. See Appendix A.2 for complete preprocessing details.

4.2 Models

We evaluate one venerable classical model and two newer neural models:

Gibbs-LDA As a strong classical baseline, we use the widely-loved Mallet (McCallum, 2002) implementation of Gibbs-sampling for LDA (Griffiths and Steyvers, 2004). Mallet produces topics of (qualitatively) competitive quality to neural models (Srivastava and Sutton, 2017).

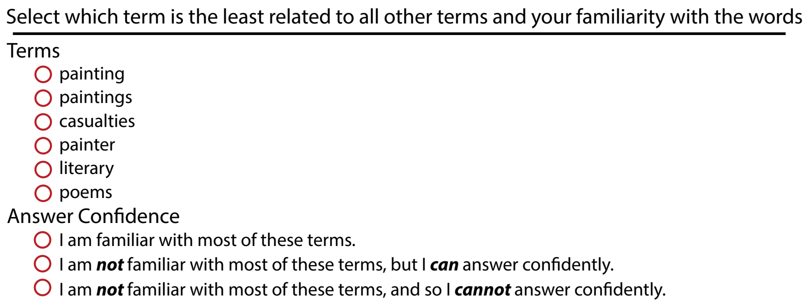
Dirichlet-VAE We reimplement Dirichlet-VAE (Burkhardt and Kramer, 2019), a state-of-the-art NTM. For simplicitly, we use pathwise gradients for the Dirichlet (Jankowiak and Obermeyer, 2018), rather than the rejection sampling variational inference of the authors' primary variant.[11] Dirichlet-VAE is a wholesale improvement on one of the first successful NTMs, the popular ProdLDA (Srivastava and Sutton, 2017), and is competitive against recent models on automated coherence. The generative

[8] `github.com/jhlau/topic_interpretability`

[9] Our preprocessing pipeline is agnostic to dataset and easily portable. `github.com/ahoho/topics`

[10] We target vocabularies approximating the number of words known by an adult English-speaker (Brysbaert et al., 2016): roughly 40k for WIKI and 35k for NYT.

[11] We replicate their NPMI and redundancy scores on 20 newsgroups. `github.com/ahoho/dvae`


Select which term is the least related to all other terms and your familiarity with the words

Terms
- painting
- paintings
- casualties
- painter
- literary
- poems

Answer Confidence
- I am familiar with most of these terms.
- I am ***not*** familiar with most of these terms, but I ***can*** answer confidently.
- I am ***not*** familiar with most of these terms, and so I ***cannot*** answer confidently.

Figure 1: The word intrusion task presented to crowdworkers (the ratings task is in Appendix A.4).

model is simple and retains a broad similarity to LDA. The primary difference is that it does not constrain the estimated topic-word distributions to the simplex.

ETM Thanks to their improved flexiblity, many NTMs incorporate external word representations, on the premise that large-scale, general language knowledge improves topic quality (Bianchi et al., 2021; Hoyle et al., 2020). The Embedded Topic Model (Dieng et al., 2020) is a popular NTM that relies on word embeddings in its generative model.[12]

We maintain a fixed computational budget per model following the exhortation of Dodge et al. (2019) and use a random set of 164 hyperparameter settings across datasets for each model type.[13] We train models for a variable number of steps (a hyperparameter); to calculate automated coherence for the model, we use the topics produced at the last step. For human evaluations, we select the models that maximize NPMI, estimated using the reference corpus with a ten-word window over the top ten topic words, per Lau et al. (2014). We follow the recommendation of Dieng et al. (2020) and learn skip-gram embeddings on the training corpus for ETM (experiments with external pretrained embeddings did not yield substantially different results). As in Hoyle et al. (2020), we eliminate models with highly redundant topics, a known degeneracy of NTMs (Burkhardt and Kramer, 2019): (a) models in which any of the top five words of one topic overlap with another and (b) models that have a topic uniqueness score (Nan et al., 2019) above 0.7. Ranges for hyperparameters and other details are in Appendix A.3.

5 Human Evaluations of Topic Quality

We use the *ratings* and *word intrusion* tasks from Section 2.2 as human evaluations of topic quality. We recruit crowdworkers using Prolific.co, an online panel provider and collect data with the Qualtrics survey platform. We pay workers 2.5 USD per ratings survey and 3 USD per word intrusion survey, equivalent to 15 USD/hour.

In order to draw meaningful conclusions from human annotations, we require an adequate number of participants to ensure acceptable statistical power. However, Card et al. (2020) show that many NLP experiments, including those relying on human evaluation, are insufficiently powered to detect model differences at reported levels. Adopting a straightforward generative model of annotations (Appendix A.5), we select enough crowdworkers per task to ensure sufficient statistical power (at least $1-\beta = 0.9$) to obtain significance at $\alpha = 0.05$, resulting in a minimum of fifteen crowdworkers per topic for both tasks. On this criterion, both Chang et al. (2009) and thus Lau et al. (2014), with eight annotators, are underpowered.

For each of our two datasets, we generate fifty topics each from the three models in Section 4.2. In the word intrusion task, we sample five of the top ten topic words plus one intruder; for the ratings task, we present the top ten words in order (Figure 4). We separate the datasets for each task and

[12] `github.com/adjidieng/ETM`

[13] While runtimes can vary drastically by model, this study is not concerned with implementation efficiency (although efficiency matters, see Ethayarajh and Jurafsky, 2020).

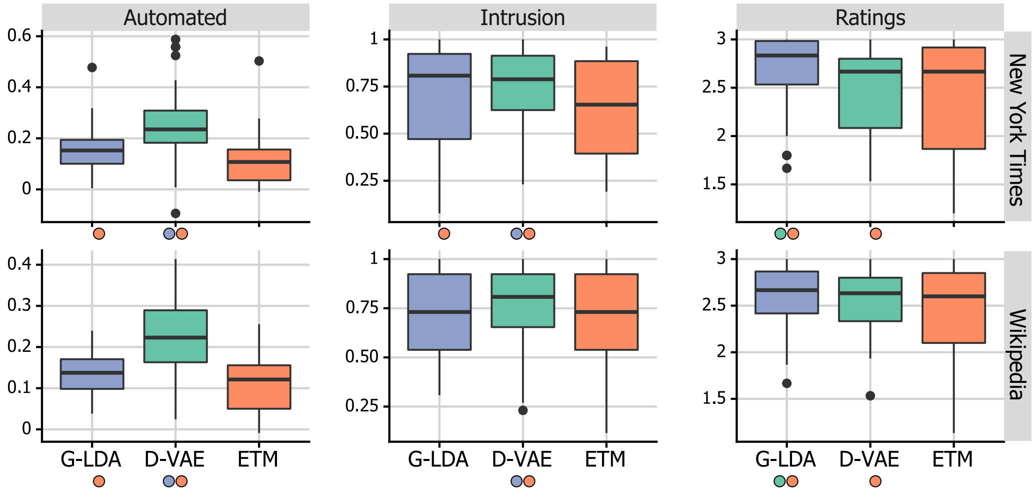


Figure 2: While automated evaluations (here, NPMI) suggest a clear winner between models, human evaluation is more nuanced. Human judgments exhibit greater variability over a smaller range of values. Colored circles correspond to pairwise one-tailed significance tests between model scores at $\alpha = 0.05$; for example, the rightmost orange circle at bottom right shows that human intrusion ratings for D-VAE are significantly higher than ETM for topics derived from Wikipedia.

randomly sample 40 of the 150 topics. In the ratings task, we include an additional sixteen synthetic poor-quality topics to help calibrate scores and filter out low-quality respondents.[14]

Phrasing of questions closely follows the wording used by Chang et al. (2009), and crowdworkers received detailed instructions with examples (Appendix A.4) before responding to items.[15] As topics can be esoteric (e.g., last columns of Table 1), we ask crowdworkers about their familiarity with the words in each question. We speculate that this question can help protect against spurious low scores for otherwise coherent topics, as real-world users of topic models are usually familiar with domain-specific terminology (see further discussion in Section 7).

