Aligning Human and Computational Coherence Evaluations

Jia Peng Lim
Singapore Management University
School of Computer and Information Systems
PreferredAI Research Group
jiapeng.lim.2021@smu.edu.sg

Hady W. Lauw
Singapore Management University
School of Computer and Information Systems
PreferredAI Research Group
hadywlauw@smu.edu.sg

Automated coherence metrics constitute an efficient and popular way to evaluate topic models. Previous work presents a mixed picture of their presumed correlation with human judgment. This work proposes a novel sampling approach to mining topic representations at a large scale while seeking to mitigate bias from sampling, enabling the investigation of widely used automated coherence metrics via large corpora. Additionally, this article proposes a novel user study design, an amalgamation of different proxy tasks, to derive a finer insight into the human decision-making processes. This design subsumes the purpose of simple rating and outlier-detection user studies. Similar to the sampling approach, the user study conducted is extensive, comprising 40 study participants split into eight different study groups tasked with evaluating their respective set of 100 topic representations. Usually, when substantiating the use of these metrics, human responses are treated as the gold standard. This article further investigates the reliability of human judgment by flipping the comparison and conducting a novel extended analysis of human response at the group and individual level against a generic corpus. The investigation results show a moderate to good correlation between these metrics and human judgment, especially for generic corpora, and derive further insights into the human perception of coherence. Analyzing inter-metric correlations across corpora shows moderate to good correlation among these metrics. As these metrics depend on corpus statistics, this article further investigates the topical differences between corpora, revealing nuances in applications of these metrics.

Action Editor: Ekaterina Shutova. Submission received: 19 September 2023; revised version received: 5 February 2024; accepted for publication: 29 March 2024.

https://doi.org/10.1162/coli_a_00518

© 2024 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

1. Introduction

Topic modeling is an important tool in the analysis and exploration of text corpora in terms of their salient topics (Blei, Ng, and Jordan 2003). To evaluate the effectiveness of topic models, the preponderance of topic modeling literature relies on automated coherence metrics. A key benefit is convenience, allowing researchers to sidestep expensive and time-consuming user studies. The basis for this reliance is the assumption that the coherence metrics correlate with human judgment (Mimno et al. 2011a; Lau, Newman, and Baldwin 2014; Röder, Both, and Hinneburg 2015).

The presumed correlation with human judgment should not be taken for granted. There are recent studies that challenge the assumption. Doogan and Buntine (2021) highlight the inconsistencies of automated coherence metrics via correlation analysis within each metric. Hoyle et al. (2021) claim some disagreement between human judgment and automated coherence metrics.

We postulate that the reasons behind such a mixed picture could be the differences in the topic samples and underlying corpora-dependent statistics, resulting in localized "biases" that affect the conclusions reached by respective studies. Given the importance of these metrics, we seek to conduct an extended analysis of automated coherence metrics on a larger scale than anything previously attempted. This study includes orders of magnitude greater than the number of topics typically analyzed, covering three large corpora, utilizing a comprehensive user study with extensive labels, across most of the widely used metrics.

There is a strong case for quantity. Given a vocabulary, a combinatorially large number of possible topics exist. If each topic representation is a vector of its scores on different metrics, the resulting curse of dimensionality (Bellman and Kalaba 1959) necessitates a larger sample size. We claim that evaluating thousands of topics might not be sufficient, and a larger sample size is required to approximate a diverse distribution, where sampled topics are representative of the corpus and the metrics. We surmise that the previous practice of using topic models to generate topics could introduce a biased result in the analysis. Topic models vary in performance, with a lengthy list of models compiled by Hoyle et al. (2021). There is also emerging debate on the performance between traditional and neural topic models (Doogan and Buntine 2021). For instance, Hoyle et al. (2022) find evidence that neural topic models are more inconsistent than traditional topic models, producing more variant topic sets across different runs. The stability of a topic in traditional topic models across different runs has been shown to correlate to its quality (Xing and Paul 2018). Because different topic models have different characteristics and performances, we propose generating candidate topic representations independent of topic models to evaluate the usability of coherence metrics.

1.1 Existing Contributions

We have three contributions in our previous work (Lim and Lauw 2023b). First, we begin by analyzing the inter-*metric* correlations (see Section 7). We propose a novel approach to sample representations of "topics" for the purpose of evaluating automated coherence metrics (see Section 4.1). Compared with prior works, we sample these representations free from topic model bias, and in a meaningful and diverse manner. With evaluations on three large corpora, we reaffirm that selected metrics do not contradict each other. We also highlight the underestimated effects (see Section 7.1) of $\epsilon = 1e{-12}$, used in the calculation of Normalized Pointwise Mutual Information (NPMI) (see Section 3.1) to avoid undefined logarithm zero.

Second, we extend our analysis to investigate inter-*corpus* correlations (see Section 7.2). We examine the understated differences of corpora statistics on the metrics by comparing the correlations across corpora. While such correlations exist to some degree, these metrics strongly depend on the evaluation corpus. Thus, any expectation that these metrics would correlate uniformly with human judgment on *all* possible corpora may be misplaced.

Finally, pivotal to any interpretability research, we design and conduct a user study, which is the keystone of our work (see Section 5). Compared with prior work, its design is more complex as we seek to benchmark human judgment at a finer granularity across different random user study groups (see Section 5.1). We analyze the user study results via a few novel proxy measures, revealing that human judgment is nuanced and varies between individuals. The metric correlation to human judgment is corpus-dependent, with the average participant likely to be attuned to the generic corpora (see Section 6). Accompanying the work, we also released our toolkit to enable convenient coherence evaluation of topic representations and to advance interpretability research. Our implementation and releasable resources can be found here.[1]

1.2 Extended Contributions

In this article, motivated by the results reported in our previous work, we conduct additional investigations into our previous findings to account for the effect of possible confounding factors. We also take this opportunity to include extra details and explanations that further elucidate the findings. Our previous work focuses more on automated coherence metrics, using human judgment as its benchmark. In contrast, this article reverses the direction of analysis, placing a heavier interest in human judgment while using automated coherence metrics as a tool for human judgment analysis.

1.2.1 Investigating Hyperparameter Effects. We notice one of the automated coherence metrics, C_{UMass} (Mimno et al. 2011b; Röder, Both, and Hinneburg 2015) (defined in Section 3.1), exhibits an unexpectedly lower correlation with human judgment, compared with other shortlisted automated coherence metrics. As C_{UMass} remains a popular metric, with recent studies (such as Meng et al. 2022; Zhang et al. 2022) using it as an evaluation criterion, we believe that further investigation is required to ascertain its efficacy. As the selected metrics are count-based and reliant on the corpus, different hyperparameters controlling window size *wsz* and minimum frequency *mf* of word pairs might affect the scores. Deviating from recommended hyperparameter settings, we evaluate the correlation of automated metrics against human judgment at different hyperparameter settings, showing the importance of setting a low *wsz* and *mf* (see Section 6.3). Additionally, some metrics might be order-dependent, with the ordering of vocabulary in the topic representation affecting the scores. C_{UMass} is one of these order-dependent metrics. We conduct additional analysis to determine the effect of ordering in these metrics. Our findings from this analysis show that the effect of ordering might be overstated and unnecessary (see Section 6.4). Using the newfound information, we update the results of our previous work and further strengthen its findings (see Section 7).

[1] Github repository for our toolkit: `https://github.com/PreferredAI/topic-metrics`.

1.2.2 Investigating Human Variability. Among the eight user study groups, one of the user study groups has low to no correlation with automated coherence metrics. We find a correlation between user study groups' inter-rater reliability and their self-reported English proficiency (see Section 6.1). Included in their responses, we collected our study participants' self-rated English proficiency, which we can use as a proxy measure of their true English proficiency. Via statistical analyses, we show that their reported English proficiency affects the correlation between automated coherence metrics and human judgment (see Section 6.2). When we exclude the outlier group, there is a significant increase in correlation between automated coherence metrics and human judgment. Beyond reporting correlation scores, we also conduct novel analyses to examine the human perception of coherence benchmark against computed coherence. In addition to quantifying surface-level differences (see Section 8.1), we analyze individual participants' responses to derive insights into their perception of coherence, building a framework to estimate subjectivity between individuals based on corpus statistics (see Section 8.2). With the framework, we investigate the possible strategies our study participants used to complete the task (see Section 8.3) and the effect of our study participants' subject matter experience (see Section 8.5).

2. Related Work

2.1 Topic Models

There are many approaches for topic modeling (Blei, Ng, and Jordan 2003), from non-neural based (Hoffman, Blei, and Bach 2010; Zhao, Tan, and Xu 2017), to many other neural-based methods, via autoencoders (Kingma and Welling 2014) such as Miao, Yu, and Blunsom (2016), Srivastava and Sutton (2017), Dieng, Ruiz, and Blei (2020), Zhang and Lauw (2020), and Bianchi et al. (2021), via graph neural networks (Yang et al. 2020; Shen et al. 2021; Zhang and Lauw 2022), and hierarchical methods (Meng et al. 2020). A common factor among these works is the usage of automated coherence metrics in their evaluation of topic representations produced. We select several popular metrics for evaluation as listed in Section 3. Coherent topic representations allow us to interpret the topic embeddings and reduces the opacity of models.

2.2 Applications of Topic Models

Topic models are applicable in numerous downstream tasks and play a supporting role in augmenting other models. Lau, Baldwin, and Cohn (2017) use topic modeling in neural language modeling. Wang et al. (2019) and Xu et al. (2023) proposed different topic-guided variational autoencoders for text generation. For abstractive document summarization, Wang et al. (2020) generate topic embeddings to improve the performance of transformer-based models. On stance detection (Mohammad et al. 2016), Arakelyan, Arora, and Augenstein (2023) use topic sampling and contrastive learning to achieve state-of-the-art results. On multimodal relation extraction (Zheng et al. 2021; see Section 2.3), Wu et al. (2023) use multimodal topic features in their framework to reduce reliance on internal data and exploit external data.

2.3 Relation to Knowledge Graphs

A knowledge graph is defined as a set of Entities E, Relations R, and Facts F, with a fact being represented as a triple of $\{head, relation, tail\}$, where $head \in E$ and $tail \in E$ (Ji et al.

2022). Compared with knowledge graphs, the associations within a topic representation are implicit rather than explicit. For example, WordNet (Miller 1995) is a knowledge graph that organizes words using their semantic relations (such as synonyms). Between vocabulary pairs, they may have multiple explicit relations. In topic modeling, when we consider their relation using corpus statistics, the single implicit relation assumes multiple explicit relations to varying degrees depending on its context. From the structured knowledge graphs, we can learn knowledge graph embeddings (Wang et al. 2017), which can be used in downstream tasks such as knowledge graph completion tasks to identify missing facts (Bordes et al. 2013). In contrast, we obtain topic embeddings from corpora that do not have an explicit graph structure but seek to interpret it via topic representations with the quality of its implicit graph structure measured on automated coherence metrics.

2.4 Relation to Mechanistic Interpretability

The field of Mechanistic Interpretability attempts to reverse the computations within models in pursuit of explainability (Olah et al. 2020), with efforts directed towards transformer-based models (Elhage et al. 2021; Geva et al. 2022; Cunningham et al. 2023; Bricken et al. 2023). We believe there is potential to apply topic modeling methodologies and evaluation in this area. In Lim and Lauw (2023a), we proposed an approach that assigns multiple topic representations to single neurons residing in the multilayer perceptron layers in decoder-only transformer models. Automated coherence metrics enable the construction and evaluation of millions of topic representations, optimized for human coherence, from millions of neurons.

2.5 User Studies in Metric Evaluation

Mimno et al. (2011a) utilize expert annotators to independently label 148 topic representations, using another ten expert annotators to evaluate topic representations via intruder word detection tasks. Röder, Both, and Hinneburg (2015) benchmark topics against different permutations of metrics with an evaluation set containing 900 topics with human ratings aggregated from prior studies (Aletras and Stevenson 2013; Lau, Newman, and Baldwin 2014; Rosner et al. 2014). In Hoyle et al. (2021), a minimum of 15 crowd workers were employed in simple rating and word intrusion tasks evaluating 40 topic-model-generated (Griffiths and Steyvers 2004; Burkhardt and Kramer 2019; Dieng, Ruiz, and Blei 2020) and 16 synthetic random topics. In Doogan and Buntine (2021), their largest user study required four subject matter experts to create 3,120 labels across 390 topics generated via topic models (Blei, Ng, and Jordan 2003; Zhao et al. 2017). In comparison, our study has large quantities in both topic representations and study participants, annotating 800 unbiasedly sampled topic representations split between 40 user study participants with at least an undergraduate level of education, generating 180K word pairs. Each question has 45 possible combinations of word pairs, with binary labels denoting coherence relations between word pairs. Our automated experiments deal with hundreds of thousands of unique topics.

2.6 Linguistics Research

Garimella, Banea, and Mihalcea (2023) use word models with topic-based features to analyze the relationship between language and demographics. Similar to our work, they also utilized context windows in text. Beyond the demographic level, we investigate the relationship with language at the individual level, specific to the perception of coherence in topic representations. Bo, Fu, and Lim (2023) extensively analyzed the relationship between English language proficiency and academic performance, showing that proficiency scores strongly predict academic results. Their study is particularly relevant as our user study participants are from a similar population in terms of culture, geography, and education level.

3. Preliminaries

In this section, we define the selected automated coherence metrics that utilize corpus statistics and describe the corpora we use to obtain the word probabilities.

3.1 Coherence Metrics

We follow the definition styles of Röder, Both, and Hinneburg (2015), where direct confirmation measure m is a function of a word-pair statistic. Direct coherence metric is defined as a mean aggregation of m between word pairs Equation (1), where topic representation t is a k-sized set of words. Let $p = \frac{|t| \cdot (|t|-1)}{2}$, representing the number of word pairs in a topic.

$$C(t, m) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i > j}} m(w_i, w_j) \tag{1}$$

C_{NPMI} Equation (2) is the mean aggregation of m_{nlr}.

$$C_{\text{NPMI}}(t) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i > j}} m_{nlr}(w_i, w_j) \tag{2}$$

m_{nlr} is defined as the NPMI (Bouma 2009) value between word-pair statistics in a topic Equation (3). The numerator is the Pointwise Mutual Information (PMI) (Church and Hanks 1990) between a word-statistic pair. The denominator, a negative logarithm of the word-statistic pair's joint probability, normalizes the value's range to between 1 and -1. The inclusion of $\epsilon = 1\text{e}{-12}$ is used to avoid undefined logarithm zero when $P(w_i, w_j) = 0$.

$$m_{nlr}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \tag{3}$$

C_{UCI} (Lau, Newman, and Baldwin 2014) is another coherence metric that is similar to C_{NPMI}. Instead of NPMI, C_{UCI} uses PMI without normalization. Since there is a strong correlation between PMI and NPMI, we omit C_{UCI} from our analysis.

