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. This paper builds on our 2009 that showed perplexity was not a good evaluation of interpretability for topic models; while the field adopted automatic topic coherence as a result of that 2009 paper, this paper argues that automatic topic coherence is not a good metric for neural topic models (even though it worked for probabilistic topic models).

Links:
- Research Talk (students): https://youtu.be/op1DkSB2VdA
- Research Talk (Jordan): https://www.youtube.com/watch?v=4KO2T0_cm2I
- Code: http://github.com/ahoho/topics

Downloaded from http://umiacs.umd.edu/~jbg/docs/2021_neurips_incoherence.pdf

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 Philip Resnik
CS, iSchool, UMIACS, LSC UMIACS, Linguistics
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...
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.

<table>
<thead>
<tr>
<th>Classical</th>
<th></th>
<th>Neural</th>
</tr>
</thead>
<tbody>
<tr>
<td>station</td>
<td>album</td>
<td>tropical</td>
</tr>
<tr>
<td>line</td>
<td>band</td>
<td>storm</td>
</tr>
<tr>
<td>bridge</td>
<td>music</td>
<td>hurricane</td>
</tr>
<tr>
<td>railway</td>
<td>song</td>
<td>cyclone</td>
</tr>
<tr>
<td>trains</td>
<td>released</td>
<td>depression</td>
</tr>
</tbody>
</table>

NPMI 0.274 0.285 0.394 0.446 0.456 0.470

Our findings challenge the validity of fully-automated evaluations as currently practiced: automated evaluation declares winners between models when the corresponding human evaluations cannot.
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, $\beta_k$, and that the documents $D$ are admixtures over the topics, $\theta_d$. 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. 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:

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

---

6This 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.”

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

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of human evaluations</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Automated Coherence</strong></td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td></td>
</tr>
<tr>
<td>NPMI</td>
<td>26 (72%)</td>
</tr>
<tr>
<td>Other</td>
<td>22 (61%)</td>
</tr>
<tr>
<td>Explicit implementation</td>
<td>22 (61%)</td>
</tr>
<tr>
<td>Explicit ref. corpus</td>
<td>10 (28%)</td>
</tr>
<tr>
<td>Perplexity w/o coherence</td>
<td>3 (8%)</td>
</tr>
</tbody>
</table>

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

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.
reference Lau et al. (2014) and the supporting code, 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.

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. 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 simplicity, we use pathwise gradients for the Dirichlet (Jankowiak and Obermeyer, 2018), rather than the rejection sampling variational inference of the authors’ primary variant. 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
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 flexibility, 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).
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.\(^\text{14}\)

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.\(^\text{15}\) 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.\(^\text{16}\)

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.\(^\text{17}\) 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.

---

\(^\text{14}\)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.

\(^\text{15}\)Code to convert topic model output into deployable questionnaires is at github.com/ahoho/topics.

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

\(^\text{17}\)Newman et al. (2010) and Lau et al. (2014) do not report an average.
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.

<table>
<thead>
<tr>
<th></th>
<th>NPMI (10-token window)</th>
<th></th>
<th></th>
<th>C_v (110-token window)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NYT</td>
<td>WIKI</td>
<td>Train</td>
<td>Val</td>
</tr>
<tr>
<td>Intrusion</td>
<td>0.27</td>
<td>0.43</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
<td>0.36</td>
<td>0.39</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>0.39</td>
<td>0.40</td>
<td>0.32</td>
<td>0.17</td>
</tr>
<tr>
<td>Rating</td>
<td>0.37</td>
<td>0.48</td>
<td>0.37</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
<td>0.41</td>
<td>0.44</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>0.44</td>
<td>0.41</td>
<td>0.35</td>
</tr>
</tbody>
</table>

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. 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.

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 $N = 1,000$ iterations, we report the proportion of significant differences detected using

---

18 We use gensim (Rehurek 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 $C_{UCL}$ (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).
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.