6 Human Judgment Differs From Automated Metrics

We compare human judgments to automated methods on topics estimated using our three models.

6.1 Human Assessment

To establish model differences using human ratings, we use pairwise significance tests: a proportion test for the intrusion scores, a U test (Mann and Whitney, 1947) for the ratings, and a t-test for automated metrics (Figure 2), using one-tailed tests for each pair in both directions. Although D-VAE fares better on the intrusion task, evaluation using ratings favors G-LDA.[16]

Our human evaluation results are consistent with past iterations of the ratings and word intrusion tasks for topic models. Mimno et al. (2011) report an average of 2.36 on the ratings task on a dataset of medical paper abstracts.[17] Our ratings means are 2.5 to 2.8 across all variations (Figure 2). Our word intrusion means range from 0.7 to 0.8, which is comparable to the roughly 0.8 accuracy on the LDA model evaluated in Chang et al. (2009). Median time taken on the tasks was 8–9 minutes.

[14]For generating *synthetic poor-quality topics*, we use random high-probability words appearing in topics from other hyperparameter settings, but that have low probability among selected topics. Eight topics each are generated from the vocabularies of NYT and WIKI.

[15]Code to convert topic model output into deployable questionnaires is at `github.com/ahoho/topics`.

[16]These discrepancies among human tasks support the argument that standard coherence metrics alone may be insufficient for automated model selection (Doogan and Buntine, 2021).

[17]Newman et al. (2010) and Lau et al. (2014) do not report an average.

	Ref. Corpus → Train Corpus ↓	NPMI (10-token window) NYT	WIKI	Train	Val	C_v (110-token window) NYT	WIKI	Train	Val
Intrusion	NYT	0.27	0.43	0.27	0.24	0.34	**0.45**	0.35	0.34
	WIKI	0.34	0.36	**0.39**	0.17	0.32	0.34	0.34	0.20
	Concatenated	0.29	0.40	0.32	0.17	0.32	**0.40**	0.35	0.24
Rating	NYT	0.37	**0.48**	0.37	0.39	0.41	0.46	0.44	0.45
	WIKI	0.34	0.41	**0.44**	0.28	0.32	0.40	0.40	0.34
	Concatenated	0.37	**0.44**	0.41	0.35	0.38	0.42	0.42	0.42

Table 3: Spearman correlation coefficients between mean human scores and automated metrics. Underlined values have overlapping bootstrapped 95% confidence intervals with that of the **largest** value in each row. "Concatenated" refers to correlations computed on a concatenation of values for the NYT and WIKI items. "Val" is a small held-out set of 15% of the training corpus. Using the more data-appropriate logistic and ordered probit regressions for word intrusion and ratings data leads to different conclusions about relative metric strength (Appendix Table 10). CIs are estimated using 1,000 samples.

Following Aletras and Stevenson (2013), we calculate inter-annotator agreement with the mean Spearman correlation between each respondent's score per topic and the average of other respondent scores, obtaining a value of 0.75 (compare to their value of 0.7 on the NYT corpus). Additionally, we include synthetic poor-quality topics (footnote 14)—correctly identified by annotators—and we monitor the duration taken for the survey to hedge against insincere submissions.

6.2 Automated Metrics

NPMI declares D-VAE the unequivocal victor among the three models (with G-LDA a clear second), a very different story from the human judgments. To understand the relationship between automated metrics and human ratings, we estimate the Spearman correlation between the two sets of values for each task and dataset for metric variants (Table 3). Although previous studies have used mean human ratings over topics, this decision obscures the inherent variance of the human ratings and leads to overconfident estimates. We therefore construct 95% confidence intervals by resampling ratings, with replacement, equal to the number of annotators per task (Table 3). We estimate NPMI with the standard 10-word window and C_v (Röder et al., 2015) with the recommended 110-word window.[18] The Wikipedia corpus appears to be best correlated with human judgments, even for the models trained on the NYT corpus—this contradicts Lau et al. (2014), where within-domain data have the highest correlations.

While all correlation coefficients are statistically significant, the strength of the correlation alone does not justify their use in model selection, as is standard in the NTM literature (Section 3). In particular, the inherent uncertainty of human judgments means that it is difficult to determine when an increase in a model's mean automated coherence implies a significant improvement in the corresponding human scores.[19]

As noted above (Figure 2), automated metrics exaggerate model differences compared to human judgments. To help clarify the utility of automated metrics for model selection, we ask how often an automated metric incorrectly asserts that one model is superior to another. To do so, we generate a bootstrapped estimate of the false discovery rate of each model. First, for each dataset, we randomly sample two independent sets of $K = 50$ topics (without replacement) from the original pool of 150, along with their corresponding automated and human scores (resampled with replacement, as in Table 3). Treating the two sampled sets as outputs from two different models, we compute pairwise significance tests between each set for both the K automated metrics and $K \times M$ human scores (using a proportions z-test for the intrusion scores and t-tests for all other values). After repeating this process for $N = 1000$ iterations, we report the proportion of significant differences detected using

[18] We use gensim (Řehůřek and Sojka, 2010) to calculate coherence. We process the reference corpora identically to the training data, retaining only terms that exist in the training vocabulary. Other metrics, like C_{UCI} (Newman et al., 2010) and C_{UMASS} (Mimno et al., 2011), show low correlations.

[19] Better models of human scores could help quantify this relationship (e.g., GLMs, see Appendix A. 10).

	Ref. Corpus → Train Corpus ↓	NPMI (10-token window) NYT	WIKI	Train	C_v (110-token window) NYT	WIKI	Train
Intrusion	NYT	46 / 53	34 / 48	48 / 50	35 / 38	**30** / 29	34 / 35
	WIKI	44 / 76	33 / 78	33 / 75	45 / 48	38 / 49	**37** / 45
	Concatenated	42 / 67	40 / 66	41 / 64	36 / 46	**31** / 44	30 / 45
Rating	NYT	45 / 50	45 / 51	41 / 47	27 / 29	26 / 26	**21** / 26
	WIKI	40 / 73	31 / 73	33 / 71	38 / 40	31 / 40	**28** / 34
	Concatenated	39 / 66	36 / 66	37 / 62	31 / 38	28 / 38	**19** / 36

Table 4: False discovery rate (1−precision, lower is better) and false omission rate of significant model differences when using automated metrics; automated metrics often overstate meaningful model differences. **Bolded** values are those with the lowest geometric mean of FDR and FOR. We sample two independent sets of 50 topics along with their human scores and automated metrics; these sets act as the outputs of two "models". We then compute significance tests between sets (per Figure 2) on both the automated scores and human scores. A false positive occurs when one set has significantly larger automated scores despite no meaningful difference in actual human scores. Estimates are over 1,000 samples.