C_{UMass} is the mean ordinal aggregation of m_{lc} (Mimno et al. 2011a) Equation (4). Within any given topic, words are ordered based on $P(w|\text{topic})$ in descending order. The summation notations of Equation (4) reflect the ordering of words.

$$C_{\text{UMass}}(t) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i > j}} m_{lc}(w_i, w_j) \tag{4}$$

m_{lc} measures the logarithmic conditional probability between ordered word-pair in a topic, where $\epsilon = 1\text{e}{-12}$ is used to avoid undefined logarithm zero when $P(w_i, w_j) = 0$ Equation (5).

$$m_{lc}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \tag{5}$$

C_P is the mean ordinal aggregation of Fitelson's coherence (Fitelson 2003) m_f Equation (6). Similar to C_{UMass} Equation (4), C_P also uses the same ordering criteria.

$$C_P(t) = \frac{1}{p} \sum_{w_i \in t} \sum_{\substack{w_j \in t \\ i > j}} m_f(w_i, w_j) \tag{6}$$

m_f can be interpreted as the degree to which w_i supports w_j, between ordered word pairs in a topic Equation (7).

$$m_f(w_i, w_j) = \frac{P(w_i|w_j) - P(w_i|\neg w_j)}{P(w_i|w_j) + P(w_i|\neg w_j)} \tag{7}$$

C_V (Equation 8) is the final metric that we are evaluating. C_V is defined as an indirect coherence metric, as it uses word-group relations instead of word pair relations used in aforementioned direct coherence metrics.

$$C_V(t, \gamma) = \frac{\sum_{w_i \in t} s_{cos}(v(w_i, t, \gamma), \bar{v}(t, \gamma))}{|t|} \tag{8}$$

Intuitively, it measures the mean cosine similarity Equation (9) between each word's feature vector and the topic's feature vector.

$$s_{cos}(\vec{v_i}, \vec{v_j}) = \frac{\sum \vec{v_i} \cdot \vec{v_j}}{||\vec{v_i}||_2 \cdot ||\vec{v_j}||_2} \tag{9}$$

A topic feature vector $\bar{v}$ represents the sum of all of its words' feature vectors Equation (10).

$$\bar{v}(t, \gamma) = \sum_{w_j \in t} v(w_j, t, \gamma) \tag{10}$$

A word's (w) feature vector v is defined as a vector of its NPMI value with the other words w_j from its topic representation t Equation (11).

$$v(w, t, \gamma) = \{m_{nlr}(w, w_j)^{\gamma} \ \forall w_j \in t\} \tag{11}$$

For indirect confirmation measure $\tilde{m}$, instead of directly using word probabilities, it uses m to create a vector of features v (Aletras and Stevenson 2013) that represent a word w from the topic t it belongs to, distorted by hyperparameter γ Equation (11). We

Table 1
Numerical descriptions of the corpora used. Lemmatized variants are similar except for ArXiv-lemma, with a vocabulary size of 22K.

Corpus	#Docs.	Mean Doc. Size	Vocab. Size
ArXiv	2.09M	75	26K
PubMed	1.07M	1,500	39K
Wiki	5.51M	217	40K

will evaluate γ at 1 and 2, denoted by superscript C_V^γ.[2] Subscript $\not\in$ denotes the absence of epsilon. In this article, $\not\in$-variants are $C^\gamma_{V,\not\in}$ and $C_{\mathrm{NPMI},\not\in}$.

3.2 Corpora

We use word co-occurrences statistics obtained from three large corpora: ArXiv,[3] PubMed,[4] and Wiki.[5] Table 1 numerically describes the corpora. We further elaborate on each corpus:

1. **ArXiv**. We use the ArXiv abstracts dataset, considering each abstract as a document. These abstracts mainly constitute research work related to non-medical science disciplines.
2. **PubMed**. We use the PubMed Central Open Access Subset that contains journal articles and preprints related to medical research and information. We consider each article body as a document, removing citations within it.
3. **Wiki**. We use the English-Wikipedia dump of August 2022 processed using Attardi (2015). We consider the content of the article as a document. To evaluate for correctness of our toolkit, we use the popular benchmark Palmetto (Röder, Both, and Hinneburg 2015), which uses a subset of Wikipedia 2011.

For each corpus, we apply processing steps suggested in Hoyle et al. (2021), retaining up to 40K frequently occurring words.[6] Additionally, we generate a lemmatized (denoted with the suffix -lemma) and unlemmatized variant (original) for further analysis. Appendix B contains additional information on the shared vocabulary between corpora.

[2] Prior to version 0.1.4 (released 21 September, 2022), Palmetto's (Röder, Both, and Hinneburg 2015) γ was set to 2.

[3] ArXiv dataset downloaded from `https://www.kaggle.com/datasets/Cornell-University/ArXiv`.

[4] PubMed articles obtained at `ncbi.nlm.nih.gov/pmc/tools/openftlist`.

[5] Wikipedia dumps obtained at `dumps.wikimedia.org`.

[6] We use spaCy (`https://spacy.io/`) for tokenization, named entity recognition, and lemmatization.

4. Sampling Topic Representations

Intuitively, if two different metrics are to correlate with human judgment, we would expect correlations between the scores of these metrics. However, it is claimed in Doogan and Buntine (2021) that these metrics do not correlate well. There are a few tested methods to generate topics: from topic models (Aletras and Stevenson 2013; Lau, Newman, and Baldwin 2014), beam search optimized on coherence (Rosner et al. 2014), and random sampling of words (Hoyle et al. 2021). Considering only optimized and random (incoherent) topic representations will result in a skewed distribution. In contrast, we seek to mine topic representations that emulate a balanced distribution for a meaningful comparison. Furthermore, there is also a desire for uniqueness among topics, which avoids repetition and is representative of the corpus. We propose a new non-topic modeling approach to sample topics to evaluate these metrics.

4.1 Approach: Balanced Sampling

The problem of mining topics of k words can be mapped to the classical k-clique listing problem (Chiba and Nishizeki 1985; Danisch, Balalau, and Sozio 2018). To generate a meaningful distribution of topic representations, we map the corpus-level information as a graph, treating each word from its vocabulary set V as a vertex. Each word will share an edge with every other word. m_{nlr} is used to determine the value of the edges between two vertices as its normalized range is intuitive, allowing easy identification of the value ranges for the generation of sub-graphs. Using m_{lc} and m_f, on the other hand, increases the sampling's complexity as they are order-dependent. As a result, their sub-graph contains bi-directional edges. Conducting sampling runs using any m, not only m_{nlr}, might introduce sampling bias, favoring certain topic representations, which our approach seeks to mitigate. Figure 1 illustrates an overview of our approach.

The initial graph will be a complete graph of $|V|$ vertices, where its k-sized sub-graph is a topic representation of k words. Combinatorially, there are $C_k^{|V|}$ possible unique topics. It is practically infeasible and unnecessary to list all k-cliques. For a more tractable approach, we modify the routine from Yuan et al. (2022) (Algorithm 1, 2) to include the following properties:

1. **Sub-graphs of varying quality**. This routine seeks to generate smaller graphs from the original complete graph to cover the spectrum of topic

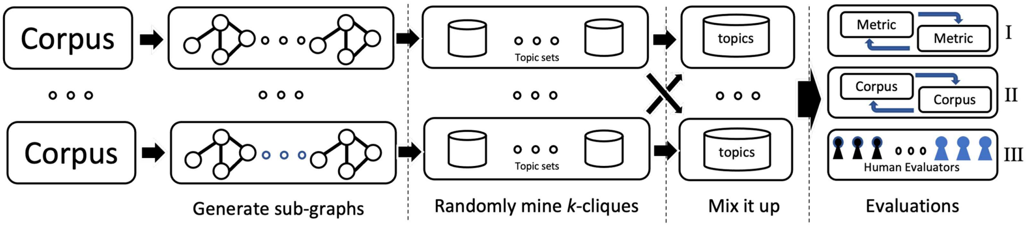


Figure 1
Illustration of our balanced sampling approach. For each corpus, we generate sub-graphs and randomly mine k-cliques as detailed in Section 4.1. For evaluation, we mix sampled topics from different corpora and conduct three evaluations: inter-metric (see Section 7), inter-corpus (see Section 7.2), and human evaluation (see Section 5).

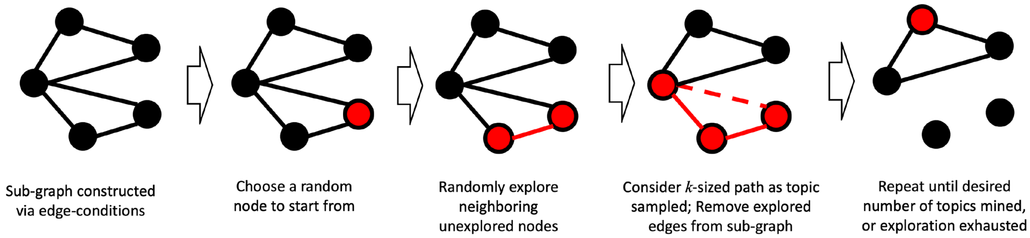

Sub-graph constructed via edge-conditions
Choose a random node to start from
Randomly explore neighboring unexplored nodes
Consider k-sized path as topic sampled; Remove explored edges from sub-graph
Repeat until desired number of topics mined, or exploration exhausted

Figure 2
The topic representation sampling process. We initially construct a sub-graph via edge conditions, which we then use to explore vertices in a clique-constrained manner, starting from a random vertex. The exploration stops once we obtain a k-clique (topic representation). We repeat the search, stopping the process when it is impossible to construct another k-clique or when we retrieve enough k-cliques from the sub-graph.

quality. Edges on the graph are eliminated conditionally via their value, and the remaining edges and connected vertices constitute the new sub-graph, enabling the creation of three different kinds of sub-graphs: *pos* where edge-values are above a given lower-bound, *mid* where edge-values are between threshold values, and *neg* where edges are below an upper-bound (see Appendix B).

2. **Topic extraction**. Inspired by Perozzi, Al-Rfou, and Skiena (2014), instead of iterating through all the neighboring nodes or searching for the next best node, we randomly select its unexplored neighbor node, which shares an edge with all explored nodes, to explore. We extract the explored k-path as our sampled topic.
3. **Topic uniqueness**. To attain a variety of topics, we remove all edges in a mined clique, making it impossible to sample a similar topic representation from the same sub-graph. Figure 2 illustrates this feature.
4. **Balanced distribution of topics**. For a given corpus, we introduce additional common topics sampled from different corpora, which differ in word distribution. We refer to this segment of external topic representations as *ext*. Lastly, *random* segment comprises groups of random words, included to represent a distribution absent from the other sampled segments. Table 2 shows the result from this mining approach. The final distribution would thus be more balanced, comprising topics of varying scores along the spectrum.

Table 2
The average quantity of topics mined by our balanced sampling approach by segments per corpus from the five independent sampling runs. Quantities of lemmatized variants are similar except for *ext* segment where it has half the numbers.

Corpus	*neg*	*pos*	*mid*	*random*	*ext*	Total
ArXiv	66,007	2,120	14,436	10,000	49,777	142,340
PubMed	10,450	3,310	8,218	10,000	61,035	93,013
Wiki	56,903	21,698	35,195	10,000	136,036	259,832

Algorithm 1 SDegreeList(k, T, C, $\vec{G}$) ▷ k denotes remaining vertices to be added to T

```
for $v \in$ Permutate($C$) do                          ▷ Randomly pick a vertex $v$ from Candidates $C$
    $\hat{C} \leftarrow N_v^+ \cap C$                  ▷ Find common vertices in $v$'s neighbors $N_v^+$ and $C$
    if $k = 2 \wedge |\hat{C}| > 0$ then               ▷ Base Case: completed, only 2 vertices required
        $T \leftarrow T \cup \{v$, a random vertex from $\hat{C}\}$
        remove $(u_i, v_j)$ from $\vec{G}\ \forall v_i, v_j \in T$      ▷ Sub-graph reduction
        Return $T$
    end if
    if $|\hat{C}| > k - 2$ then
        $\hat{T} \leftarrow$ SDegreeList($k-1$, $T \cup \{v\}$, $\hat{C}$, $\vec{G}$)   ▷ Recursive step to explore further
        if $|\hat{T}| \neq 0$ then                     ▷ Propagate completed result
            Return $\hat{T}$
        end if
    end if
end for
Return $\emptyset$                                     ▷ Base Case: exploration exhausted
```

Algorithm 2 Main(G, k, target)

```
$G \leftarrow$ PRE-CORE($G, k$)                       ▷ Prune vertices with less than $k$ edges from $G$
Generate DAG $\vec{G}$ from $G$                        ▷ DAG: Directed Acyclic Graph
$R \leftarrow \emptyset$
for $v \in$ Permutate($\vec{G}$) do                    ▷ Randomly explore $\vec{G}$
    $T \leftarrow$ SDegreeList($k-1, \{v\}, N_v^+, \vec{G}$)   ▷ Enter recursive step
    if $|T| = k$ then                                  ▷ It is possible that exploration yields $\emptyset$
        $R \leftarrow R \cup \{T\}$
    end if
    if target $= |R|$ then   ▷ Early stop when we extract enough topic representations $R$
        return $R$
    end if
end for
```

In Algorithm 2, the pre-processing step, PRE-CORE, to reduce complexity remained unchanged. This step is skippable when the graph is large and dense, such as during *neg* sub-graphs generation. Our modification in Algorithm 1 and Algorithm 2 introduces randomness via permutations and early stopping when Algorithm 1 finds a k-clique and a desired number of k-cliques found in Algorithm 2. The sub-graph reduction routine is in Algorithm 1.

4.2 Hyperparameter Threshold Selection

From Table 3, we have three different thresholds for *pos*. As the lower bound increases for *pos*, it becomes increasingly difficult to sample cliques that meet the criteria, with some sampling runs returning no samples. Sampling for *mid* is easier, in terms of speed and quantity sampled, compared to *pos* as more edges fit the relaxed criteria. For ArXiv and Wiki, we use bounds $(-0.05, 0.15)$ while for PubMed, we use bounds $(0, 0.15)$.

Table 3
Hyperparameter threshold for different sub-graphs. Multiple thresholds are indicative of multiple runs. *random* and *ext* are not hyperparameter dependent. When possible, we choose hyperparameters to control sub-graph density.

	ArXiv	PubMed	Wiki
pos (>)		$0.05, 0.1, 0.15$	
mid	$(-0.05, 0.15)$	$(0, 0.15)$	$(-0.05, 0.15)$
neg (<)	$-0.2, -0.4$	-0.2	$-0.1, -0.4$

We use a stricter bound for PubMed as relaxing the bounds will result in a sample similar to a random distribution. While *neg* seems easier to sample, compared with *pos* or *mid*, it is not always guaranteed to be easier. When we set *neg*'s upper bound to -0.4 for PubMed, sampling cliques became challenging. Selected hyperparameter thresholds should return feasible samples while allowing us to avoid emulating a random distribution, which we can trivially obtain.

4.3 Optimizing Position-Dependent Scoring

Since our approach (see Section 4.1) does not produce $P(w|t)$, we can locally optimize the word positions within a topic to obtain the best possible score for position-sensitive metrics C_{UMass} and C_P. This additional step ensures fairness in our evaluation of order-dependent metrics. We use subscript s to denote alphabetical order and subscript o to denote optimized positions. In our correlation evaluations, we use topic representation t of size $k = 10$, with words w arranged based on $P(w|t)$ in descending order.