<table>
<thead>
<tr>
<th></th>
<th>Ref. Corpus →</th>
<th>NPMI (10-token window)</th>
<th>C_v (110-token window)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NYT</td>
<td>WIKI</td>
</tr>
<tr>
<td>Intrusion</td>
<td></td>
<td>46 / 53</td>
<td>34 / 48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44 / 76</td>
<td>33 / 78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42 / 67</td>
<td>40 / 66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45 / 50</td>
<td>45 / 51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40 / 73</td>
<td>31 / 73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39 / 66</td>
<td>36 / 66</td>
</tr>
<tr>
<td>Rating</td>
<td></td>
<td>45 / 50</td>
<td>34 / 48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45 / 51</td>
<td>41 / 47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39 / 66</td>
<td>36 / 66</td>
</tr>
</tbody>
</table>

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.

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 ρ = −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,

20Details on testing equivalence are in Section A.5.1.
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.\(^\text{21}\) 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


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.


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


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.


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.


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” → 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” → united_states_of_america).
- We filter out tokens with two or fewer characters.
- We retain only tokens that are matched by the regular expression `^[a-zA-Z\-]+\w-]*\w-$`.
- We remove tokens that appear in more than 90% of documents.
- We remove tokens that appear in fewer than $2(0.02\log_{10}D)$ 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).

22github.com/ahoho/topics
23github.com/explosion/spaCy/blob/v3.0.5/spacy/lang/en/stop_words.py
24Standard 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.
|------------------------------|-------------|------------------|------------------------------------|------------------------|--------------------------|------------------------|------------------|--------------------------|-------------|-------------------|-----------------|--------------|----------------|
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).

<table>
<thead>
<tr>
<th>Domain</th>
<th>WIKI</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Docs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>28.5k</td>
<td>273.1k</td>
</tr>
<tr>
<td>Reference</td>
<td>4.62M</td>
<td>1.82M</td>
</tr>
<tr>
<td>Mean Tokens / Doc.</td>
<td>1291</td>
<td>281</td>
</tr>
<tr>
<td>Vocab. Size</td>
<td>39.7k</td>
<td>34.6k</td>
</tr>
</tbody>
</table>

G-LDA We use gensim (Řehůřek and Sojka, 2010) as a Python wrapper for running Mallet. In Table 8a, we tune hyperparameters $\alpha$ (topic density parameter) and $\beta$ (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 ($\alpha$), the learning rate ($\eta$), 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 ($\gamma_{BN}$, comes from Card et al., 2018), and the number of epochs to anneal the KL-divergence term in the loss ($\gamma_{KL}$) (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 ($\eta$), the $L_2$ regularzation constant for the Adam (Kingma and Ba, 2015) optimizer ($W_{reg}$), and a boolean indicator of whether to anneal the learning rate ($\gamma_\eta$). 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. G-LDA (Mallet) ran on CPU only, with m5d.2xlarge instances (with 32 GiB memory, 8 CPUs).

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).

---

25https://aws.amazon.com/hpc/parallelcluster/
26See https://aws.amazon.com/ec2/instance-types/ for further details.
Model: **G-LDA**

<table>
<thead>
<tr>
<th>α</th>
<th>β</th>
<th>Optim. Interval</th>
<th>#Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>{0.01, 0.05, 0.1, 0.25, 1.0, 5.0}</td>
<td>{0.01, 0.05, 0.1}</td>
<td>{0, 10, 100, 500}</td>
<td>{1000, 2000}</td>
</tr>
</tbody>
</table>

(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**

<table>
<thead>
<tr>
<th>α</th>
<th>η</th>
<th>β_{reg}</th>
<th>γ_{HN}</th>
<th>γ_{KL}</th>
<th>#Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>{0.001, 0.01, 0.1}</td>
<td>{0.001, 0.01}</td>
<td>{0.01, 0.01, 1.0}</td>
<td>{0, 1, 100, 200}</td>
<td>{200, 500}</td>
<td></td>
</tr>
</tbody>
</table>

(b) Hyperparameter ranges for D-VAE. α is the Dirichlet prior. η is the learning rate. β_{reg} is the L_1-regularization of the topic-word distribution. γ_{HN} 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**

<table>
<thead>
<tr>
<th>η</th>
<th>W_{decay}</th>
<th>γ_{η}</th>
<th>#Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>{0.001, 0.002, 0.01, 0.02}</td>
<td>{1.2e^{-5}, 1.2e^{-6}, 1.2e^{-7}}</td>
<td>{0, 1}</td>
<td>{500, 1000}</td>
</tr>
</tbody>
</table>

(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. †: best setting for NYT; based on NPMI estimated with a 10-token sliding window over the reference corpus.