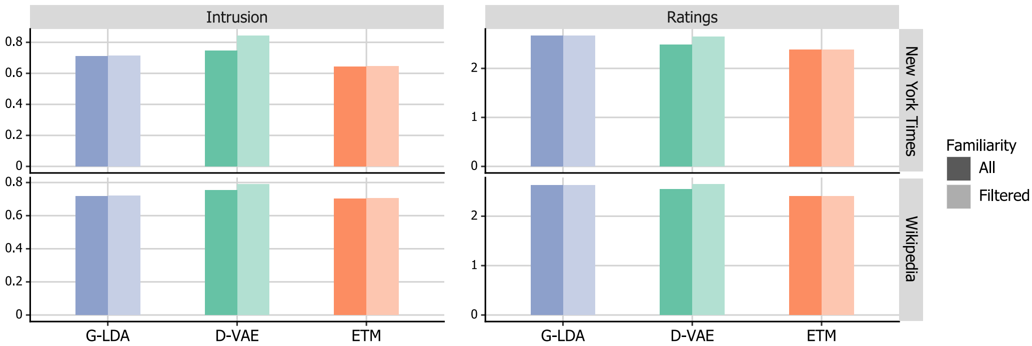


Figure 3: Mean human evaluation on the ratings and word intrusion tasks, after filtering out respondents who reported a lack of familiarity with the topic words. When filtering, D-VAE scores improve, highlighting its tendency to produce esoteric topics.

the predicted scores despite *equivalent* human scores (after correcting for the probability of type I errors, $\alpha = 0.05$).[20] Even the best-performing automated metrics predict significant differences absent a meaningful human effect roughly one-fifth of the time (Table 4).

These results suggest that automated metrics alone may be inadequate for model comparison.

6.3 Explaining the discrepancy

One reason for the discrepancy between human judgments and automated metrics is that metrics favor more esoteric topics. Specifically, there is a significant negative correlation between a topic's NPMI or C_v and the share of respondents reporting familiarity with topic words (Pearson's $\rho = -0.29$). And while D-VAE achieves the highest automated metric scores of the three models, it produces topics with the fewest familiar words: respondents report familiarity with terms over 90% of the time on both tasks for G-LDA and ETM, but they do so only 70% of the time for D-VAE. This difference suggests that the topics selected by D-VAE are narrower in scope than those of the other models. As shown in Figure 3, removing item annotations where respondents indicate unfamiliarity causes both accuracy in the word intrusion task and the ratio of "Very related" terms in the ratings task for D-VAE to increase substantially.

Qualitatively, this result is apparent when examining topics with a high NPMI but low humans ratings. In Table 5, the top rows consists of financial terms that frequently appear *together* in NYT articles,

[20]Details on testing equivalence are in Section A.5.1.

Data	Model	Topic	NPMI	Rat.	Int.
NYT	D-VAE	inc 6mo earns otc rev qtr 9mo nyse outst dec	0.56	1.60	0.77
WIKI	D-VAE	waterline conning turrets boilers amidships aft knots armament guns mounts	0.33	1.93	0.65
NYT	G-LDA	bedroom room bath taxes year market listed kitchen broker weeks	0.30	2.00	0.23
NYT	D-VAE	condolences mourns mourn board_of_directors heartfelt deepest esteemed	0.38	2.60	0.23
NYT	D-VAE	shareholders earnings federated mci shares takeover new_york_stock_exchange	0.18	3.00	0.81
WIKI	D-VAE	continental_army expedition militia frigate musket frigates muskets skirmish	0.11	3.00	0.69
NYT	D-VAE	medicaid medicare hospitals welfare uninsured patients	0.13	2.80	0.96
NYT	G-LDA	city mayor state new_york new_york_city officials county yesterday governor	0.09	2.53	1.00

Table 5: Topics with the largest human–NPMI discrepancies; top half are topics where NPMI is high and human preferences are low, bottom half is the reverse. NPMI favors esoteric and corpus-specific topics. **NPMI** is calculated with a 10-token sliding window over the in-domain reference corpus, **Rat.** is the average 3-point rating for a topic, and **Int.** refers to the percentage of annotators who identify the intruder word.

and the second row contains rare terms about boating—arguably both are reasonable topics for their respective corpora. We can also see instances where words are qualitatively very related (bottom half of table), but that NPMI fails to score high—perhaps because these words, while related, may not frequently appear together within a ten-word sliding window (Equation 1).

Even for familiar words, some topics may be sensible in the context of the specific corpus, despite their component words lacking an immediately obvious semantic relationship. For example, the topic words in the third and fourth rows appear somewhat unrelated (e.g., "taxes" and "bedroom" in the third row), but they are in fact characteristic of common document types in the New York *Times*: real estate listings and obituaries. Topics like these render the word intrusion task more difficult: only 23% of crowdworkers identified the intruder for both topics.

Furthermore, using term familiarity as a proxy for domain expertise does not address the key problems with topic model evaluation: even after filtering out respondents who are not familiar with topic terms, automated metrics still overstate model differences (Appendix A.7). The problems with topic model evaluation may therefore extend to our choice of *human* evaluations as well.

7 So… is Automated Topic Modeling Evaluation Broken?

To the extent that our experimentation accurately represents current practice, our results do suggest that topic model evaluation—both automated and human—is overdue for a careful reconsideration. In this, we agree with Doogan and Buntine (2021), who write that "coherence measures designed for older models [. . .] may be incompatible with newer models" and instead argue for evaluation paradigms centered on corpus exploration and labeling. The right starting point for this reassessment is the recognition that both automated and human evaluations are abstractions of a real-world problem. The familiar use of precision-at-10 in information retrieval, for example, corresponds to a user who is only willing to consider the top ten retrieved documents. In future work, we intend to explore automated metrics that better approximate the preferences of real-world topic model users.

One primary use of topic models is in computer-assisted content analysis. In that context, rather than taking a methods-driven approach to evaluation, it would make sense to take a needs-driven approach.[21] Generic evaluation of topic models using domain-general corpora like NYT needs to be revisited, since there is no such thing as a "generic" corpus for content analysis, nor a generic analyst. *Content analysis* can be formulated in a broad way, as Krippendorff (2004) has shown, but its actual application is always in a domain, by people familiar with that domain. This fact stands in tension with the desirable practicalities of general corpora and crowdworker annotation, and the field will need to address this tension. We have identified "coherence" as calling out a latent concept in the mind of a reader. It follows that we must think about who the relevant human readers are and the conceptual spaces that matter to them.

[21] These needs also have a computational component: neural models usually have longer runtimes even when accelerated with GPUs, whereas many practitioners work in local, CPU-only, environments. See Appendix A.3 for additional details on runtimes.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants 2031736, 2008761, 1822494, ARLIS, and by an Amazon Research Award. We thank Sweta Agrawal for her suggestion to conduct a meta-analysis. We owe much appreciation to Dallas Card for his keen advice on power analyses. Thanks to Frank Fineis for help on several statistical questions, as well as Shuo Chen for his suggestions regarding the false discovery rate calculations. Finally, we thank Caitie Doogan for her helpful comments on the clarity of argumentation, as well as our anonymous reviewers.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics.

Daniel Allington, Sarah Brouillette, and David Golumbia. 2016. Neoliberal tools (and archives): A political history of digital humanities. In *LA Review of Books*.

Areej Alokaili, Nikolaos Aletras, and Mark Stevenson. 2019. Re-ranking words to improve interpretability of automatically generated topics. In *International Conference on Computational Semantics*. Association for Computational Linguistics.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2017. An automatic approach for document-level topic model evaluation. In *Conference on Computational Natural Language Learning*, 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 Association for Computational Linguistics*, Online. Association for Computational Linguistics.

David M. Blei, Andrew Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. *Applications of Topic Models*. NOW Publishers.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. In *Frontiers in Psychology*.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model. In *Journal of Machine Learning Research*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

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

Jason Chuang, John D. Wilkerson, Rebecca Weiss, Dustin Tingley, Brandon M. Stewart, Margaret E. Roberts, Forough Poursabzi-Sangdeh, Justin Grimmer, Leah Findlater, Jordan Boyd-Graber, and Jeff Heer. 2014. Computer-assisted content analysis : Topic models for exploring multiple subjective interpretations. In *Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning*.