Given a set of k words as a topic, we seek to optimize the position-dependent score (see Algorithm 3). This problem is reducible to a weighted activity selection problem,

```
Algorithm 3 OptPlacements($\vec{G}$)                ▷ Where $\vec{G}$ is a word-score graph
  order ← ()
  for $i \in [0, |\vec{G}|)$ do:
      best_edge_score ← $-$inf
      best_vertex ← $-1$
      for $v \in \vec{G}$ do
          incoming_edges_score ← $\sum_{n \in \vec{G}, v \neq n} e_{n,v}$
          outgoing_edges_score ← $\sum_{n \in \vec{G}, v \neq n} e_{v,n}$
          difference ← incoming_edges_score $-$ outgoing_edges_score
          if difference > best_edge_score then
              best_edge_score ← difference
              best_vertex ← $v$
          end if
      end for
      order$_i$ ← best_vertex
      Remove best_vertex and its edges $\in \vec{G}$
  end for
  return order
```

akin to finding a max-weight independent set in an interval graph, and solvable in polynomial time (Bar-Noy et al. 2001). Consider a word w at the j^{th} position, with index j starting from 0, its j incoming edges and $k - j + 1$ outgoing edges representing its rank in the order. The incoming edges indicate the subset of preceding words to w, while the outgoing edges indicate the subset of the remaining succeeding words. We define an activity's position in the ordering using its preceding and succeeding activities. Each activity has an equal interval with its weight determined by the difference of outgoing and incoming edges to all other words scored via m. We can transform the activities into an interval graph, with $|C^l_j| \cdot |C^l_{l-j+1}|$ combinatorial number of possible instances for each word per interval in the schedule. Our transformation will result in an interval graph of k disjoint graphs. While the number of activities might seem combinatorially explosive, selecting the first activity only involves k activities. Each selection will prune multiple branches of possible orderings, resulting in $k - j$ choices after j selection. Hence, we are only required to select the best activity within each disjoint graph conditioned on availability (for unselected words).

5. User Study Design

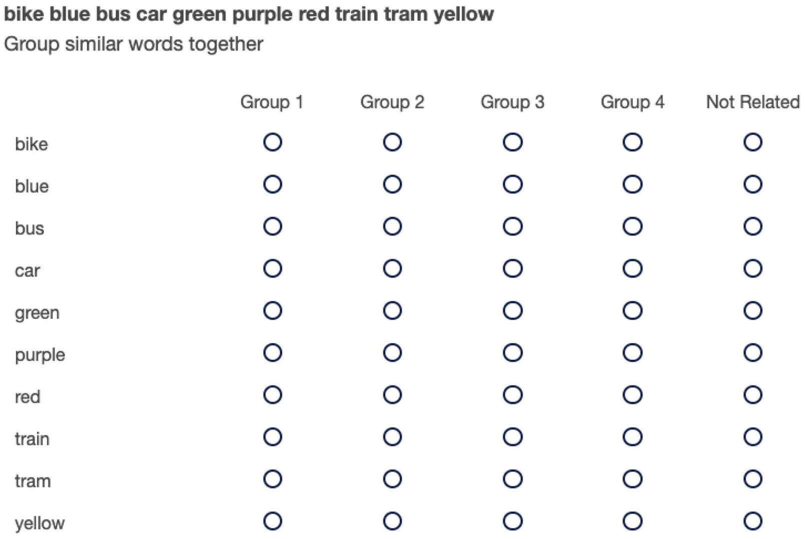
Previous studies measure human judgment through simple evaluation tasks such as rating the coherence of a topic on a few-point ordinal scale (Mimno et al. 2011a; Aletras and Stevenson 2013), identifying the intruder word introduced into the topic representation (Chang et al. 2009), or both (Lau, Newman, and Baldwin 2014; Hoyle et al. 2021). For word intrusion, outlier detection signals the cohesiveness of the topic representation, which is similar to rating topics on an ordinal scale. However, for both tasks, qualitative gaps might exist. In word intrusion, study participants are restricted to just one outlier per topic, assuming perfect coding, resulting in an exponential drop in scoring. For simple ratings, topic representations of differing qualities might get the same score as its ordinal scale is rigid in its score increments.

Additionally, while the decisions between human annotators might be equivalent, it is not evident if their thought processes are similar. The key reason for this line of inquiry stems from the observation that everyone is different in some aspects, such as knowledge, culture, and experiences. Assuming our prior beliefs influence our understanding of words, what and how we perceive similarity and coherence might differ from person to person. For these reasons, we are motivated to design a user study that combines both word intrusion and topic rating tasks but is differentiated at a finer granularity such that we can quantify the decision-making process. We instruct study participants to cluster word groups that indicate coherent and outlier word groups. We then examine the relationships between automated coherence metrics and different proxy tasks derived from the user study.

User studies are typically not replicated precisely. While the methodology may be similar, the participants and questions asked are likely to differ. We designed our study to encompass multiple groups to investigate replicability. By examining the variation resulting from the difference in participants and question sets, we hope to ensure some degree of confidence in the study.

5.1 User Study Definitions

For our study, we recruit eight user study groups $U = \{U_1, \ldots, U_8\}$, with five study participants per group, totaling 40 study participants. Most of our study participants



bike blue bus car green purple red train tram yellow

Group similar words together

	Group 1	Group 2	Group 3	Group 4	Not Related
bike	○	○	○	○	○
blue	○	○	○	○	○
bus	○	○	○	○	○
car	○	○	○	○	○
green	○	○	○	○	○
purple	○	○	○	○	○
red	○	○	○	○	○
train	○	○	○	○	○
tram	○	○	○	○	○
yellow	○	○	○	○	○

Figure 3
Format of questions presented to study participants. Each word is only assigned to the group when the word is deemed coherent with other words belonging to the group. The topic displayed in this example was manually curated to serve as a verification question and not included in the evaluation. Refer to Lim and Lauw (2023b) for actual examples shown.

recruited have completed or are pursuing an undergraduate program from tertiary institutions where English is the primary language of instruction, requiring sufficient English competency for admissions (specified in Bo, Fu, and Lim 2023). As such, we expect our applicants to have adequate competency in English, sufficient to understand the tasks, and to recognize most, if not all, of the words presented. Before distributing the study, participants are allocated randomly to only one study group and disallowed to attempt another set of questions.

For each study group, we prepared eight unique question sets $T = \{T_1, \dots, T_8\}$, each containing 100 10-word topic representations, $T_i = \{t_{1,i}, \dots, t_{100,i}\}$ and $t = \{w_{0,j,i}, \dots, w_{9,j,i}\}$. For each participant $U_{u,i} \in U_i$, we present each $t_{j,i} \in T_i$ individually and sorted alphabetically. We ask participants to cluster words in $t_{j,i}$ that they deem similar to form coherent word groups g, where their response $R_{u,j,i}$ to $t_{j,i}$ is a set of unique g. To limit the complexity of the given task, we constrain each word to belong to one coherent word group. Additionally, when a study participant determines that the word is unrelated to the given topic representation, it forms its group of one. We format the response as a Likert matrix grouping coherent words (see Figure 3), mandating a response for each word $w_{k,j,i} \in t_{j,i}$. Refer to Table 4 for the table of notations and Appendix A for actual instructions given to the study participants.

5.2 Topic Representation Selection

We construct an initial pool of 1,000 topics. We randomly sampled 400 common topic representations from Wiki, ArXiv, and PubMed for parity between corpora. To represent non-scientific topics, we randomly sampled 200 topics from Wiki that do not appear in

Table 4
Table of notations.

Variable	Definition
C	Automated coherence measure
U	Sets of user study group
T	Sets of 100 questions
$t_{j,i}$	j^{th} indexed 10-word topic representations $\in T_i$
$w_{k,j,i}$	k^{th} indexed word in topic representation $t_{j,i}$
$R_{u,j,i}$	u^{th} indexed participant in study U_i response to topic representation $t_{j,i}$
g	group of words clustered together $\in R_{u,j,i}$
P	Proxy tasks based on human responses

ArXiv/PubMed. For ArXiv/PubMed exclusive topics, we randomly sampled 200 topic representations each, with these topics also appearing in Wiki. Since most coherence evaluations occur in the positive domain of C_{NPMI}, we conduct sampling in a 7:1:1:1 ratio of *pos/mid/neg/random* segments of the corpus, seeking to emulate a uniform score distribution. To account for word familiarity, we select lemmatized topics with words found in 20K most frequently used words, based on Corpus of Contemporary American English. The 14K most frequently used words, including proper nouns, will account for 99% of running words (Nation 2006; Beglar 2010), where 99 out of 100 consecutive words in any text will belong to the 14K set. In our manual review of the topic representations, we do not find any unrecognizable words. However, if there were difficult words, the study allowed participants to look up the definitions of words.[7] We pair each user study group with a set of randomly sampled 100 topics from the pool without replacement. For topic representations not found in ArXiv or PubMed, we exclude them during the evaluation of those corpora.

5.3 Proxy Tasks

Representing coherence as word clusters allows us to derive deeper insight into what we perceive as human judgment. We distill our user study into a few proxy tasks, measuring the correlation (Spearman's ρ) of the user responses to the automated coherence metrics.[8] We propose three topic-level human coherence measures. Using the density of human agreement, we define P_1 Equation (12) as the mean agreement of U_i on all possible word pairs on any topic $t_{j,i}$.

$$P_1(t_{j,i}) = \frac{1}{|U_i|} \sum_{u=1}^{|U_i|} \frac{\sum_{g \in R_{u,j,i}} |g|(|g|-1)}{|t_{j,i}|(|t_{j,i}|-1)} \tag{12}$$

[7] See Lim and Lauw (2023b) for topic representations T_1 presented to U_1.

[8] We use Spearman's ρ instead of Pearson's r, as we generally obtain a better r (than ρ shown) through distortion of scores. To ensure parity, we use ρ instead.

If $t_{j,i}$ have perfect agreement on coherence, we expect $P_1(t_{j,i})$ to have a value of 1, and for incoherence, a value of 0. Subsequently, we consider the largest selected word group within $t_{j,i}$, and define P_2 Equation (13) as the mean of this measure among U_i.

$$P_2(t_{j,i}) = \frac{1}{|U_i|} \sum_{u=1}^{|U_i|} \max\{|g| : g \in R_{u,j,i}\} \tag{13}$$

A value of 1 will suggest that each word in $t_{j,i}$ has no relations to each other, and a value of $|t_{j,i}|$ suggests perfect agreement on coherence. Lastly, we define P_3 Equation (14) as the mean number of annotated word groups among U_i.

$$P_3(t_{j,i}) = \frac{1}{|U_i|} \sum_{u=1}^{|U_i|} |R_{u,j,i}| \tag{14}$$

The interpretation of P_3 is the inverse of P_2. While these group-wise measures might seem similar, they measure different nuances of human-annotated data. P_1 evaluates the sizes of multi-word groups, weighted towards larger groups. P_2 only accounts for the maximum word group size, which ignores the properties of the other remaining group. P_3 disregards group sizes to a certain extent and includes single-word "outlier" groups. We evaluate these measures' correlation against various $C(t_{j,i})$.

6. User Study Results

From our previous work (Lim and Lauw 2023b), scores from the proxy tasks correlate to automated coherence metrics. In Figure 4, the mean results of study groups on the three proxy tasks indicate correlations between human judgment and some automated coherence metrics (see Lim and Lauw [2023b] for individual group results). These results include user study group U_3, which exhibits outlier results, prompting us to investigate the possible reason for the observations (see Section 5). Since most of our study participants have some science-related background, we are surprised by ArXiv's lower correlation scores relative to Wiki in each proxy task. Additionally, ArXiv's correlation scores have a higher variance when compared to PubMed and Wiki. Lastly, we also observed a weak correlation between C_{Umass} and the proxy tasks. Corpus hyperparameters could be a confounding factor, and we examine its effect (see Section 6.3).

6.1 Inter-Rater Reliability (IRR)

Many factors will affect the variation for IRR (Belur et al. 2021). For our user study, we attempted to mitigate some of these factors. Considering the framing and education factors, we include a short introductory primer and some example questions. Study participants will view these materials before starting the user study (Appendix A). To alleviate fatigue, we allowed the study participants up to a week to work on the task, pausing and continuing at their own pace. We were not concerned about the learning effect, as topic representations of various themes exist in our question sets. Additionally, the correctness of the task is subjective to their personal preference. As our objective is to poll for their beliefs, with many possible valid answers, reviewing their responses to enforce consistency between study participants is unnecessary. While

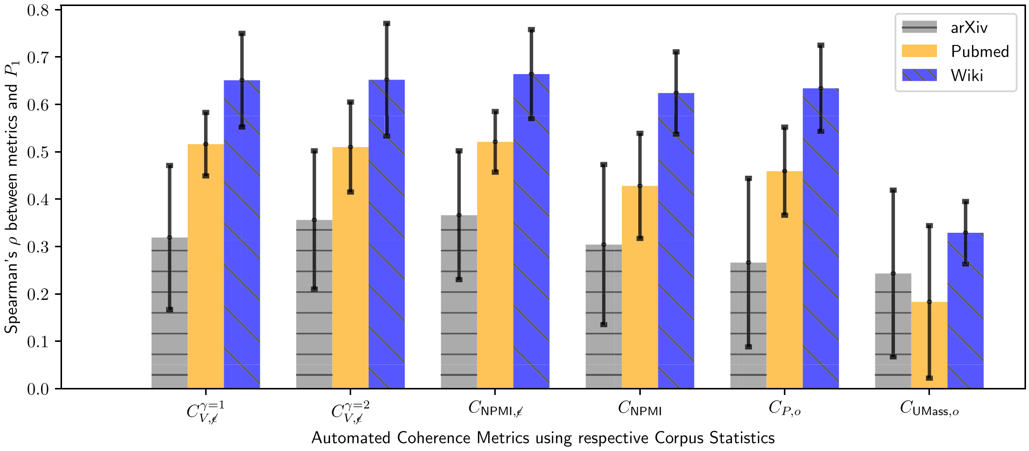


Figure 4
Bar graph presenting evaluation results of Spearman's ρ between automated coherence metrics and density of agreement among study participants (P_1). The results visualized are the mean correlation scores from the 8 study groups with error bars. We omit reporting the lemmatized version of the corpus as its values are similar to the original. $C_{\text{UMass},s}$ and $C_{P,s}$ omitted as they are almost identical to their o variant. The strength of correlation scores from using maximum coherent group size (P_2) and coherent group count (P_3) are similar to these results (Appendix Table B1). These results are from Lim and Lauw (2023b).

there are correlations between proxy measures and coherence metrics, it does not imply that individual study participants have similar responses. IRR allows us to quantify the difference in responses across different study groups.

6.1.1 Methodology. To measure IRR, we use Krippendorf's α (Krippendorff 2011), defining pair-wise rater similarity with two different metrics to measure common answers between raters: Jaccard distance (Jaccard 1912) and Measuring Agreement on Set-valued Items (MASI) distance (Passonneau 2006). We treat each $w_{k,j,i} \in t_{j,i}$ as a multi-classification question, comprising other words (in $t_{j,i}$) and "not related" as categories, producing Boolean vector representations. When given two sets A and B, Jaccard distance calculates their intersection over their union, $(A \cap B)/(A \cup B)$, while MASI distance gives fixed scores: 0 for disjoint sets, 1/3 for intersecting sets where neither sets are a subset of the other, 2/3 when a set is a subset of another, and 1 for identical sets. Comparing the two metrics, they are equivalent in the cases of disjoint and identical sets. However, MASI rewards subsets but penalizes other non-empty intersections.

6.1.2 Analysis. From Table 5, except U_5 and U_7, the ordering of the study groups is similar for both metrics. U_1 scored the highest for both distance metrics, while U_3 scored

Table 5
Individual Krippendorf's α for each user study using Jaccard and MASI distance.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean $\bar{\alpha}$ (S.D.)
α (Jaccard)	0.463	0.391	0.323	0.376	0.325	0.366	0.333	0.347	0.366 ± 0.04
α (MASI)	0.372	0.309	0.224	0.284	0.239	0.277	0.237	0.261	0.276 ± 0.05

the lowest. α (Jaccard) is higher than α (MASI) across the different user study groups. A completely random study response will have an α (Jaccard) of 0.12 and α (MASI) of 0.092, significantly less than the study's respective $\bar{\alpha}$, giving us some confidence about the reliability of the responses.