<table>
<thead>
<tr>
<th>WIKI</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-LDA</td>
<td>~2 minutes</td>
</tr>
<tr>
<td>D-VAE</td>
<td>~45 minutes</td>
</tr>
<tr>
<td>ETM</td>
<td>~260 minutes</td>
</tr>
</tbody>
</table>

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 |

<table>
<thead>
<tr>
<th>Rating</th>
<th>Answer Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very related</td>
<td>I am familiar with most of these terms, but I can answer confidently.</td>
</tr>
<tr>
<td>Somewhat related</td>
<td>I am not familiar with most of these terms, but I can answer confidently.</td>
</tr>
<tr>
<td>Not very related</td>
<td>I am not familiar with most of these terms, and so I cannot answer confidently.</td>
</tr>
</tbody>
</table>

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 ∼ Bern(0.5) (“coherent” or “incoherent”) which indexes a parameter p_{zk} ∈ [0, 1]. Annotator i samples an answer to the intruder task x_{ik} ∼ Bern(p_{zk}). 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.
- Do not base your decisions on how the word is pronounced or written or its grammatical function. For example, if you saw {apple, apricot, amel, 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 “amel” because it is not a fruit.

Here are some examples:

"baby", "crib", "diaper", "bear", "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) (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.

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 13”), sports (“1992 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.

(a) (b)

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

which has a sample of $K = 50$ binary topic labels, $z_k^{(A)}$; 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 $x_k^{(A)}$ and $x_k^{(B)}$ 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 $z_k \sim \text{Cat}(1/3, 1/3, 1/3)$. Annotator scores are noisy, so true labels are corrupted according to probabilities $p_z \in \Delta^2$. 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.
<table>
<thead>
<tr>
<th>Ref. Corpus →</th>
<th>NPMI (10-token window)</th>
<th>C (110-token window)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NYT WIKI Train Val</td>
<td>NYT WIKI Train Val</td>
</tr>
<tr>
<td>Intrusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYT</td>
<td>2.42</td>
<td>4.16</td>
</tr>
<tr>
<td>WIKI</td>
<td>4.11</td>
<td>5.08</td>
</tr>
<tr>
<td>Concatenated</td>
<td>2.82</td>
<td>4.56</td>
</tr>
<tr>
<td>Rating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYT</td>
<td>1.92</td>
<td>2.08</td>
</tr>
<tr>
<td>WIKI</td>
<td>2.97</td>
<td>4.10</td>
</tr>
<tr>
<td>Concatenated</td>
<td>2.20</td>
<td>2.75</td>
</tr>
</tbody>
</table>

Table 10: Logistic (intrusion) and ordinal probit (ratings) regression coefficients of automated metrics on human annotations. Underlined 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 $\epsilon$ 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, $z^{(A)} = z^{(B)}$. We estimate one-sided tests for each sample of human scores, with the null $H_0: \mu_1^{(B)} - \mu_1^{(A)} > \epsilon$ for some bound $\epsilon$. We minimize $\epsilon$ 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 $k = 1 \ldots K$ at each iteration, then conduct a t-test between each set. $\alpha$ and $\beta$ 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
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 rating task (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...
Table 12: Spearman correlation coefficients between mean human scores for a five-point ratings scale (rather than three), compare to Table 3. Underlined values have overlapping 95% confidence intervals with that of the largest value in each row.

<table>
<thead>
<tr>
<th></th>
<th>Ref. Corpus →</th>
<th>NPMI (10-token window)</th>
<th>C_v (110-token window)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NYT Train Val</td>
<td>NYT Train Val</td>
<td>NYT Train Val</td>
</tr>
<tr>
<td>Rating (5-pt.)</td>
<td>0.27</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>NYT</td>
<td>0.15</td>
<td>0.21</td>
<td>0.29</td>
</tr>
<tr>
<td>WIKI</td>
<td>0.21</td>
<td>0.30</td>
<td>0.28</td>
</tr>
</tbody>
</table>

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.