Matthew J Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. In *Political Analysis*. Cambridge University Press.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? Re-evaluating semantic interpretability measures. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the International Conference of Machine Learning*. Omnipress.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboard design. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alan H. Feiveson. 2002. Power by simulation. In *The Stata Journal*.

Jiachun Feng, Zusheng Zhang, Cheng Ding, Yanghui Rao, and Haoran Xie. 2020. Context reinforced neural topic modeling over short texts. In *ArXiv*.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences*. National Academy of Sciences.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. In *Political Analysis*. Cambridge University Press.

Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. 2019. Neural topic model with reinforcement learning. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pankaj Gupta, Yatin Chaudhary, F. Buettner, and Hinrich Schütze. 2019a. textTOvec: Deep contextualized neural autoregressive models of language with distributed compositional prior. In *Proceedings of the International Conference on Learning Representations*.

Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schütze. 2019b. Document informed neural autoregressive topic models with distributional prior. In *Association for the Advancement of Artificial Intelligence*. AAAI Press.

Ruifang He, Xuefei Zhang, Di Jin, Longbiao Wang, Jianwu Dang, and Xiangang Li. 2018. Interaction-aware topic model for microblog conversations through network embedding and user attention. In *International Conference on Computational Linguistics*. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving Neural Topic Models using Knowledge Distillation. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xuemeng Hu, Rui Wang, Deyu Zhou, and Yuxuan Xiong. 2020. Neural topic modeling with cycle-consistent adversarial training. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Karoliina Isoaho, Daria Gritsenko, and Eetu Mäkelä. 2021. Topic modeling and text analysis for qualitative policy research. In *Policy Studies Journal*.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-Structured Neural Topic Model. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Martin Jankowiak and Fritz Obermeyer. 2018. Pathwise derivatives beyond the reparameterization trick. In *Proceedings of the International Conference of Machine Learning*. PMLR.

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. 2020. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514.

Namkyu Jung and Hyeong In Choi. 2017. Continuous semantic topic embedding model using variational autoencoder. In *ArXiv*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Katsiaryna Krasnashchok and Salim Jouili. 2018. Improving topic quality by promoting named entities in topic modeling. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: an Introduction to its Methodology*. SAGE.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Lihui Lin, Hongyu Jiang, and Yanghui Rao. 2020. Copula guided neural topic modelling for short texts. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Tianyi Lin, Zhiyue Hu, and Xin Guo. 2019. Sparsemax and relaxed wasserstein for topic sparsity. In *International Conference on Web Search and Data Mining (WSDM)*. ACM.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. In *Queue*. ACM.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. In *SpringerPlus*.

Luyang Liu, Heyan Huang, Yang Gao, Yongfeng Zhang, and Xiaochi Wei. 2019. Neural variational correlated topic modeling. In *Proceedings of the World Wide Web Conference*. ACM.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Emily Hales, and Kevin Seppi. 2019. Cross-referencing using fine-grained topic modeling. In *Proceedings of the Association for Computational Linguistics*, Minneapolis, Minnesota. Association for Computational Linguistics.

Henry Berthold Mann and Donald Ransom Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. In *The Annals of Mathematical Statistics*. Institute of Mathematical Statistics.

Stephen Marche. 2012. Literature is not data: Against digital humanities. In *LA Review of Books*.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit.

Elijah Meeks and Scott B Weingart. 2012. The digital humanities contribution to topic modeling. In *Journal of Digital Humanities*.

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

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the International Conference of Machine Learning*. PMLR.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the International Conference of Machine Learning*. PMLR.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

John W. Mohr and Petko Bogdanov. 2013. Introduction—topic models: What they are and why they matter. In *Poetics*.

Fred Morstatter and Huan Liu. 2018. In search of coherence and consensus: Measuring the interpretability of statistical topics. *Journal of Machine Learning Research*.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*.

Xuefei Ning, Y. Zheng, Zhuxi Jiang, Y. Wang, H. Yang, and J. Huang. 2020. Nonparametric topic modeling with neural inference. In *Neurocomputing*.

Madhur Panwar, Shashank Shailabh, Milan Aggarwal, and Balaji Krishnamurthy. 2020. TAN-NTM: Topic attention networks for neural topic modeling. In *Proceedings of the Association for Computational Linguistics*.

Min Peng, Qianqian Xie, Yanchun Zhang, Hua Wang, Xiuzhen Zhang, Jimin Huang, and Gang Tian. 2018. Neural sparse topical coding. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the Language Resources and Evaluation Conference*. ELRA.

Mehdi Rezaee and Francis Ferraro. 2020. A discrete variational recurrent topic model without the reparametrization trick. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *International Conference on Web Search and Data Mining (WSDM)*. ACM.

Evan Sandhaus. 2008. The New York Times annotated corpus. In *Linguistic Data Consortium*.

Benjamin M Schmidt. 2012. Words alone: Dismantling topic models in the humanities. In *Journal of Digital Humanities*.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*.

Donald J Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. In *Journal of pharmacokinetics and biopharmaceutics*. Springer.

Denys Silveira, André Carvalho, Marco Cristo, and Marie-Francine Moens. 2018. Topic Modeling using Variational Auto-Encoders with Gumbel-Softmax and Logistic-Normal Mixture Distributions. In *International Joint Conference on Neural Networks (IJCNN)*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the International Conference on Learning Representations*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Marilyn Strathern. 1997. Improving Ratings: Audit in the british university system. In *European Review*. Cambridge University Press.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple! In *Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Laure Thompson and D. Mimno. 2020. Topic modeling with contextualized word representation clusters. In *ArXiv*.

Runzhi Tian, Yongyi Mao, and Richong Zhang. 2020. Learning VAE-LDA models with rounded reparameterization trick. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

William E Underwood. 2017. A genealogy of distant reading. In *Digital Humanities Quarterly*. Alliance of Digital Humanities Organisations.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020a. Neural topic modeling with bidirectional adversarial training. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Rui Wang, Deyu Zhou, and Yulan He. 2020b. ATM: Adversarial-neural topic model. In *Proceedings of the Association for Computational Linguistics*.

Stefan Wellek. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC.

Jiemin Wu, Yanghui Rao, Zusheng Zhang, Haoran Xie, Qing Li, Fu Lee Wang, and Ziye Chen. 2020a. Neural mixed counting models for dispersed topic discovery. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020b. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph attention topic modeling network. In *Proceedings of the World Wide Web Conference*. ACM.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. 2018. WHAI: weibull hybrid autoencoding inference for deep topic modeling. In *Proceedings of the International Conference on Learning Representations*.

He Zhao, Lan Du, Wray L. Buntine, and Mingyuan Zhou. 2018. Dirichlet belief networks for topic structure learning. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2021a. Neural topic model via optimal transport. In *Proceedings of the International Conference on Learning Representations*.

He Zhao, Dinh Q. Phung, Viet Huynh, Y. Jin, Lan Du, and W. Buntine. 2021b. Topic modelling meets deep neural networks: A survey. In *ArXiv*.

Deyu Zhou, Xuemeng Hu, and Rui Wang. 2020. Neural topic modeling by incorporating document relationship graph. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.

A Appendix

A.1 List of Neural Topic Modeling Works used in our Meta-Analysis

In Table 6, we report the forty publications used in our meta-analysis (Section 3), which are sourced from a survey of neural topic models (Zhao et al., 2021b).

A.2 Preprocessing Details

Our steps are delineated in our implementation,[22] but we list our choices here for easy reference. Corpus statistics are in Table 7. We use the default `en-core-web-sm` spaCy model (Honnibal et al., 2020), version 3.0.5, throughout.