6.1.3 Discussion. There might be two plausible reasons for the difference in the α metrics observed across the different user study groups. When two responses have a non-empty intersection with neither being a subset of the other, the results suggest that most intersections of such nature are more than 1/3 of the union of sets. Alternatively, when either response is a subset of the other, the results also suggest that the subset is likelier to have more than 2/3 of the elements in the superset. Considering the exponential number of possible combinations for each topic response, these values show that our study participants have varied responses but with some overlapping similarities.

6.2 English Proficiency Effect

Even though we recruit from an English-speaking population, there could still be some variation in writing and reading abilities. Before the study, we asked prospective applicants to rate their English proficiency using a sliding scale from 0 to 100. Self-ratings of English proficiency can estimate the study participants' English competency. Xu (1991) found that self-rated English Proficiency was the most significant predictor of the perceived level of academic difficulty for reading tasks. In a study conducted by Takahashi (2009), students who rated themselves higher also scored better academically. Our participants would have had some form of standardized English test before being given admission into a tertiary education program. However, for privacy reasons, we refrained from asking participants for their academic grades or test scores for English. As anonymity is guaranteed, reporting self-assessed English proficiency is a reasonable way to assess their opinion of their abilities relative to their peers, giving us a viable proxy estimate of their true English proficiency. We reiterate that self-reporting a low score does not mean they are incompetent in English.

6.2.1 Group Analysis. Figure 5 shows the distribution of participants' self-rated English proficiency. The distribution is left-skewed, with many participants rating themselves proficient in English. From Figure 6, out of all the user groups, U_3 seems to buck the trend, reporting lower to no correlation with the proxy tasks. Its correlation scores are two to three times the standard deviation less than the mean, a clear outlier. No participants from U_3 self-rate their proficiency score above 80, whereas the rest of the groups contain multiple participants highly confident in their mastery of English.

6.2.2 Linear Regression Analysis. Using the mean self-rated proficiency scores (see Table 6), we conduct a simple linear regression analysis against their respective IRR scores and the correlation scores of C_{NPMI} from the three different proxy tasks. We select C_{NPMI} as it has moderate to good correlation with other automated coherence metrics on Wiki (Lim and Lauw 2023b). For the regression analysis, we set the bias parameter to 0 as we expect user groups with zero English proficiency to have low IRR and no correlation to the various proxy tasks. The mean self-rated English proficiency of the group (X) is assumed to be normally distributed, with the Shapiro-Wilk test (Shapiro and Wilk 1965) giving a p-value of 81.61%. We analyze the relationship of X with different dependent variables (Y): IRR and the different correlation scores from the various proxy tasks. From

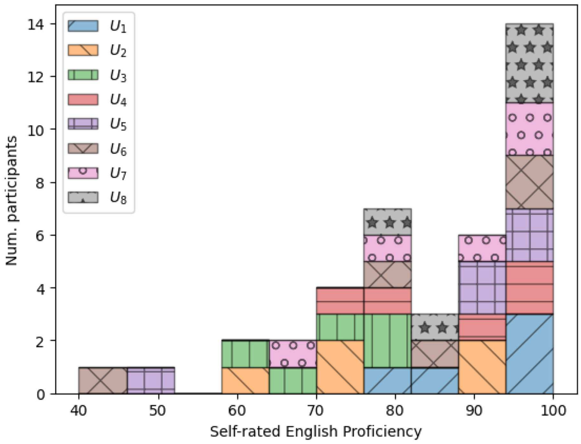


Figure 5
This histogram visualizes the self-reported English proficiency of our study participants (exact values in Table B14). Most of our study participants report high confidence in their English proficiency, with most user study groups containing some of these individuals. However, study group U_3 is a notable exception, where its participants rated their English proficiency ≤ 80.

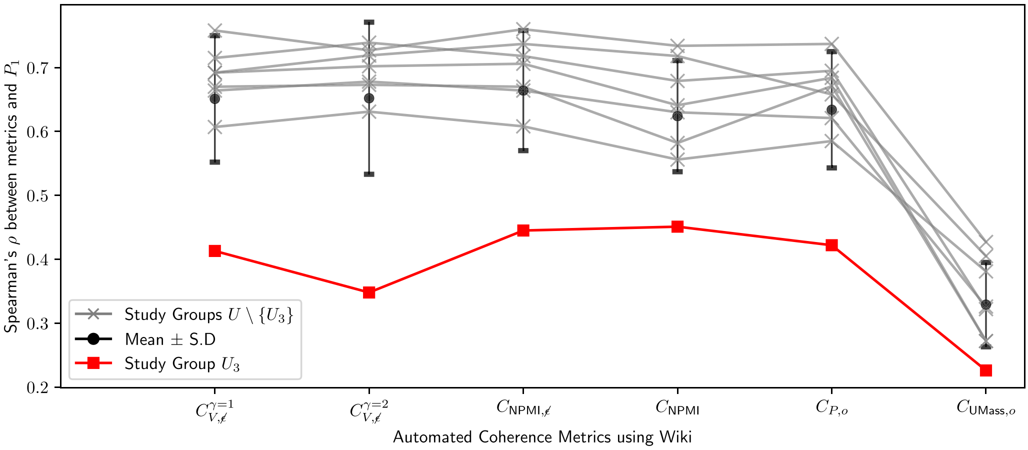


Figure 6
Line plot presenting a detailed breakdown of Proxy Task I using Spearman's ρ of density of agreement and coherence scores on Wiki corpus. Each line represents a user study group, with emphasis on U_3. The ranking of the study groups based on these scores is similar on ArXiv and PubMed corpora (Appendix Table B2). $C_{\text{UMass},s}$ and $C_{P,s}$ ommited as they are almost identical to their o variant. The scores visualized are from Lim and Lauw (2023b), with similar trends for Proxy Task II and III.

Table 7, the adjusted R^2 of the various analyses is very high at 98%, implying a strong correlation between X and Y with most of the variance explained. The different linear regression tests report very low p-values, confirming that the effect from X is significant, suggesting goodness of fit. For validation, we examine the residual normality and variance homogeneity. We can assume that the residuals are normally distributed, with

Table 6
Mean self-rated proficiency in English for each user study group.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D.)
Mean self-rated proficiency	93.0	76.4	71.6	87.8	85.4	81.2	84.8	92.2	84.1 ± 7.4

Table 7
Simple linear regression analysis ($Y = mX$) against different dependent variables Y with X as the group mean self-rated English proficiency. corr$_i$ is the correlation score (Spearman's ρ) of proxy tasks P_i and C_{NPMI} scores. Values for goodness of fit, residual normality, and homoscedasticity are p-values for the respective tests.

$Y = mX$	Y			
X = self ratings	IRR	corr$_1$	corr$_2$	corr$_3$
Coefficient m	0.00434	0.00743	0.00732	−0.00745
Adjusted R^2	98.6%	98.7%	98.6%	98.7%
Goodness of fit	6.00E-06%	4.59E-06%	5.79E-06%	4.98E-06%
Residual normality	35.6%	56.6%	99.1%	96.0%
Homoscedasticity	41.2%	80.6%	31.3%	54.4%

the Shapiro-Wilk test giving a p-value $>$ 5%. We can similarly assume that the variance is homogeneous, with the White test (White 1980) giving p-value $>$ 5%.

6.2.3 Discussion on the Effects of English Proficiency. The linear regression analysis (Section 6.2.2) shows that group mean self-rated English proficiency (X) has a significant effect on the correlation scores. As X is assumed to have a normal distribution, it suggests that our user study simulates drawing correlation studies from an English-competent population with this effect significant among such a population. Some of these study groups with higher self-rated English proficiency also contain participants who rated their proficiency low. Fortunately, averaging responses may have reduced outlier effects from individual study participants.

6.3 Hyperparameter Search

There are two corpus hyperparameters: window size and minimum frequency. Window size influences the word pair counts as it determines which neighboring words to count as occurring together. Minimum frequency is a count threshold that excludes word pairs with a lower word pair count. Intuitively, window size determines the locality of the context, that is, sentence, paragraph, or document scope. Minimum frequency regulates and excludes rare word pairs. As these hyperparameters may affect the results obtained, it is imperative that we verify their effect.

6.3.1 Analysis. Using our collected responses, we can map human judgment against various automated coherence scores computed on Wiki at different corpus hyperparameters. From Figure 7, for all metrics, correlation scores improve or remain consistent as minimum frequency decreases to 0 and window size decreases to 10. We exclude U_3

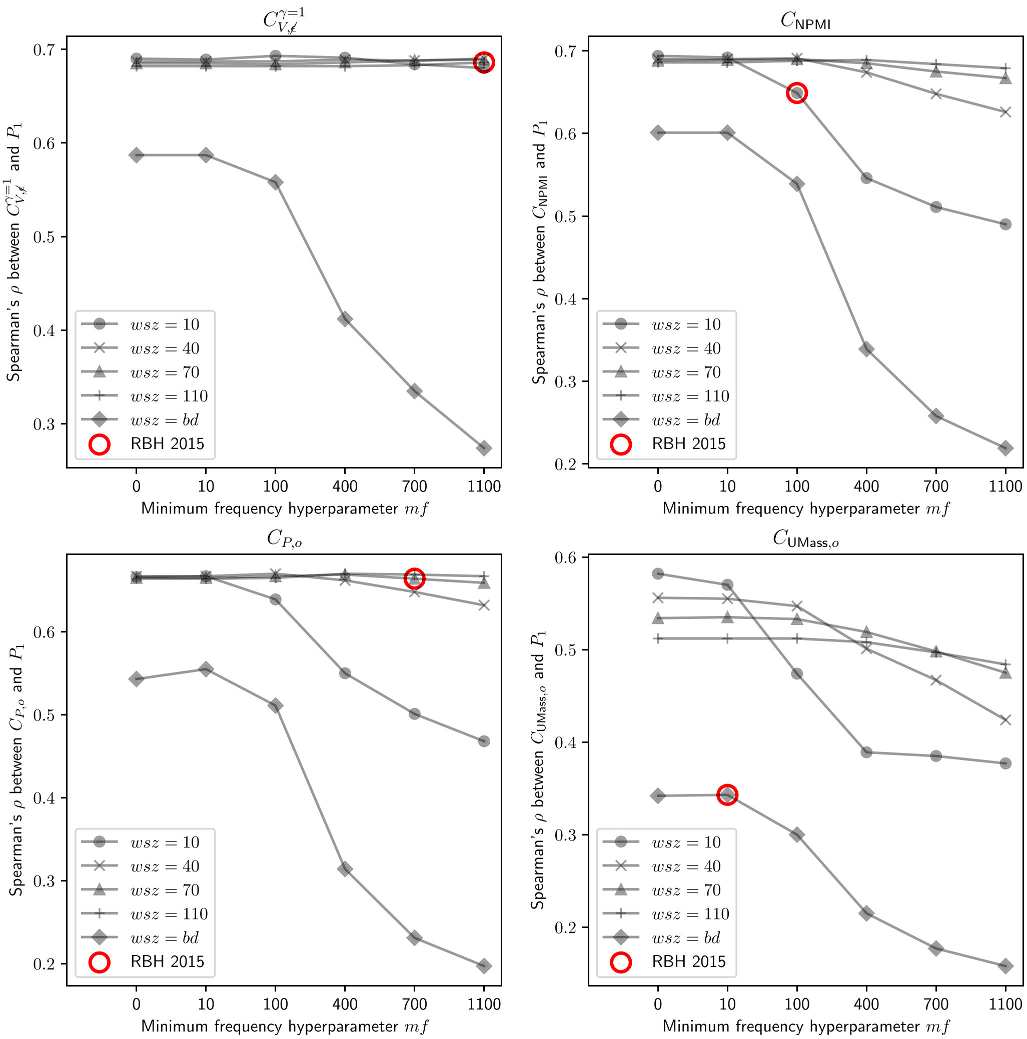


Figure 7
Visual results of average correlation scores (Spearman's ρ) on Proxy Task I, the density of agreement on Wiki-based coherence scores, when corpus hyperparameters are adjusted. Correlation scores on $C_{V,\not\in}^{\gamma=2}$ and $C_{\text{NPMI},\not\in}$ are similar to $C_{V,\not\in}^{\gamma=1}$, and thus omitted. Results on Proxy Task II and III exhibit similar trends as well. Complete Wiki, ArXiv, and PubMed-based results are in Appendix B. These results exclude outlier group U_3. Hyperparameters adjusted are window size, *wsz*, to determine counts and minimum frequency of word occurrences. Boolean document, *bd*, is where we treat the entire document length as the window size. We highlight plots that uses the hyperparameters defined in Röder, Both, and Hinneburg (2015).

in the hyperparameter search and we obtained similar hyperparameter search results in ArXiv and PubMed (see Appendix B).

6.3.2 Further Analysis on Minimum Frequency. When we decrease the minimum frequency, scores of rare-occurring word pairs increase, as we had previously treated them as non-occurring, increasing the overall score of the topic representation. We focus

Table 8
This table compares sets of results between coherence metrics differing in minimum frequency (*mf*) with U_3 excluded. p-values are from the Wilcoxon signed-rank test, testing if scores calculated at a lower minimum frequency $mf = 10$ have a stronger correlation with human judgment than scores calculated at a higher $mf = 100$. Results of PubMed are excluded as it is larger than Wiki. The results with $mf = 10$ is similar to $mf = 0$.