Document processing

- We do not process documents with fewer than 25 whitespace-separated tokens.
- Following processing (e.g., stopword removal), we remove documents with fewer than five tokens.
- We truncate documents to 5,000 whitespace-separated tokens for NYT and to 19,000 for WIKI (in both cases affecting less than 0.15% of documents).

Vocabulary creation

- We tokenize using spaCy.
- We lowercase terms.
- We *do not* lemmatize.
- We detect noun entities with spaCy, keeping only the `ORG`, `PERSON`, `FACILITY`, `GPE`, and `LOC` types, joining constituent tokens with an underscore (e.g, "New York City" $\rightarrow$ `new_york_city`).

Vocabulary filtering

- The vocabulary is created from the training data. The reference texts used in coherence calculations are processed identically and use the same vocabulary.
- We filter out stopwords using the default spaCy English stopword list.[23] Stopwords are retained if they are contained within detected noun entities (e.g., "The United States of America" $\rightarrow$ `united_states_of_america`).
- We filter out tokens with two or fewer characters.
- We retain only tokens that are matched by the regular expression `^[\w-]*[a-zA-Z][\w-]*$`
- We remove tokens that appear in more than 90% of documents.
- We remove tokens that appear in fewer than $2(0.02|D|)^{1/\log 10}$ documents, where $|D|$ is the corpus size.[24]

A.3 Training Details

Expanding Section 4.2, we detail the hyperparameter tuning for each of our three topic models, along with other pertinent details about runtimes and compute resources. Scripts used to run the models with all the various hyperparameter configurations are released as part of our code; this section is also included for reference.

Our general strategy, especially with the neural models, is to select different values around the reported optimal settings in original papers. For all three models, we try two different values for the number of training iterations (G-LDA) or epochs (D-VAE, ETM).

[22] `github.com/ahoho/topics`

[23] `github.com/explosion/spaCy/blob/v3.0.5/spacy/lang/en/stop_words.py`

[24] Standard rules-of-thumb for vocabulary pruning, like removing terms that appear in fewer than 0.5% of documents (Denny and Spirling, 2018), ignore the power-law distribution of word frequency Zipf (1949), and hence do not scale to large corpora. To keep vocabulary sizes roughly consistent across datasets, we set the minimum document-frequency for terms as a (power) function of the total corpus size. This has the intuitive appeal of increasing proportional to the order of magnitude of the number of total documents, starting at a minimum document-frequency of 2 for a 50-document corpus and reaching about 110 for a corpus of 500,000.

Source	Human Evals?	Perplexity	Coherence	Implementation Specified	Ref. Corpus Specified ?	Consistent Preproc?	Hparam search?	>1 run / err. bars?	LDA Implementation?	Baseline w/in 2 yr?
Bianchi et al. (2021)	No	No	NPMI, Embed-sim	None	Internal, External-GoogleNews	Yes	No	Yes	Variational	No
Zhao et al. (2021a)	No	No	NPMI	Palmetto	No	Unclear	No	Yes	N/A	Yes
Feng et al. (2020)	No	Yes	NPMI	None	No	Yes	No	No	N/A	Yes
Hoyle et al. (2020)	No	No	NPMI	In paper	External NYT, Internal	No	Yes	Yes	N/A	Yes
Hu et al. (2020)	No	No	C_p, C_a, NPMI	Palmetto	External WIKI	No	Likely no	No	Sampling	Yes
Isonuma et al. (2020)	No	Yes	NPMI	None	No, likely external	Unclear	No	No	Sampling	No
Joo et al. (2020)	No	Yes	NPMI	None	No, likely internal	No	Likely yes	Yes	N/A	Yes
Lin et al. (2020)	No	Yes	NPMI	None	No, likely internal	Unclear	Yes	Yes	N/A	Yes
Ning et al. (2020)	No	Yes	NPMI	Lau github	No	Yes	Likely no	Yes	Variational	No
Panwar et al. (2020)	No	No	NPMI	Lau github	No	Yes	Likely no	No	Sampling	Yes
Rezaee and Ferraro (2020)	No	No	N/A	N/A	N/A	Yes	Likely no	Yes	Variational	No
Thompson and Mimno (2020)	No	No	Coherence, PMI	In paper	External NYT	No	No	Yes	Sampling	No
Tian et al. (2020)	No	Yes	NPMI	None	No	No	Yes	No	Variational	Yes
Wang et al. (2020a)	No	No	C_p, C_a, NPMI, UCI	Palmetto	No	No	No	No	Sampling	Yes
Wu et al. (2020a)	No	Yes	NPMI	None	No	No	Yes	No	N/A	Yes
Wu et al. (2020b)	No	No	C_v	Palmetto	No	Yes	No	No	Unspecified	Yes
Yang et al. (2020)	No	Yes	Coherence	In paper	No, likely internal	Yes	No	No	Unspecified	No
Zhou et al. (2020)	No	No	NPMI, C_p	Palmetto	External WIKI	No	Likely no	No	Unspecified	Yes
Burkhardt and Kramer (2019)	No	Yes	NPMI	None	No, likely internal	Unclear	Yes	No	Variational	Yes
Dieng et al. (2020)	No	Yes	Coherence	In paper	No, likely internal	Yes	No	No	Unspecified	No
Gui et al. (2019)	No	No	C_v	None	External WIKI	Yes	Likely no	No	Unspecified	Yes
Gupta et al. (2019b)	No	Yes	C_v	Gensim	No, likely internal	Unclear	Likely no	No	N/A	No
Gupta et al. (2019a)	No	Yes	C_v	Gensim	No, likely internal	Unclear	Likely no	No	Sampling	Yes
Lin et al. (2019)	No	Yes	PMI	In paper	No, likely external	Unclear	No	No	Variational	Yes
Liu et al. (2019)	No	Yes	NPMI	Lau github	No, likely internal	Yes	No	No	Variational	Yes
Nan et al. (2019)	No	No	NPMI	None	No	No	No	No	Sampling	Yes
Wang et al. (2020b)	No	No	C_p, C_a, UCI, NPMI, UMASS	Palmetto	No	No	No	No	Unspecified	Yes
Card et al. (2018)	No	Yes	NPMI	In paper	External-gigaword	Yes	Likely yes	No	Sampling	Yes
Ding et al. (2018)	No	Yes	NPMI	Lau github	No, likely external	No	Likely no	No	Sampling	Yes
He et al. (2018)	No	No	Coherence	None	No, likely internal	Yes	No	No	N/A	Yes
Peng et al. (2018)	No	Yes	N/A	N/A	N/A	Yes	Likely no	No	Variational	Yes
Silveira et al. (2018)	No	Yes	NPMI	Lau github	Internal	Yes	No	Yes	N/A	Yes
Zhang et al. (2018)	No	Yes	N/A	N/A	N/A	Unclear	Likely no	No	N/A	Yes
Zhao et al. (2018)	No	Yes	NPMI	Palmetto	External WIKI	Unclear	No	Yes	N/A	Yes
Zhu et al. (2018)	No	No	Coherence	None	No, likely internal	Yes	Likely no	No	Variational	Yes
Jung and Choi (2017)	No	Yes	NPMI, PMI, UMASS	None	No	Yes	No	No	Sampling	Yes
Miao et al. (2017)	No	Yes	NPMI	In paper	No	No	Likely no	No	Variational	Yes
Srivastava and Sutton (2017)	No	Yes	NPMI	None	No	Yes	No	No	Sampling	Yes
Miao et al. (2016)	No	Yes	N/A	N/A	N/A	Yes	Likely no	No	Unspecified	Yes
Nguyen et al. (2015)	No	No	NPMI	Lau github	External WIKI	Yes	No	Yes	Sampling	No

Table 6: Papers used in meta-analysis, Section 3

	WIKI	NYT
Domain	Encyclopedia	News
Number of Docs.		
Training	28.5k	273.1k
Reference	4.62M	1.82M
Mean Tokens / Doc.	1291	281
Vocab. Size	39.7k	34.6k

Table 7: Corpus statistics. Datasets vary in domain, average document length, and total number of documents. WIKI is from Merity et al. (2017) and NYT is from Sandhaus (2008).