	Proxy Task I			Proxy Task II			Proxy Task II		
ArXiv	$mf = 10$	$mf = 100$	p	$mf = 10$	$mf = 100$	p	$mf = 10$	$mf = 100$	p
$C_{V,\not\in}^{\gamma=1}$	0.444 ± 0.080	0.408 ± 0.085	**2%**	0.440 ± 0.088	0.408 ± 0.092	**2%**	−0.525 ± 0.066	−0.493 ± 0.072	**1%**
$C_{V,\not\in}^{\gamma=2}$	0.423 ± 0.073	0.364 ± 0.069	**4%**	0.416 ± 0.078	0.365 ± 0.067	11%	−0.511 ± 0.066	−0.412 ± 0.069	**1%**
$C_{\text{NPMI},\not\in}$	0.440 ± 0.078	0.406 ± 0.081	**2%**	0.436 ± 0.087	0.407 ± 0.089	**2%**	−0.524 ± 0.064	−0.491 ± 0.068	**1%**
C_{NPMI}	0.434 ± 0.110	0.439 ± 0.076	66%	0.434 ± 0.118	0.439 ± 0.085	66%	−0.507 ± 0.083	−0.513 ± 0.065	66%
$C_{P,s}$	0.432 ± 0.076	0.341 ± 0.069	**2%**	0.424 ± 0.078	0.341 ± 0.067	**2%**	−0.522 ± 0.066	−0.403 ± 0.074	**1%**
$C_{P,o}$	0.432 ± 0.076	0.342 ± 0.069	**2%**	0.424 ± 0.079	0.341 ± 0.067	**2%**	−0.522 ± 0.066	−0.403 ± 0.074	**1%**
$C_{\text{UMass},s}$	0.354 ± 0.099	0.302 ± 0.087	**2%**	0.354 ± 0.099	0.304 ± 0.087	**2%**	−0.392 ± 0.096	−0.315 ± 0.073	**2%**
$C_{\text{UMass},o}$	0.353 ± 0.084	0.298 ± 0.075	**2%**	0.353 ± 0.087	0.300 ± 0.075	**2%**	−0.396 ± 0.076	−0.309 ± 0.066	**1%**
Wiki									
$C_{V,\not\in}^{\gamma=1}$	0.693 ± 0.043	0.695 ± 0.048	71%	0.689 ± 0.035	0.691 ± 0.039	66%	−0.680 ± 0.057	−0.683 ± 0.060	83%
$C_{V,\not\in}^{\gamma=2}$	0.694 ± 0.044	0.649 ± 0.061	**1%**	0.690 ± 0.036	0.644 ± 0.052	**1%**	−0.684 ± 0.057	−0.645 ± 0.074	**2%**
$C_{\text{NPMI},\not\in}$	0.690 ± 0.044	0.693 ± 0.051	71%	0.687 ± 0.036	0.689 ± 0.042	71%	−0.678 ± 0.059	−0.682 ± 0.065	83%
C_{NPMI}	0.689 ± 0.036	0.707 ± 0.041	100%	0.690 ± 0.026	0.707 ± 0.030	100%	−0.682 ± 0.051	−0.701 ± 0.054	100%
$C_{P,s}$	0.667 ± 0.039	0.639 ± 0.055	6%	0.661 ± 0.034	0.632 ± 0.046	6%	−0.657 ± 0.056	−0.633 ± 0.071	**4%**
$C_{P,o}$	0.666 ± 0.039	0.639 ± 0.055	6%	0.661 ± 0.034	0.632 ± 0.046	6%	−0.657 ± 0.056	−0.633 ± 0.071	**4%**
$C_{\text{UMass},s}$	0.566 ± 0.062	0.462 ± 0.074	**1%**	0.559 ± 0.057	0.457 ± 0.064	**1%**	−0.590 ± 0.069	−0.487 ± 0.077	**1%**
$C_{\text{UMass},o}$	0.582 ± 0.044	0.474 ± 0.061	**1%**	0.576 ± 0.036	0.469 ± 0.050	**1%**	−0.609 ± 0.042	−0.499 ± 0.063	**1%**

our observations on ArXiv benchmarks (see Table 8), as its corpus size is smaller than Wiki, with many of the topic representations independently sampled for the user study from Wiki. An increase in correlation scores indicates a likelihood that the previously omitted scores are informational. We see an improvement in correlation scores across most ArXiv-based automated coherence metrics, compared to those based on Wiki. Since the topic representations selected for the user study comprise common vocabulary between the corpora, this may positively influence the informativeness of the observed rare-occurring word pairs. We use the Wilcoxon signed rank test (Wilcoxon 1945) to test the significance of the difference between pairs of correlation scores between different automated coherence metrics and human judgment from the same study group.

6.3.3 Discussion. Even though there might be different optimal settings for the various automated coherence metrics, the difference is marginal and within error bars. The only exception is $C_{\text{UMass,o}}$ where there is a gap between the recommended settings and its optimal settings, possibly explaining its weakness in correlation to human judgment and other automated coherence metrics. The advantage of setting a minimum frequency is to prevent rare word pairs from skewing the scores. However, this omitted information may negatively affect the score. Setting this hyperparameter is about deciding which scenario is likelier. For window sizes, word pair occurrences in close proximity suggest greater relevance. With larger document sizes, having large window sizes might result in associating irrelevant word pairs and penalizes frequently occurring word pairs within the document. For larger corpora, our results demonstrate that it is unnecessary to set a large minimum frequency or window size.

6.4 Further Analysis on Word Ordering

By the definitions of C_P and C_{UMass} (Section 3.1), different ordering of words might produce different scores. Our approach (Section 4.1) samples unordered topic representations, which we then optimize for ordering (Section 4.3). However, our user study presented these topic representations alphabetically, necessitating further investigation.

6.4.1 Analysis. Table 9a and 9b show the respective mean correlation scores of optimally ordered and sorted variants of C_P and C_{UMass} across different corpora and proxy tasks. Excluding U_3, we compare the pairs of related correlation scores across variants, using the Wilcoxon signed rank test, to test whether the optimally ordered variant (subscript o) has a stronger correlation score than the alphabetically sorted variant (subscript s). Most tests report a high p-value, unable to reject the null hypothesis that the s-variant has a stronger or similar correlation score to the o-variant. Even for those tests that are statistically significant, it appears that the difference in magnitude is marginal.

6.4.2 Discussion. From our results, it seems that ordering of words for automated coherence metrics does not significantly affect the correlation with human judgment. Presenting the topic representation alphabetically in our user study is essential from

Table 9
This table compares sets of results between similar corpus-based coherence metrics with U_3 excluded. p-values are from the Wilcoxon signed-rank test. Minimum frequency (mf) in parenthesis, otherwise $mf = 10$. Corpus hyperparameter window size = 10. Bolded p-values are $< 5\%$. The results using $mf = 10$ are similar to $mf = 0$.

(a) Comparison of correlation scores on three Proxy Tasks between Corpus-based coherence $C_{P,s}$ and $C_{P,o}$ with p-value testing if $C_{P,o}$ has a stronger correlation with human judgment than $C_{P,s}$.

	Proxy Task I			Proxy Task II			Proxy Task III		
Corpus (mf)	$C_{P,s}$	$C_{P,o}$	p	$C_{P,s}$	$C_{P,o}$	p	$C_{P,s}$	$C_{P,o}$	p
ArXiv	0.432 ± 0.076	0.432 ± 0.076	84%	0.424 ± 0.078	0.424 ± 0.079	84%	−0.522 ± 0.066	−0.522 ± 0.066	84%
ArXiv-l	0.435 ± 0.074	0.435 ± 0.074	50%	0.428 ± 0.082	0.427 ± 0.082	67%	−0.524 ± 0.054	−0.524 ± 0.054	57%
PubMed	0.473 ± 0.047	0.473 ± 0.047	91%	0.461 ± 0.034	0.461 ± 0.034	84%	−0.500 ± 0.051	−0.500 ± 0.052	92%
PubMed-l	0.518 ± 0.043	0.518 ± 0.043	50%	0.503 ± 0.038	0.503 ± 0.037	50%	−0.553 ± 0.051	−0.553 ± 0.050	28%
Wiki	0.667 ± 0.039	0.666 ± 0.039	92%	0.661 ± 0.034	0.661 ± 0.034	50%	−0.657 ± 0.056	−0.657 ± 0.056	72%
Wiki-l	0.675 ± 0.044	0.675 ± 0.044	9%	0.671 ± 0.040	0.672 ± 0.040	**5%**	−0.664 ± 0.057	−0.665 ± 0.058	**3%**
Wiki (100)	0.639 ± 0.055	0.639 ± 0.055	28%	0.632 ± 0.046	0.632 ± 0.046	28%	−0.633 ± 0.071	−0.633 ± 0.071	8%
Wiki-l (100)	0.672 ± 0.047	0.672 ± 0.047	**2%**	0.666 ± 0.040	0.667 ± 0.041	**4%**	−0.667 ± 0.060	−0.668 ± 0.060	**2%**

(b) Comparison of correlation scores on three Proxy Tasks between Corpus-based coherence $C_{\text{UMass},s}$ and $C_{\text{UMass},o}$ with p-value testing if $C_{\text{UMass},o}$ have a stronger correlation with human judgment than $C_{\text{UMass},s}$.

	Proxy Task I			Proxy Task II			Proxy Task III		
Corpus (mf)	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$	p	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$	p	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$	p
ArXiv	0.354 ± 0.099	0.353 ± 0.084	53%	0.354 ± 0.099	0.353 ± 0.087	66%	−0.392 ± 0.096	−0.396 ± 0.076	38%
ArXiv-l	0.338 ± 0.060	0.368 ± 0.059	**1%**	0.335 ± 0.068	0.367 ± 0.071	**1%**	−0.392 ± 0.072	−0.419 ± 0.060	**2%**
PubMed	0.382 ± 0.081	0.387 ± 0.081	23%	0.367 ± 0.064	0.374 ± 0.065	15%	−0.398 ± 0.085	−0.407 ± 0.086	11%
PubMed-l	0.391 ± 0.099	0.394 ± 0.081	53%	0.372 ± 0.089	0.375 ± 0.072	53%	−0.414 ± 0.090	−0.419 ± 0.075	47%
Wiki	0.566 ± 0.062	0.582 ± 0.044	19%	0.559 ± 0.057	0.576 ± 0.036	19%	−0.590 ± 0.069	−0.609 ± 0.042	11%
Wiki-l	0.574 ± 0.049	0.597 ± 0.040	15%	0.567 ± 0.044	0.591 ± 0.032	11%	−0.602 ± 0.059	−0.626 ± 0.040	15%
Wiki (100)	0.462 ± 0.074	0.474 ± 0.061	11%	0.457 ± 0.064	0.469 ± 0.050	11%	−0.487 ± 0.077	−0.499 ± 0.063	11%
Wiki-l (100)	0.536 ± 0.043	0.559 ± 0.033	6%	0.529 ± 0.041	0.554 ± 0.028	6%	−0.569 ± 0.043	−0.591 ± 0.032	6%

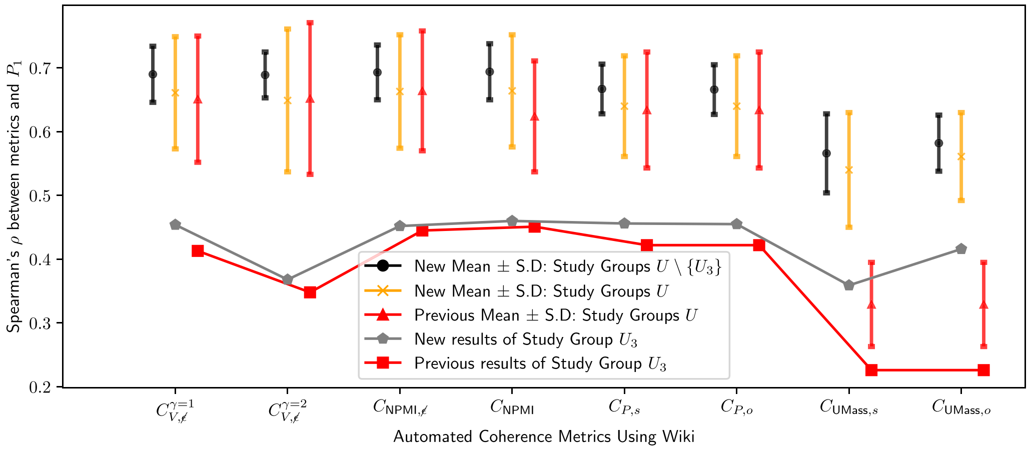


Figure 8
Visual breakdown of the three Proxy Tasks, plotting Spearman's ρ of the density of agreement across Wiki-based coherence scores, with window size 10 and minimum frequency 0. We also plot mean and error bars to account inclusion or exclusion of outlier group U_3. Proxy Task II and III have similar results and thus omitted. Wiki-lemma has similar results. The complete tables for Wiki, ArXiv, and PubMed are in Appendix B.

a user experience perspective, allowing our participants to locate specific words easily when entering their response. The marginal difference might imply that the benefits of ordering only applies to a minority of topic representations.

6.5 Notes on User Study Results

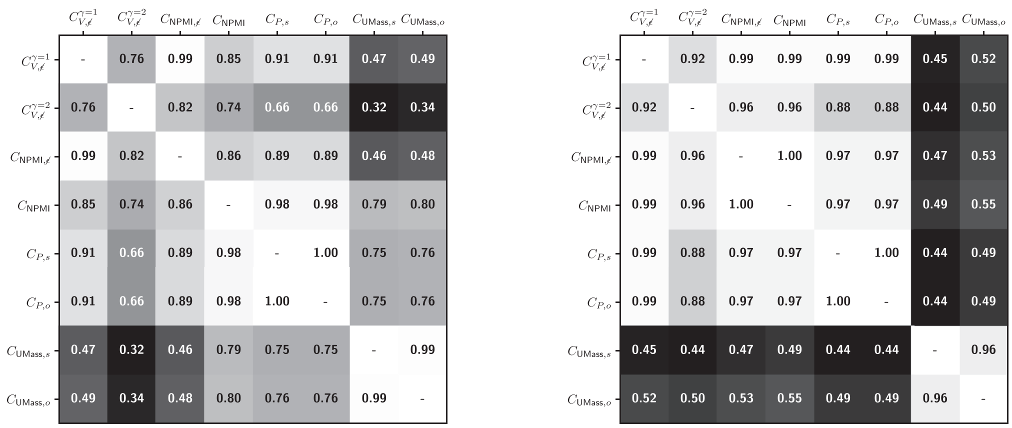
In this section, we identified and analyze four different factors: English proficiency, window size, minimum frequency, and ordering of words. Figure 8 visualizes the individual group correlation scores against the various automated coherence scores on Wiki on the newly selected corpus hyperparameters. Excluding U_3, the mean correlation score increases with a reduction in variation. For future experiments, we standardize corpus hyperparameter settings, unless otherwise specified, to window size $= 10$ and minimum frequency $= 0$, as all of our selected automated coherence metrics seem to work well in those settings. We recommend setting a low minimum frequency of $[0, 100]$ and a window size $= 10$ for large corpora.

7. Automated Metrics Correlation Analysis

Previously, we examined the relation between human and metric. In this section, we examine the behavior of automated coherence metrics, in relation to each other and across corpora.

7.1 Examining Inter-Metric Correlations

In our previous work (Lim and Lauw 2023b), we determined that inter-metric correlations exist between automated coherence metrics, identifying how including or



(a) Correlation scores of metrics measured on Wiki.

(b) Correlation scores of metrics on subsection of data used in Figure 9a where $C_{\text{NPMI}} > 0$.

Figure 9
Heat maps comparing correlations (mean of 5 independently sampled sets of topic representations) between selected Wiki-based coherence metrics with window size = 10 and minimum frequency = 0. Error bars omitted as S.D ≤ 0.02. The results on ArXiv and PubMed are similar.

excluding $\epsilon = 1\text{e}{-12}$ affects these correlations. Considering the effects of hyperparameters, we re-evaluate the correlation (Pearson's r[9]) between different automated metrics measured on Wiki (see Figure 9), PubMed, and ArXiv (see Appendix B). We expect a high positive correlation score between metrics if they are both purportedly measuring for coherence.

7.1.1 Analysis Between Metrics. Our first inter-metric analysis (see Figure 9a) on the better ϵ variant uses the entire sampled set of topic representations. We observe that C_{NPMI} and C_P correlate well with other metrics. However, $\notin$-variant metrics do not correlate strongly with $C_{\text{UMass},o}$. The difference in correlation scores between C_{NPMI} and C_{UMass} and $C_{\text{NPMI},\notin}$ and C_{UMass} suggests that the weaker correlation is due to the removal of ϵ. When ϵ is removed from C_{NPMI}, edges between word pairs with no occurrence will have a C_{NPMI} score of 0. This indifference to non-occurring word pairs decreases the correlation score. This explanation is applicable to $C_{V,\notin}$ as it uses $C_{\text{NPMI},\notin}$. The removal of ϵ is important for C_V, as non-occurring word pairs have similar negative NPMI vectors, resulting in a high C_V score that is contradicting.

7.1.2 Analysis Between Metrics on Coherent Subset. Our second inter-metric analysis (see Figure 9) uses a subset of coherent topic representations with $C_{\text{NPMI}} > 0$. We observe that C_{NPMI} and C_P have similar weaker correlation with C_{UMass} compared to $\notin$-variants. Building on the previous analysis, we can attribute the better correlation scores of C_{NPMI} and C_P to C_{UMass} on incoherent topic representations, suggesting that these metrics agree on incoherence. In both analyses (see Table 9), for position-dependent metrics

[9] Based on reasons provided in Doogan and Buntine (2021), with the main argument that datasets (scores) are continuous and have a bi-variate normal distribution.