G-LDA We use gensim (Řehůřek and Sojka, 2010) as a Python wrapper for running Mallet. In Table 8a, we tune hyperparameters α (topic density parameter) and β (word density parameter) which can be thought of as "smoothing parameters" that reserve some probability for the topics (words) unassigned to a document (topic) thus far. Mallet internally optimizes hyperparameters, and the Optimization Interval controls the frequency of hyperparameter updates, measured in training steps.

D-VAE Our reimplementation of Dirichlet-VAE (Burkhardt and Kramer, 2019) largely uses the same hyperparameters as reported in that work. As shown in Table 8b, we vary the prior for the Dirichlet distribution (α), the learning rate (η), the L_1-regularization constant for the topic-word distribution ($\beta_{reg.}$, not in the original model but inspired by Eisenstein et al., 2011), the number of epochs to anneal the use of batch normalization in the decoder (γ_{BN}, comes from Card et al., 2018), and the number of epochs to anneal the KL-divergence term in the loss (γ_{KL}) (it needs to be introduced slowly in the loss function due to the component collapse problem in VAEs (Bowman et al., 2016)).

ETM Following Dieng et al. (2020), we learn skip-gram embeddings on the training corpus using the provided script, which relies on gensim. As shown in Table 8c, we vary the learning rate (η), the L_2 regularization constant for the Adam (Kingma and Ba, 2015) optimizer (W_{decay}), and a boolean indicator of whether to anneal the learning rate (γ_η). If annealing is allowed, the learning rate gets divided by 4.0 if the loss on the validation set does not improve for more than 10 epochs, per the default settings of the model (preliminary experiments showed that annealing did not attain higher NPMI).

The runtimes for each of the models on each dataset are in Table 9. We used AWS ParallelCluster to provide a cloud-computing computing cluster. Neural models ran on NVIDIA T4 GPUs using `g4dn.xlarge` instances with 16 GiB memory and 4 CPUs.[25] G-LDA (Mallet) ran on CPU only, with `m5d.2xlarge` instances (with 32 GiB memory, 8 CPUs).[26]

A.4 Instructions for Crowdworkers

Recruiting participants on Prolific.co for a Qualtrics survey produced results with higher inter-worker agreement than Mechanical Turk, based on a pilot test. Using the Prolific.co platform, we recruited respondents that met the criteria of living in the United States and listing fluency in English. Each respondent was paid through Prolific upon completion of the survey, at a rate corresponding to $15 an hour. The total amount spent on conducting all the surveys, including our pilot test, was $2084.91. We used automated scripts to generate separate Qualtrics surveys for each task that contained the topics for evaluation, available in our released code. Each respondent was shown 25% of the questions in each survey; the question selection and answer display order was chosen randomly via the survey configuration on Qualtrics. Figures 1 and 4 depict our word intrusion and ratings tasks, respectively. Crowdworkers receive instructions explaining the task (Figure 5) and the dataset (Figure 6).

[25] `https://aws.amazon.com/hpc/parallelcluster/`

[26] See `https://aws.amazon.com/ec2/instance-types/` for further details.

Model: **G-LDA**			
α	β	Optim. Interval	#Steps
$\{0.01, 0.05, 0.1, 0.25^{\dagger}, 1.0^{*}, 5.0\}$	$\{0.01, 0.05^{*}, 0.1^{\dagger}\}$	$\{0, 10^{\dagger}, 100, 500^{*}\}$	$\{1000^{\dagger}, 2000^{*}\}$

(a) Hyperparameter ranges for G-LDA. α is the topic density parameter. β is the word density parameter. Optim. Interval sets the number of iterations between Mallet's own internal hyperparameter updates. #Steps are training iterations.

Model: **D-VAE**					
α	η	$\beta_{reg.}$	γ_{BN}	γ_{KL}	#Steps
$\{0.001, 0.01^{*\dagger}, 0.1\}$	$\{0.001, 0.01^{*\dagger}\}$	$\{0.0^{*}, 0.01, 0.1^{\dagger}, 1.0\}$	$\{0, 1^{*}, 100, 200^{\dagger}\}$	$\{100^{*}, 200^{\dagger}\}$	$\{200, 500^{*\dagger}\}$

(b) Hyperparameter ranges for D-VAE. α is the Dirichlet prior. η is the learning rate. $\beta_{reg.}$ is the L_1-regularization of the topic-word distribution. γ_{BN} and γ_{KL} are the number of epochs to anneal the batch normalization constant and KL divergence term in the loss, respectively. #Steps are training epochs.

Model: **ETM**			
η	W_{decay}	γ_{η}	#Steps
$\{0.001^{*}, 0.002, 0.01, 0.02^{*\dagger}\}$	$\{1.2e^{-5*}, 1.2e^{-6\dagger}, 1.2e^{-7}\}$	$\{0^{*\dagger}, 1\}$	$\{500, 1000^{*\dagger}\}$

(c) Hyperparameter ranges for ETM. η is the learning rate. W_{decay} is the L_2 regularization constant. γ_{η} is an indicator of whether learning rate is annealed. #Steps are training epochs.

Table 8: Hyperparameter settings for G-LDA, D-VAE, and ETM. $*$: Best setting for WIKI, $\dagger$: best setting for NYT; based on NPMI estimated with a 10-token sliding window over the reference corpus.

	WIKI	**NYT**
G-LDA	$\sim$ 2 minutes	$\sim$ 9 minutes
D-VAE	$\sim$ 45 minutes	$\sim$ 330 minutes
ETM	$\sim$ 260 minutes	$\sim$ 1300 minutes

Table 9: Runtimes for the three topic models on each of the two datasets. G-LDA requires CPUs only while the neural models use a single GPU. Compute resources detailed at the end of Section A.3.

Please rate how related the following terms are to each other and how familiar you are with the terms

concerto, balanchine, mozart, orchestra, brahms, beethoven, recital, choreographers, schubert, composers

Rating
- ◯ Very related
- ◯ Somewhat related
- ◯ Not very related

Answer Confidence
- ◯ I am familiar with most of these terms.
- ◯ I am ***not*** familiar with most of these terms, but I ***can*** answer confidently.
- ◯ I am ***not*** familiar with most of these terms, and so I ***cannot*** answer confidently.

Figure 4: Ratings task presented to crowdworkers.

A.5 Power Analysis for Human Evaluation Tasks

To select the number of crowdworkers, we conduct a power analysis with simulated data (Feiveson, 2002) by formulating a generative model of annotations (implementation included in released code). Card et al. (2020) find that many NLP experiments, including those relying on human evaluation, are insufficiently powered to detect model differences at reported levels.

Word Intrusion. Topic k has a true latent binary label $z_k \sim \text{Bern}(0.5)$ ("coherent" or "incoherent") which indexes a parameter $p_{z_k} \in [0, 1]$. Annotator i samples an answer to the intruder task $x_{ik} \sim \text{Bern}(p_{z_k})$. We therefore run a simulation of annotator data for two different models: MODEL A, This survey asks you to look at lists of words produced by an automatic computer program. For each list, you'll be answering the question: "Which word doesn't belong?"