C_P and C_{UMass}, we observe that their correlation between their position-optimized (subscript o) and sorted (subscript s) variants have very high correlation. This high correlation suggests that word positioning in the topic representation might be unimportant for a larger corpus (see Section 6.4). The C_V, C_{NPMI}, $C_{\text{NPMI},\notin}$, and C_P of both analyses have a similarly strong correlation, albeit stronger with coherent topics.

7.1.3 Discussion. Comparing our updated analyses to Lim and Lauw (2023b), we see an improvement of inter-metric correlation scores between C_{UMass} and the other metrics on Wiki, ArXiv, and PubMed. We attribute this improvement to the standardization of corpus hyperparameters: window size and minimum frequency. Choosing C_{NPMI}, $C_{V,\notin}$, or C_P for coherence evaluation does not seem to matter as they are highly correlated with each other and have similar correlation scores with human judgment. C_{UMass} has a weaker correlation to human judgment and moderately correlates with other metrics. However, it could be meaningful to use C_{UMass} in conjunction with another metric after accounting for its sensitivity to corpus hyperparameters.

7.2 Examining Inter-Corpus Correlations

A natural extension after inter-metric comparison is to compare metrics measured on different corpora. It is a common expectation that research works would use multiple corpora, with the differences between corpora quantified superficially (such as in Section 3.2). Between corpora, we can quantify their differences at a topical level, using common topics evaluated on automated coherence metrics. If the corpora are thematically similar, we expect a high correlation. Again, using the new corpus hyperparameters, we update our findings.

7.2.1 Inter-Corpus Analysis. We posit that variance in scores measured on different corpora lowers correlation scores due to the missing themes within the shared vocabulary space in either corpus. Using the common topics from the paired corpora, we conduct a correlation analysis on the scores measured on each corpus per metric. Different from Lim and Lauw (2023b), where we only include topics evaluated on both corpus pairs, we seek out all topics present in either corpus that can exist in the other. As such, we increased the number of topic representations evaluated in each corpus pair. Compared with Lim and Lauw (2023b), Table 10 shows a wider range of correlations between each corpus pair on different automated coherence metrics. While these correlation scores are positive, these correlations do not imply identical statistics in various corpora. The control analysis in Table 11a shows a strong correlation score. Compared with the scores found in Table 10, the lower correlation scores confirm the presence of topical differences between the various corpora.

7.2.2 Lemmatization Evaluation Methodology. While we know how lemmatization affects topic modeling (Schofield and Mimno 2016), its effect on evaluation is unclear. We carried out two additional ablations simulating lemmatizing topics for evaluation. We evaluate these topics on pairs of corpora differing in lemmatization while originating from the same documents. These corpus pairs would be thematically identical in knowledge and organization while superficially different in corpus statistics, with the lemmatized vocabulary having larger counts when compared to the unlemmatized counterpart. For the first ablation, we shortlist topics containing at least one unlemmatized word, where if lemmatized, the word exists in the same unlemmatized corpus. We compare correlations of the original set of topic representations and its corresponding

Table 10
Pearson's r between exact automated coherence metric measured on different corpus pairs. We fully sample unique topic representations that may exist in both corpora aggregated from the five different independent samples totaling $|T|$ topics. Corpus hyperparameters are window size = 10 and minimum frequency = 0. In Lim and Lauw (2023b), the shortlisted topics consists of common topic representations evaluated on both corpora. Here, we shortlist common topic representations that are at least present in one but can appear in the other.

Corpus Pairs		$\lvert T\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
ArXiv	PubMed	394K	0.394	0.573	0.562	0.615	0.573	0.574	0.500	0.596
ArXiv	Wiki	477K	0.525	0.625	0.608	0.584	0.612	0.612	0.336	0.464
PubMed	Wiki	468K	0.537	0.592	0.578	0.628	0.601	0.601	0.572	0.631
ArXiv-l	PubMed-l	249K	0.479	0.595	0.575	0.633	0.595	0.596	0.728	0.799
ArXiv-l	Wiki-l	279K	0.521	0.641	0.627	0.599	0.606	0.607	0.499	0.629
PubMed-l	Wiki-l	252K	0.515	0.643	0.632	0.643	0.576	0.576	0.713	0.756

Table 11
Pearson's r (mean from 5 independently sampled sets of size $|\bar{T}|$) of automated coherence metric measured on different scenarios. Each selected topic representation will have two variants, producing two sets of scores for each metric. We compare the correlation of the two sets of scores for the same set of topic representations. Error bars omitted as S.D ≤ 0.01. Corpus hyperparameters are window size = 10 and minimum frequency = 0.

(a) Comparing correlation scores from selected sets of topic representation measured on both lemmatized and unlemmatized corpus.

Corpus	$\lvert\bar{T}\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
ArXiv	80K	0.97	0.98	0.97	0.88	0.94	0.94	0.91	0.92
PubMed	29K	0.97	0.98	0.98	0.95	0.96	0.96	0.95	0.95
Wiki	143K	0.99	0.99	0.99	0.96	0.98	0.98	0.96	0.96

(b) The selected set of unlemmatized topic representations compared to the set of corresponding lemmatized variants, with both variants evaluated on the unlemmatized corpus.

Corpus	$\lvert\bar{T}\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
ArXiv	111K	0.96	0.96	0.96	0.95	0.96	0.96	0.97	0.96
PubMed	60K	0.97	0.98	0.98	0.92	0.94	0.94	0.94	0.94
Wiki	150K	0.98	0.98	0.98	0.97	0.98	0.98	0.97	0.97

(c) The selected set of topic representations, measured on the unlemmatized corpus, is compared with its lemmatized variants, evaluated on the lemmatized corpus.

Corpus	$\lvert\bar{T}\rvert$	$C_{V,\notin}^{\gamma=1}$	$C_{V,\notin}^{\gamma=2}$	$C_{\text{NPMI},\notin}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
ArXiv	126K	0.95	0.96	0.96	0.91	0.93	0.93	0.90	0.92
PubMed	75K	0.95	0.97	0.96	0.84	0.88	0.88	0.79	0.81
Wiki	245K	0.97	0.97	0.97	0.94	0.96	0.96	0.89	0.93

lemmatized set, with their scores measured on the same unlemmatized corpus. In the second ablation, we use the same shortlisting process but with lemmatized topics evaluated on the lemmatized corpus statistics.

7.2.3 Lemmatization Evaluation Analysis. For the first ablation, their scores exhibit a strong correlation (see Table 11b), suggesting that the difference between lemmatized and unlemmatized sets of topics is marginal. For the second ablation, our results (see

Table 11c) show a strong correlation across the various metrics, implying that it is viable to post-process topics for evaluation.

7.2.4 Discussion. When we compare corpora to human judgment, we indirectly quantify the difference in knowledge representation and organization. Similarly, comparing between corpora achieves the same goal. In Section 7.2.1, the low to moderate correlation between corpus pairs using the same automated coherence metrics quantifies the thematic differences between corpora. This result implies that automated coherence metrics calculated on some corpora might not be able to approximate human judgment. To evaluate for coherence, we can choose between conducting a user study or using a different reference corpus, already benchmarked to human judgment. For large corpora, the impact of text lemmatization is minimal, possibly from the law of averages, where common words have similar occurrence frequencies in both lemmatized and unlemmatized corpora. Both lemmatized and unlemmatized corpus statistics report similar correlations of automated coherence to human judgment.

8. Investigating Human Responses

Our experiments suggest that the Wiki corpus can emulate the average human mental model of coherence. We can analogize the human mental model as a corpus, each individual having a unique mental model. Previously, we benchmarked human-corpus and inter-corpus coherence relations. In this section, we seek to investigate human-human coherence relations. Most user studies, including ours, analyze human responses in a group setting, obtaining an average representation of the group. Regrettably, we claim that in doing so, we are removing a defining feature of humans, which is their individuality. The question of *who* is important as different individuals have different knowledge and perceptions of coherence. Our IRR (see Section 6.1) shows that the responses are different but still correlate well with Wiki (Figure 8). Given that it is possible to compare the knowledge representation between corpora (see Section 7.2), we extend the analysis to individual study participants' perception of coherence. For these analyses, we reverse the roles, using Wiki as the benchmark frame of reference and human judgment as the subject of interest.

8.1 Surface-Level Differences

The user study design allows us to investigate our study participants' responses, where we attempt to gain additional insights into their thought processes and how they decide which words within the topic representation are coherent.

8.1.1 Methodology. To get a sense of the difference in responses, we select four quantifiable metrics and visualize their results in Figure 10.

1. **Self-rated English proficiency** (see Table B14, appeared in Figure 5, reappears in Figure B14). Study participants rated their English Proficiency on a slider, ranging from 0 to 100. We use this as a proxy measure for their true English proficiency, and their group-mean self-ratings have a strong correlation with IRR and C_{NPMI} on the three proxy tasks (see Table 7).

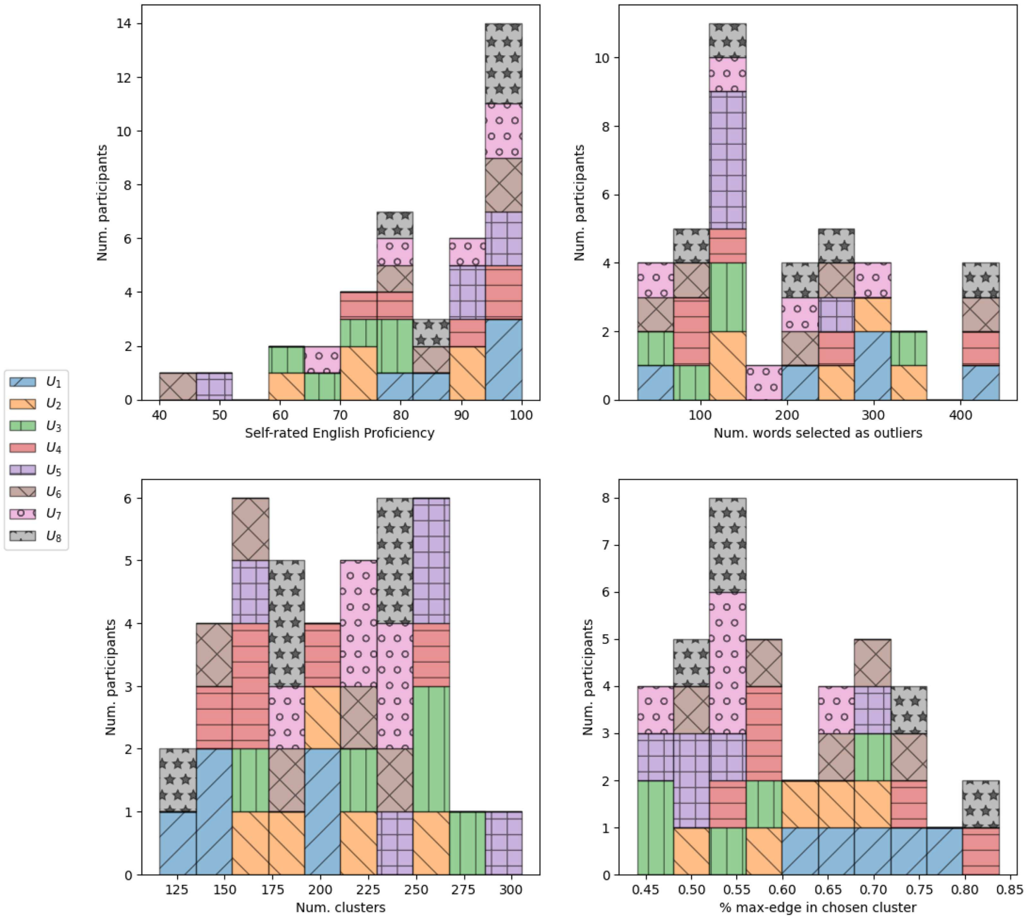


Figure 10
Histogram visualization on selected metrics showing that study participants from the same study group respond differently in each metric. See Appendix Table B14 for the raw data.

2. **Total number of outliers** (see Table B14). The propensity of deciding which words are outliers can indicate a difference in the thought processes of our study participants. The decision of deciding that the word is not coherent with any word groups is influenced by which linguistic content they consume. There are possibly many factors influencing the choice of linguistic content, such as but not limited to education, culture, interests, and so forth.

3. **Total number of coherent groups selected** (see Table B14). Similar to measuring the total number of outliers, this metric is an alternative difference measure. The definition of what is coherent might differ between individuals. Assuming similar outlier counts, some may have stricter criteria for coherence and thus have many smaller coherent groups. Others may have a loose interpretation of coherence, seeking global themes and building fewer but larger coherent groups.

4. **Word-edge score agreement within coherent groups** (see Table B14). We measure the percentage of words in our study participants' chosen coherent groups that share the group with the word's highest-scoring C_{NPMI} paired word, quantifying how frequently our user study participants group frequently occurring word pairs together. Anchoring the statistics on Wiki as a reference point, we can observe the differences between individual study participants.

8.1.2 Analysis and Discussion. From the visualization in Figure 10, within each study group, the overall responses of its individual study participants are very different. Additionally, since the agreement will be affected by the change in window sizes, we focus our examination on two groups, U_1 and U_3 (Figure 11), as the remaining user study groups are similar to U_1. We observe a trend where word-edge score agreement at larger window sizes (70, 110, *bd*) and lower window sizes (10, 40) are very similar. Since the difference is marginal, it suggests that the grouping of best word pairs is consistent. However, from our previous results in our hyperparameter search (see Figure 7), we observe that changing the window size does have a noticeable change in correlation scores, specifically *bd*, where the difference in its scores are widest when compared to other window sizes. As this measurement only accounts for a portion of possible word pair relations, we have to extend the analysis to include word group relations and to account for outliers.

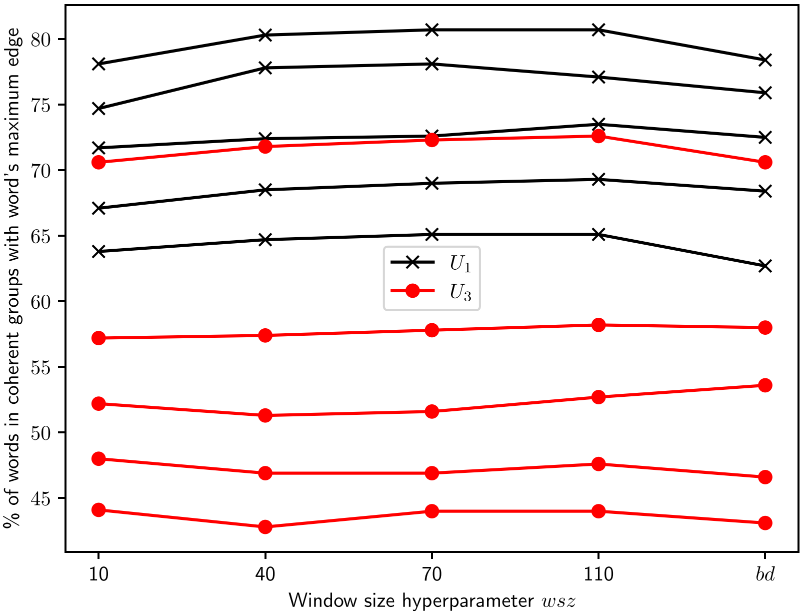


Figure 11
Percentage of words in coherent groups where the word's maximum edge (Wiki C_{NPMI}) is present, examined across corpus hyperparameter window size (wsz). Other unmentioned study groups have similar results to U_1. The Boolean document, *bd*, is where we treat the entire document length as the window size.

8.2 Consistency Within Individual Responses

In our previous work (Lim and Lauw 2023b), we did a correlation analysis on word pair scores. However, measuring individual correlations might not be meaningful due to the binary mode of choice as the user study design enforces tie-breaking. Moreover, the topic representations consist of ten words, which may provide additional context for the study participants to work with. Nevertheless, these word pair scores may provide additional information on how study participants decide which word is an outlier or belongs to a group.

8.2.1 Methodology. In our previous analyses, multiple study participants' input measures one topic. For this analysis, we use multiple topic representations to measure the perception of coherence. First, for each study participant indexed u from study i, we find $\nu_{u,i}^{\max}$ Equation (15).

$$\nu_{u,i}^{\max} = \bigcup_{j=1}^{|T|} \bigcup_{g \in O_{u,j,i}} \{\max \bigcup_{w_k \in t_{j,i}} \{e_{g_0,w_k} : g_0 \neq w_k\}\} \tag{15}$$

$\nu_{u,i}^{\max}$ Equation (15) is the set of maximum word pair edge-weights e_{w_1,w_2} (C_{NPMI}) from the set of outliers $O_{u,i}$ Equation (16).

$$O_{u,i} = \bigcup_{j=1}^{|T|} \{O_{u,j,i}\} \tag{16}$$

A word is an outlier if it belongs to a word set g consisting of only itself in the study participant's response R. $O_{u,j,i}$ is the set of outliers from $R_{u,j,i}$ Equation (17).

$$O_{u,j,i} = \bigcup \{g \in R_{u,j,i} : |g| = 1\} \tag{17}$$

Next, for each study participant indexed u from study i, we find $\nu_{u,i}^{\min}$ Equation (18).

$$\nu_{u,i}^{\min} = \bigcup_{g \in I_{u,i}} \{\min \bigcup_{w_{k_1} \in g, w_{k_2} \in g} \{e_{w_{k_1},w_{k_2}} : w_{k_1} \neq w_{k_2}\}\} \tag{18}$$

$\nu_{u,i}^{\min}$ Equation (18) is the set of minimum word pair edge-weights $e_{w_{k_1},w_{k_2}}$ (C_{NPMI}) from coherent word groups $I_{u,i}$ Equation (19).

$$I_{u,i} = \bigcup_{j=1}^{|T|} \{I_{u,j,i}\} \tag{19}$$

We consider coherent word groups as word sets g in the study participant's response R with membership greater than 1. $I_{u,j,i}$ is the set of coherent word groups from $R_{u,j,i}$ Equation (20).

$$I_{u,j,i} = \bigcup \{g \in R_{u,j,i} : |g| > 1\} \tag{20}$$

Table 12
Individual ambiguity gap results. Mean gap defined as $[\bar{v}^{\min}_{u,i}, \bar{v}^{\max}_{u,i}]$. Mean gap difference is defined as $\bar{v}^{\max}_{u,i} - \bar{v}^{\min}_{u,i}$. Values scored using Wiki C_{NPMI}. Noteworthy groups consists of: U_1 has the most domain expertise related to ArXiv, U_3 is the outlier study group, and U_4 is the polar opposite of U_1, being the most diverse in expertise and experience. See Appendix Table B11 for complete results.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8
Mean Group Gap Difference	0.082	0.088	0.111	0.078	0.095	0.087	0.078	0.084

We chose C_{NPMI} as it correlates well with both ϵ and $\not\epsilon$ metrics (see Section 7). From the sets of selected word pair edges, we derive two means, $\bar{v}^{\min}_{u,i}$ and $\bar{v}^{\max}_{u,i}$. Considering that $\bar{v}^{\max}_{u,i} > \bar{v}^{\min}_{u,i}$, we define mean ambiguity gap $[\bar{v}^{\min}_{u,i}, \bar{v}^{\max}_{u,i}]$ to map out a range of scored word pair edges where there is some level of ambiguity, indicating the degree of disagreement between the individual participant and corpus statistic. For word pair edges that lie outside the gaps, study participants align with Wiki on the relatedness within the word pairs. Below the gap, study participants agree with Wiki that the word pairs are unrelated, while above the gap, the word pairs are coherent.

8.2.2 Analysis. From the results in Table 12, study group U_3 reported the widest group mean ambiguity gap. Intuitively, a wider gap might imply more uncertainty, explaining U_3's lower correlation with automated coherence metrics. However, when we examine other groups, there are also individual study participants with equal or higher differences in the mean gap. We also examined the correlations of these measurements with the self-rated English proficiency of individual study participants and found that $\bar{v}^{\max}_{u,i}$ significantly correlates, averaging $r = 0.55$ across the various window sizes. We also observe that the different study participants have a wide range of scores for $v^{\min}_{u,i}$ (see Appendix Table B11 and Figure 12), even within the same group, suggesting differences in opinions on grouping coherent words.

8.2.3 Discussion. When deciding which words to group, one has to consider many possible combinations as there are 5^{10} possible answers using the Likert matrix format. Because our participants cannot process all possibilities, a realistic approach to the task will consist of entity-to-entity comparisons. These comparisons can be word-to-word, word-to-group, or group-to-group, with group comparisons reducing the task's overall complexity. It is also likely that these comparisons are done linearly and will result in some local optimum, where no further changes will result in a better response where its difference is detectable by the participant. Since we have already established that automated coherence scores correlate to human judgment at topic representations of size 10, the observations should be extendable to groups of smaller sizes. The ambiguity gap can quantify the preferences of each individual study participant with respect to the corpus statistics. Excluding U_3, these other study groups have participants with high and low mean gap differences, evidence that there are outlier opinions in the group. Since these groups correlate better with automated coherence metrics, these outlier opinions might be productive within the group as they introduce a statistical gradation that helps in the coherence evaluation of word pairs.

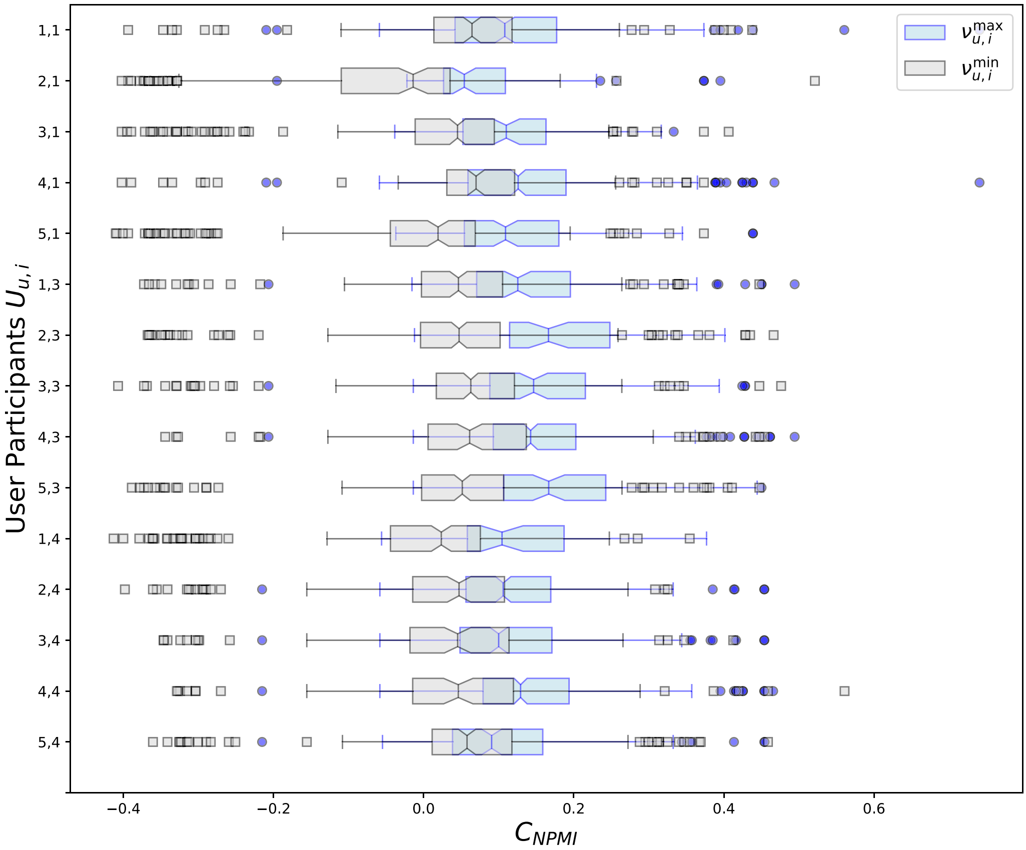


Figure 12
Box and whisker visualization of noteworthy study groups (U_1, U_3, U_4). The box plots in light blue are values from $\nu_{u,i}^{max}$, and the box plots in light gray are values from $\nu_{u,i}^{min}$. The boxes denote the interquartile range with its median notched. This visualization ignores outliers (plotted points). See Appendix C for the complete figure.

8.3 Investigating Local Optimality

We further analyze the results in a word-to-group setting using swaps inspired by the classical 2-opt algorithm (Croes 1958). $\bar{\nu}_{u,i}^{max}$ serves as an estimate for the respective study participant's coherence tipping point, an intuition that helps them decide to include or exclude words. If their responses are consistent, we expect a low percentage of swaps that produce an observable difference from their perspective, suggesting 2-optimality in their response.

8.3.1 Methodology. A swap consists of transferring a word from one group to another. After the swap, we score both groups, using C_{NPMI} for this analysis, and determine whether the sum of the change in scores for both groups is noticeable, where it exceeds a threshold personalized to each study participant. We use $\bar{\nu}_{u,i}^{max}$ as the threshold to decide if the benefit is detectable to study participant u in study group j. For a given pair of groups g_1 and g_2, we conduct swaps from g_1 to g_2, each resulting in $\hat{g_1}$ and $\hat{g_2}$ pair. We examine four different kinds of swaps:

1. **Cluster-to-cluster**. We find coherent group set $I_{u,j,i}$, where $|I_{u,j,i}| \geq 2$, that belongs to the same response to a topic representation $R_{u,j,i}$ for topic representation $t_{j,i}$. We consider all 2-permutations group-wise in $I_{u,j,i}$. We expect $|g_1| + |g_2|$ number of swaps for each group pair. We consider the scenario $(C(\hat{g_1}) + C(\hat{g_2})) - (C(g_1) + C(g_2)) > \bar{v}_{u,i}^{\max}$ as an improvement in finding a better locally optimal solution within a swap. if $|\hat{g_1}| = 1$, $C(\hat{g_1}) = 0$. This swap tests which group is the word better off in.
2. **Outlier-to-cluster**. We find outlier group set $O_{u,j,i}$ belonging to response $R_{u,j,i}$ for topic representation $t_{j,i}$. We select combinations of g_1 from $O_{u,j,i}$ and g_2 from $I_{u,j,i}$. Since $|g_1| = 1$, there is only one swap between each pair. After the swap, $\hat{g_1} = \emptyset$, we set $C(\hat{g_1}) = 0$. Since $|\hat{g_1}| = 1$, its $C(\hat{g_1}) = 0$ as well. Hence, we consider the scenario $C(\hat{g_2}) - C(g_2) > \bar{v}_{u,i}^{\max}$ as an improvement. This swap tests if we can obtain bigger and better coherent topic groups by including outliers.
3. **Cluster-to-outlier**. We conduct swaps of each g_1 from $I_{u,j,i}$ with $g_2 = \emptyset$. For each g_1, there will be $|g_1|$ number of swaps. As $|\hat{g_2}| = 1$, its $C(\hat{g_2}) = 0$. If $|\hat{g_1}| = 0$, its $C(\hat{g_1}) = 0$. We consider the scenario $C(\hat{g_1}) - C(g_1) > \bar{v}_{u,i}^{\max}$ as an improvement. This swap tests if we can obtain a smaller but more coherent topic by excluding words as outliers.
4. **Outlier-to-outlier**. As $|g_1| = 1$ and $|g_2| = 1$, let w_g be the word from g. We conduct swaps between permutations of $O_{u,j,i}$, where $|O_{u,j,i}| \geq 2$, resulting in $|O_{u,j,i}|^2$ number of swaps. As the swap combines two outlier words, the change is equivalent to its edge score $C(\hat{g_2}) > \bar{v}_{u,i}^{\max}$. This swap tests if we can get better 2-sized coherent groups from outliers. Among all the different swaps, this swap is the easiest to show improvement as it only compares the edge.

8.3.2 Analysis. From the results in Figure 13, most swaps do not result in a better overall state. Since $\bar{v}_{u,i}^{\max}$ is the threshold, larger $\bar{v}_{u,i}^{\max}$ are expected to have fewer improving swaps. We benchmark 15 random sampling runs at three different $\bar{v}_{u,i}^{\max}$ of 0.10, 0.15, and 0.20. If we treat the corpus statistics as the oracle, the low rate of better swaps in the random group, serving as baselines, suggests some difficulty in the task. Comparing the study participants to their corresponding random baseline with the next higher $\bar{v}_{u,i}^{\max}$, we find that the majority of the participants have a lower rate of improvement in outlier-to-cluster and outlier-to-outlier swaps. Whereas for cluster-to-cluster and cluster-to-outlier swaps, fewer participants have a lower rate of improvement than the random baseline.

8.3.3 Discussion. We do not expect study participants to have perfect responses as there are many possible swaps. A low rate of improving swaps suggests that most study participants reached some decision local optima on certain kinds of swaps during their interpretation of the topic representations presented. Naturally, since the user study task is on grouping words, a low rate of improvement in outlier-to-cluster and outlier-to-outlier confirms that our participants' primary focus is on assigning words to a group.

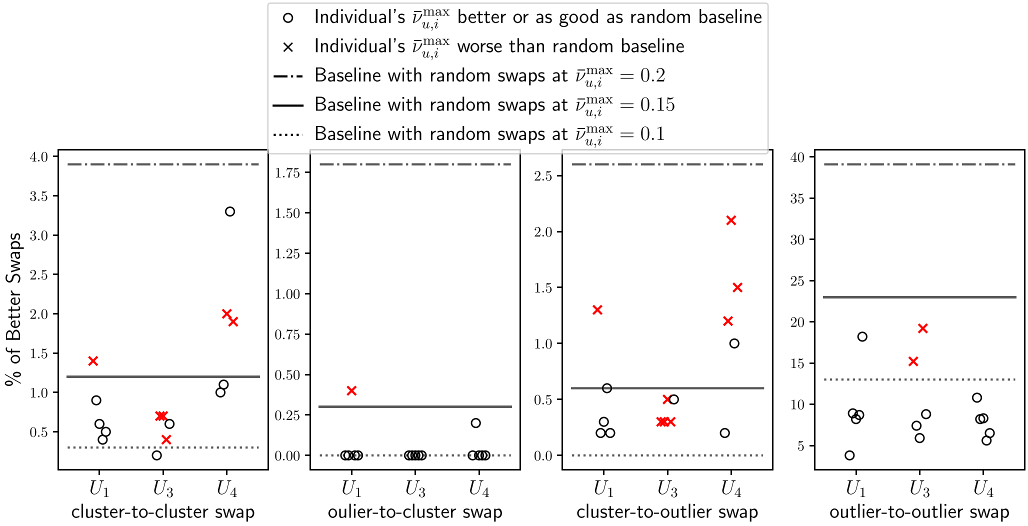


Figure 13
We examine 2-optimality in individual user study responses, where a swap is an action that transfers a word from one group to another. We define four kinds of swaps: cluster-to-cluster, outlier-to-cluster, cluster-to-outlier, and outlier-to-outlier. % better denotes the percentage of swaps that improve the overall system, where the sum of the change in scores for both groups is better than threshold $\bar{v}^{\max}_{u,i}$ tailored to each study participant. For brevity, results of U_1, U_3, and U_4 are shown (see Appendix B for full statistics). Results that are less than the next larger random threshold, suggesting better local optimality for that swap, are marked as circles. Conversely, negative results are marked as crosses.