- You will be shown ten sets of six words.
- For each set, click the word whose meaning or usage is most unlike that of the other words.
- If you feel that multiple words do not belong, choose the one that you feel is most out of place.
- Do not base your decisions on how the word is pronounced or written or its grammatical function. For example, if you saw {apple, apricot, anvils, peach}, you would not choose "peach" because it doesn't start with "a", you would not choose "apricot" because it isn't five letters long, and you would not choose "apple" because it ends with a vowel. Ideally, you would choose "anvils" because it is not a fruit.

Here are some examples:

"baby", "crib", "diaper", "beer", "pacifier", "cry"

In this example the word 'beer' is the least related. All of the other words are closely related to each other, and related to infants.

Here is another, harder, example:

"Hard Drive", "motherboard", "video card", "processor", "RAM", "USB key"

While all of these terms are related to a computer, all but one of them are components inside of a computer. The best choice is therefore 'USB key'.

You may not always know all the words and that's okay.

This study should take approximately 10-15 minutes to complete. Your response will be completely anonymous.

(a)

This survey asks you to evaluate lists of words produced by an automatic method.

The computer model we are testing seeks to identify groups of words that are highly related to each other. You will be asked to select how related groups of words are on a 3-point scale.

The rating options are: Not Very Related, Somewhat Related, Very Related.

A helpful question to ask yourself is: "what is this group of words about?" If you can answer easily, then the words are probably related. Here is some guidance on how to apply these ratings and some examples.

Very Related - Most of the words are clearly related to each other, and it would be easy to describe how they are related.

Example: "dog", "cat", "hamster", "rabbit", "snake" (An obvious way to describe the relationship here would be 'Pets')

Example: "brushwork", "canvases", "expressionism", "cubism", "modernism", "curators", "abstract_expressionism", "national_gallery_of_art", "museum", "fossils" (An obvious way to describe this would be "art", even though one or two of the words are not as clearly related to that.)

Somewhat Related - The words are loosely related to each other, but there may be a few ambiguous, generic, or unrelated words

Example: "computer", "video", "new", "plug", "screen", "model" (In this example, some of the words are generic, and seem more closely related than others)

Example: "dog", "ball", "pet", "receipt", "pen" (In this example, some of the words seem closely related, but not all of them)

Not Very Related - The words do not share any obvious relationship to each other. It would be difficult to describe how the words are related to each other.

Example: "dog", "apple", "pencil", "earth", "computer"

This study should take approximately 10-15 minutes to complete. Your response will be completely anonymous.

(b)

Figure 5: Instructions for (a) word intrusion and (b) ratings

In this survey, the word lists are based on a computer analysis of The New York *Times*.

The New York *Times* is an American newspaper featuring articles from 1987 to 2007. Sections from a typical paper include International, National, New York Regional, Business, Technology, and Sports news; features on topics such as Dining, Movies, Travel, and Fashion; there are also obituaries and opinion pieces.

(a)

In this survey, the word lists are based on a computer analysis of Wikipedia.

Wikipedia is an online encyclopedia covering a huge range of topics. Articles can include biographies ("George Washington"), scientific phenomena ("Solar Eclipse"), art pieces ("La Danse"), music ("Amazing Grace"), transportation ("U.S. Route 131"), sports ("1952 winter olympics"), historical events or periods ("Tang Dynasty"), media and pop culture ("The Simpsons Movie"), places ("Yosemite National Park"), plants and animals ("koala"), and warfare ("USS Nevada (BB-36)"), among others.

(b)

Figure 6: Descriptions for (a) NYTimes and (b) Wikipedia.

which has a sample of $K = 50$ binary topic labels, $\boldsymbol{z}^{(A)}$; and MODEL B, with r fewer "coherent" topics than A, $\sum_k z_k^{(B)} = \sum_k z_k^{(A)} - r$. After collecting pseudo-scores $\boldsymbol{x}^{(A)}$ and $\boldsymbol{x}^{(B)}$ for M annotators, we run a one-tailed proportion test on the respective sums. The power is the proportion of significant tests over the total number of simulations N (i.e., tests there where A is correctly determined to have higher scores than B). We set $p_0 = 1/6$ (chance of guessing), $p_1 = 0.85$ (roughly estimated with data from Chang et al., 2009).

Ratings. Rating scores on a 3-point scale are generated analogously, in a generalization of the above binary case. Assume that topics have true labels $\boldsymbol{z}_k \sim \text{Cat}(1/3, 1/3, 1/3)$. Annotator scores are noisy, so true labels are corrupted according to probabilities $p_{z_k} \in \Delta^2$. Here, MODEL A has a sample of $K = 50$ ratings on a 3-point scale. MODEL B has r fewer 3-ratings ("very related") and r greater 1-ratings ("not related") than A (the 2-ratings stay constant). After simulating scores for M annotators for both "models," we run a one-tailed U-test (Mann and Whitney, 1947). Again, the power is the share of significant tests over all simulations N. Probabilities are $p_1 = [3/4, 1/4, 0]; p_2 = [1/4, 2/4, 1/4]; p_3 = [0, 1/4, 3/4]$, designed to roughly approximate empirical data—if we sample scores according to them and compute inter-"annotator" agreement, the one-versus-rest Spearman correlation is $\rho \approx 0.7$, or the same as the most-correlated dataset (NYT) in Aletras and Stevenson (2013) (our final data has $\rho = 0.75$).

For both settings, we set $r = 4$, the critical value $\alpha = 0.05$, and the desired power $1 - \beta = 0.9$. This analysis suggests fifteen annotators per topic for the ratings task and twenty-five for intrusion.

	Ref. Corpus → Train Corpus ↓	NPMI (10-token window) NYT	WIKI	Train	Val	C_v (110-token window) NYT	WIKI	Train	Val
Intrusion	NYT	2.42	**4.16**	2.11	1.97	2.50	<u>3.27</u>	2.55	2.40
	WIKI	<u>4.11</u>	<u>5.08</u>	**5.45**	0.87	2.23	2.79	2.74	0.70
	Concatenated	2.82	**4.56**	3.18	0.78	2.30	3.05	2.64	0.87
Rating	NYT	1.92	<u>2.08</u>	1.77	1.85	<u>2.55</u>	<u>2.51</u>	**2.68**	<u>2.59</u>
	WIKI	2.97	<u>4.10</u>	**4.29**	1.45	2.01	2.82	2.86	0.80
	Concatenated	<u>2.20</u>	**2.75**	<u>2.52</u>	1.17	<u>2.27</u>	<u>2.60</u>	<u>2.74</u>	1.07

Table 10: Logistic (intrusion) and ordinal probit (ratings) regression coefficients of automated metrics on human annotations. <u>Underlined</u> values have overlapping 95% confidence intervals with that of the **largest** value in each row.

A.5.1 Power analysis for equivalence

To estimate the false discovery (omission) rates in Table 4, we need to determine when differences between human (automated) scores are not meaningful. Since human effects in the opposite direction of automated metrics also imply a false discovery, we conduct a test of non-inferiority; this is the same as using a large negative lower bound in the two-one-sided tests procedure for equivalence (Schuirmann, 1987; Wellek, 2010).

To determine the non-inferiority threshold—the bound ϵ below which we consider two sets of scores to be equivalent—we also conduct a power analysis, per the previous section. In this case, the simulation assumes *no* difference between the "true" labels of the model outputs, $\boldsymbol{z}^{(A)} = \boldsymbol{z}^{(B)}$. We estimate one-sided tests for each sample of human scores, with the null $H_0 : \mu_1^{(B)} - \mu^{(A)} > \epsilon$ for some bound ϵ. We minimize ϵ while maintaining $\beta > 0.9$. This process produces $\epsilon = 0.05$ for the word intrusion task and $\epsilon = 0.11$ for the ratings task (roughly equivalent to a difference of 2.5 "incoherent" topics for both tasks, respectively).

For the automated scores, we generate two sets of scores $x_k \sim \mathcal{N}(0, \sigma^2)$; $\sigma^2 \sim \text{Gamma}(\alpha, \beta)$ for $k = 1 \dots K$ at each iteration, then conduct a t-test between each set. α and β are selected such that the Gamma distribution approximately matches the empirical distribution of automated score variances. This leads to $\epsilon = 0.05$ for NPMI scores and $\epsilon = 0.06$ for the C_v scores.

A.6 Regression Results

Prior work (e.g., Röder et al., 2015) relates averaged human ratings to automated metrics using either Pearson or Spearman correlations. As an alternative that takes into account both the variation in human judgments as well as their numerical type, we estimate logistic and ordered probit regressions on the ratings and intrusion annotations, respectively. In Table 10, we report the estimated coefficients for each metric, finding that—on the whole—using the WIKI reference performs best, although the large estimated confidence intervals mitigate the strength of this conclusion.

A.7 Filtering on Term Familiarity

Several topics, particularly those produced by D-VAE, contain terms that are not well-known to annotators (6.1). When a respondent is unfamiliar with a topic's words, their ratings for that topic may not accurately reflect its true coherence. For example, a mycologist may find the words in the fifth column of Table 1 highly related, whereas someone unfamiliar with fungi-related jargon may rate it poorly—indeed, the mean rating for this topic is 2.1 for those unfamiliar with terms and 2.6 for those who are familiar.

Since automated metrics do not take into account a term's familiarity to humans, we posit that automated metrics should be more predictive of human judgments among respondents who are familiar with topic terms. To test this hypothesis, we re-evaluate the relationships between automated metrics and human judgments *after removing* respondents who state they are not familiar with a topic's terms (Table 11). On the whole, results are much clearer than above; NPMI estimated using WIKI reference counts is strongly correlated across tasks and datasets. The false discovery rate is

	Ref. Corpus → Train Corpus ↓	NPMI (10-token window)				C_v (110-token window)			
		NYT	WIKI	Train	Val	NYT	WIKI	Train	Val
Intrusion	NYT	0.34	0.51	0.32	0.25	0.44	**0.55**	0.42	0.38
	WIKI	0.39	0.39	0.40	0.14	0.38	**0.40**	0.39	0.13
	Concatenated	0.36	0.45	0.36	0.18	0.41	**0.48**	0.41	0.26
Rating	NYT	0.45	**0.59**	0.44	0.43	0.51	0.58	0.53	0.52
	WIKI	0.45	**0.51**	0.51	0.21	0.44	0.51	0.51	0.23
	Concatenated	0.47	**0.54**	0.47	0.35	0.49	0.53	0.51	0.42

(a) Spearman correlation coefficients between mean human scores and automated metrics, compare to Table 3.

	Ref. Corpus → Train Corpus ↓	NPMI (10-token window)			C_v (110-token window)		
		NYT	WIKI	Train	NYT	WIKI	Train
Intrusion	NYT	53 / 55	46 / 50	56 / 52	41 / 34	**28** / 27	42 / 33
	WIKI	38 / 76	36 / 77	32 / 76	**29** / 37	31 / 43	33 / 39
	Concatenated	54 / 70	41 / 73	51 / 70	41 / 44	**29** / 42	41 / 44
Rating	NYT	45 / 49	39 / 53	45 / 47	18 / 27	**16** / 24	17 / 25
	WIKI	37 / 73	25 / 74	30 / 70	28 / 31	19 / 33	**18** / 27
	Concatenated	45 / 64	38 / 68	42 / 64	26 / 36	**21** / 36	27 / 33

(b) False discovery rate (1−precision, lower is better) and false omission rate of significant model differences when using automated metrics, compare to Table 4.

	Ref. Corpus → Train Corpus ↓	NPMI (10-token window)				C_v (110-token window)			
		NYT	WIKI	Train	Val	NYT	WIKI	Train	Val
Intrusion	NYT	3.71	**7.14**	3.04	2.54	3.34	4.54	3.23	2.94
	WIKI	5.87	**6.46**	6.19	0.85	3.23	3.59	3.39	0.42
	Concatenated	4.24	**6.81**	4.17	0.94	3.18	4.06	3.30	0.91
Rating	NYT	4.40	**5.87**	3.85	3.93	3.97	4.44	4.03	3.89
	WIKI	4.84	**5.95**	5.65	1.33	2.96	3.73	3.69	0.62
	Concatenated	4.49	**5.80**	4.56	1.78	3.45	3.91	3.81	1.32

(c) Logistic (intrusion) and ordinal probit (ratings) regression coefficients of automated metrics on human annotations, compare to Table 10.

Table 11: Tables 3, 4, and 10 after removing respondents who report a lack of familiarity with topic words.

lower overall, although automated metrics still misdiagnose significant results at a rate of one in six in even the best case.

These findings provide further evidence—per our discussion in Section 7—that future human evaluations of topic models ought to take into account domain expertise and information need.

A.8 Five-point Ratings Scale

Although most prior work uses three-point scales for the ratingstask (Fig. 4), for comparison we also ask annotators to label the topic topic words with a five-point scale ranging from 1 ("not at all related") to 5 ("very related", no labels are given for points 2-4). Broadly, we find that values for correlations are reduced relative to the three-point scale (Table 12). We believe examining this discrepancy is an interesting direction for future work that re-visits human evaluation of topic models.

A.9 Potential Negative Impact

Our work focuses its investigation on data from the English language alone. In this way, it further entrenches English-language primacy in NLP, and more crucially, findings may not translate directly to other languages. We caution the reader against applying claims made in this work to topic modeling

	Ref. Corpus → Train Corpus ↓	NPMI (10-token window) NYT	WIKI	Train	Val	C_v (110-token window) NYT	WIKI	Train	Val
Rating (5-pt.)	NYT	0.27	**0.37**	0.28	<u>0.33</u>	0.29	<u>0.35</u>	<u>0.33</u>	<u>0.35</u>
	WIKI	0.15	0.21	0.29	<u>0.43</u>	0.10	0.16	0.17	**0.50**
	Concatenated	0.21	0.30	0.28	0.32	0.20	0.26	0.26	**0.39**

Table 12: Spearman correlation coefficients between mean human scores for a **five-point** ratings scale (rather than three), compare to Table 3. <u>Underlined</u> values have overlapping 95% confidence intervals with that of the **largest** value in each row.

on corpora of other languages. It is even possible that one of the tasks designed to elicit human judgment (e.g., word intrusion) may not be amenable for use with other languages.

Concerning topic models more broadly, we note that others question the scholarly value of "distant reading" and the digital humanities in general (Marche, 2012; Allington et al., 2016). Do topic models encourage a passive, disengaged relationship to texts—fomenting conclusions about broad, generic trends rather than idiosyncratic specifics, leading us to miss the trees for the forest? As noted by Schmidt (2012), "topics neither can nor should be studied independently of a deep engagement in the actual word counts that build them." In this light, topic models can be viewed as an extension of the insidious neoliberal trend toward mass data harvesting that blurs differences between individuals and cultures. Researchers should take care to avoid such elisions when drawing conclusions from model outputs.

A.10 NeurIPS Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 7 and relevant places throughout the paper.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Appendix A.9
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Appendices (Sections A.3 and A.2) and explanation in main paper (Section 4).
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Section A.3 and results in main paper.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Section A.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Full instructions given to crowdworkers are included in

supplemental (Section A.4), and they are told what they are evaluating. Annotators are told that their ratings will be used to judge automatic methods.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No such information or content was present in our work.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Screenshots of examples of what the task looks like are included, as are full set of instructions (Section A.4).

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] Estimated hourly wage in Section 5. Total amount spent is included in Section A.4.