In contrast, a low rate of improvement in cluster-to-cluster and cluster-to-outlier for some participants either suggests that they compared clusters and improved on their initial response or had coherent initial word clusters. Whereas for other participants, the higher rate on certain swaps suggests that they might not have considered them. The analysis of local optimality shows the difference in answering style, reinforcing the case for investigating individual study participants.

8.4 Ambiguity-Based Scoring

$\llcorner\hat{v}^{\min}\lrcorner$ Equation (21) estimates the population lower bound threshold for coherent word groups, averaging across the lowest individual threshold $\bar{v}^{\min}_{u,i}$ from each study group.

$$\llcorner\hat{v}^{\min}\lrcorner = \frac{1}{|U|}\sum_{i=1}^{|U|} \min \bigcup_{u=1}^{|U_i|} \{\bar{v}^{\min}_{u,i}\} \tag{21}$$

$\ulcorner\hat{v}^{\min}\urcorner$ Equation (22) estimates the population upper bound threshold for coherent word groups, averaging across the highest individual threshold $\bar{v}^{\min}_{u,i}$ from each study group.

$$\ulcorner\hat{v}^{\min}\urcorner = \frac{1}{|U|}\sum_{i=1}^{|U|} \max \bigcup_{u=1}^{|U_i|} \{\bar{v}^{\min}_{u,i}\} \tag{22}$$

$\llcorner\hat{v}^{\max}\lrcorner$ Equation (23) estimates the population lower bound threshold for outliers, averaging across the lowest individual threshold $\bar{v}^{\max}_{u,i}$ from each study group.

$$\llcorner\hat{v}^{\max}\lrcorner = \frac{1}{|U|}\sum_{i=1}^{|U|}\min\bigcup_{u=1}^{|U_i|}\{\bar{v}^{\max}_{u,i}\} \tag{23}$$

$\ulcorner\hat{v}^{\max}\urcorner$ Equation (23) estimates the population upper bound threshold for outliers, averaging across the highest individual threshold $\bar{v}^{\max}_{u,i}$ from each study group.

$$\ulcorner\hat{v}^{\max}\urcorner = \frac{1}{|U|}\sum_{i=1}^{|U|}\max\bigcup_{u=1}^{|U_i|}\{\bar{v}^{\max}_{u,i}\} \tag{24}$$

8.4.1 Discussion. From our user study, $\llcorner\hat{v}^{\min}\lrcorner$, $\ulcorner\hat{v}^{\min}\urcorner$, $\llcorner\hat{v}^{\max}\lrcorner$, and $\ulcorner\hat{v}^{\max}\urcorner$ are 0.005, 0.085, 0.109, and 0.152, respectively. Considering two topic representations with a slight qualitative difference, quantifiable through some metric, we may be unable to discern between the two topics. If we could not perceive the minor difference, then perhaps that difference might not be as meaningful. An ordinal grading system might be able to highlight the qualitative difference, but determining such a system is challenging. An alternative approach might involve capping the extremes. Consider two topics that have a noticeable qualitative difference, with both scoring well above $\ulcorner\hat{v}^{\max}\urcorner$ or well below $\llcorner\hat{v}^{\min}\lrcorner$. In the case of clearly coherent topics, it is unlikely that we will consider the lower-scoring topic as incoherent. To ensure robustness, we can examine multiple different caps and floors. The efficacy of using grades to score topic representations is an area for future investigation.

8.5 Effect of Individual Expertise

An important factor that influences one's decision will be their expertise and experience. Deciding whether to recruit subject matter experts is pivotal and might affect the user study results. User studies on specific domains necessitate recruiting related experts. In this section, we investigate the effect of individual expertise on our user study.

8.5.1 Methodology. It is challenging to judge expertise in Wikipedia as it encompasses a plethora of subjects. However, we can easily classify relevant expertise on domains commonly found in ArXiv. From our study participants, we identify individuals who we believe have experience with these domains. Our shortlisting criteria require these individuals to be graduates with working research experience or pursuing postgraduate studies in relevant scientific domains. We use public LinkedIn information to ascertain their experiences at the point of the study. From our pool of 40 study participants, we identified 12 potential experts. We conduct three different analyses to analyze their impact on our user study. To determine if the experts identified responded differently, we analyze their ambiguity gaps using ArXiv corpus statistics, comparing their results to non-experts'. As different user study groups have different numbers of experts, we identified the expert-dominant group U_1 and compared their responses to other groups using their responses' correlation to ArXiv corpus statistics. Lastly, since there

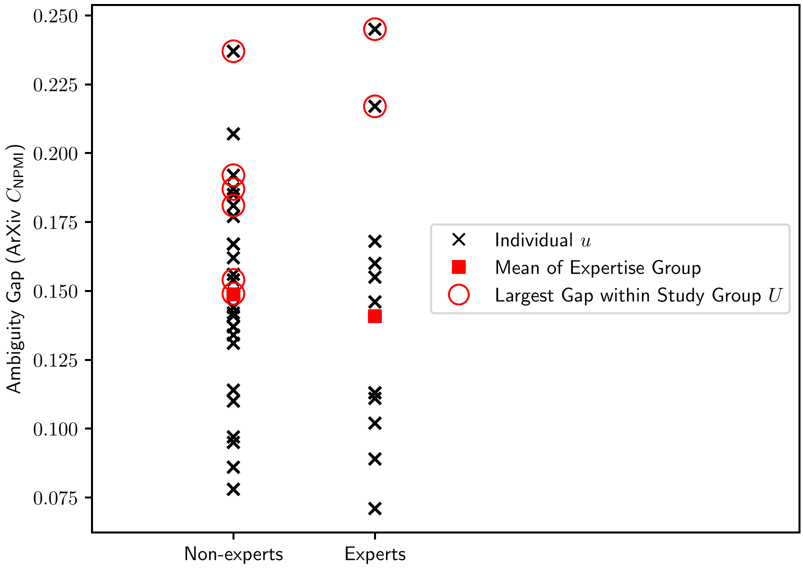


Figure 14
Scatter plot of ambiguity gap (using ArXiv C_{NPMI}) of every study participant u separated by expertise status (see Section 8.5). In each study group U, we circle the plot of the member with the highest ambiguity gap, implying that they have the greatest disagreement with the corpus statistics within their respective U. See Appendix Table B13 for exact results.

are varying levels of expertise even among experts, we examine the experts in U_1 as a case study.

8.5.2 Analysis of Ambiguity Gap. From Figure 14, of the seven study groups with experts, we have two experts having the widest ambiguity gap within their respective study group, implying that they have the greatest disagreement with the ArXiv corpus statistic. The mean ambiguity gap for experts is 0.141, and the mean ambiguity gap for non-experts is 0.149. A Mann-Whitney U test (Mann and Whitney 1947), with a null hypothesis where the ambiguity gap of experts is more than or equal to that of non-experts, returning a p-value of 0.25, and we are unable to reject the null hypothesis.

8.5.3 Analysis Between Study Groups. Ranking user group U_1, a study group with a strong majority of four experts, with other study groups (see Appendix Table B13), reveals that U_1 has the second highest correlation score ($\rho = 0.497$), with U_5 having only one expert, scored higher at ($\rho = 0.530$). However, when we consider U_3 with the weakest score ($\rho = 0.065$), it implies that not every non-expert study groups' competency of ArXiv's domains approaches that of U_1. U_4 serves as another case example, with participants from diverse educational backgrounds, containing undergraduates and graduates, in both ArXiv and non-ArXiv domains. U_4 has the highest correlation score on Wiki corpus statistic ($\rho = 0.75$), while being average on ArXiv corpus statistic ($\rho = 0.388$). U_1, on the other hand, scored similarly well on Wiki corpus ($\rho = 0.72$).

8.5.4 Case Study of Experts Within Study Groups. Even among experts, there are different levels of expertise, which we can approximate using education and career. We use U_1 as a case study (see Appendix Table B13), with u_2 and u_3 having higher levels of expertise than u_1 and u_4. Both u_2 and u_3 have a similar ambiguity gap with u_4 while having a wider ambiguity gap than u_1. This observation indicates that more expertise might not necessarily lead to better agreement with the corpus. Compared to u_5, a non-science major graduate, all experts in U_1 have a lower ambiguity gap.

8.5.5 Discussion on Individual Expertise. From ambiguity gap analysis (8.5.2), within our pool of study participants, it does not seem that expertise leads to agreement with corpus statistics. From inter-group analysis (8.5.3), recruiting experts appears to be the safer option, as recruiting non-experts might result in the outlier scenario of U_3. However, the recruited expert may be opinionated on specific areas, resulting in a large ambiguity gap (case study 8.5.4). We also observe these disagreements between corpora, with ArXiv and PubMed being more specialized than Wiki. Although recruiting randomly might entail some risk, we managed to recruit some non-experts who agreed with the corpus statistics, suggesting that they have comparable contextual knowledge in the subject areas. Overall, recruiting a mix of experts and non-experts worked out for this user study.

8.5.6 Discussion on Recruiting Experts. In a perfect user study, we would have chosen only subject matter experts. However, we realized that the choice lies in our study participants on whether they wish to partake in our user study rather than us selecting them. These experts are paid similarly, at the same rate, as the non-experts, and since they are already well-renumerated in their careers, money is not their primary motivator. As such, we did encounter many rejections, with time cost as the primary reason. For this study, recruiting experts requires more effort, but we were more confident in their quality of response. While recruiting non-experts is easier effort-wise and less time-consuming, we had concerns that they might not have pre-requisite contextual knowledge or English proficiency. In hindsight, creating non-monetary value for our user study participants is an area we could have done better.

9. Final Discussion

We condense our findings and recommendations into three themes.

9.1 Assumptions Challenged

We examine different factors that may affect our evaluation of large corpora. We show that the previous assumptions of window sizes and minimum frequency are not optimal for large corpora (see Section 6.3) , and the selection of lemmatization (see Section 7.2.2) and ordering of words (see Section 6.4) do not seem to affect evaluation. These findings are attributable to the large size of the corpus, resulting in sufficient word co-occurrences count even in a small window size setting. For this user study, the impact of experts is unclear (see Section 8.5), with some non-expert-dominated groups having similar or better correlation than the expert-dominated group. With the increasing ease of access to knowledge, traditional indicators of expertise might underestimate one's knowledge. Alternatively, it could be that only surface-level expertise, and not deep expertise, is required when discerning topic representation from niche areas.

With a better understanding of these factors, we re-examine our previous findings (see Section 7). When evaluating topic representations via large corpora, we recommend using a low window size of [0,100] and low minimum frequency ≤ 10. If the vocabulary size is not an issue, lemmatization is skippable. On the Wikipedia corpus, since most corpus-based metrics are correlated, we prefer using C_{NPMI}. When the evaluations are too close to call, C_{UMass} is a potential tie-breaker. While it is easier to model topics on small corpora, understanding the behavior of these factors may increase the appeal of large corpora.

9.2 Agree to Disagree

One substantial part of this article is to show the differences between individual study participants in their study groups. The reported inter-rater reliability scores quantify the divergence in responses within each study group (see Section 6.1). We also quantified some surface-level statistics showing varied overall responses, with English proficiency having the most significant impact, where groups with better mean proficiency correlate better with automated coherence metrics (see Section 6.2). In a user study group, at least some participants should be proficient in English. Diving deeper to analyze individual differences, we investigate individual responses in two approaches, using corpus statistics as the reference. First, we propose to measure the ambiguity gap that attempts to identify the range within the metric where the participant disagrees with the corpus statistics (see Section 8.2). We show that some participants have a wide ambiguity gap, implying disagreement with the corpora. Second, to decipher the participants' strategy for the task, we investigated the 2-optimality of their responses (see Section 8.3), with results suggesting varying degrees of preferences to the four hypothetical strategies for the user study task. Interpretation of topic representations can be subjective, with individuals having differing opinions. However, despite such disagreement and differences between individual user study participants, the proxy measures employed correlate to the different automated coherence metrics, especially on the Wikipedia corpus.

9.3 The Quest for Interpretability

This work shows that automated coherence metrics are effective, with Wikipedia as the benchmark for evaluating human-level coherence (see Section 7). After selecting appropriate corpus hyperparameters, we find stronger correlation scores. Excluding outlier group U_3, C_{NPMI} and C_V have a mean correlation $\bar{\rho} = 0.69$, C_P at $\bar{\rho} = 0.66$, and C_{UMass} at $\bar{\rho} = 0.58$. A key concern about user studies is their reproducibility. We conduct our user study across several study groups and questions that paint a realistic picture, obtaining replicable study groups with good correlation scores, such as U_4 on C_{NPMI} at $\rho = 0.75$, and outliers, such as U_3 with bad correlation score on C_{NPMI} at $\rho = 0.46$. Of the three corpora examined, Wikipedia's corpus statistics seems to be the best medium to base these corpus-reliant automated coherence metrics on. Wikipedia's size and coverage of diverse subjects might be a reflection, at least in some parts, of our organization of learned concepts.

Evaluating topic representations utilizing large reference corpora is poised to be the new requirement, evidenced by the latest approaches using pre-trained embeddings (Thielmann et al. 2024) and large language model prompts (Stammbach et al. 2023). While these black-box approaches are promising, we believe automated coherence metrics based on corpus statistics will remain relevant in evaluations. Topic representations can serve as a shared interpretation interface between humans and machines, and their evaluation may help us to pursue safer and explainable models.

Ethics Considerations

Before embarking on our user study, we sought and attained prior approval from our organization's institutional review board. We manually reviewed all topic representations in the questions, ensuring the words were not offensive. Our study participants were paid SGD 15 for the study, a rate slightly higher than the hourly rate paid by our organization for undergraduates. We strove to ensure the well-being of our study participants. They were allowed to drop out of the study at any point before completion. We also allowed study participants to complete the tasks at their own pace and place of choosing. Privacy-wise, our questions are neither personal nor sensitive, and it is improbable to de-anonymize our study participants via their responses.

Appendix A. User Study Instructions

Primer on Task

Evaluating the relations between words from a computational lens serves to further the research and understanding of artificial intelligence linguistic research.

A group of words can be considered coherent if they share a similar theme. For example, the group "apples banana coconut durian" can be considered coherent as most people would identify "fruit," "food," or "tree" as the common theme or link.

However, some group of words might be more ambiguous and the common theme might not be as straightforward. For example, "trees ore corn hydrogen" might be considered incoherent to some, while others might identify the common theme as "resources."

Ultimately, it is up to one's personal preferences and experiences to decide on whether a group of words are coherent.

Task Instructions

You will be presented with 10 English words. These words belongs to the 20,000 most frequently used words, so it is unlikely that you will encounter strange words. If you do encounter words that you have never seen before, you are free to use a dictionary or search engine (e.g., Google).

You will then be asked to assign each word to groups, where each group contains words that you think are coherent when grouped together.

Given an example: alcohol athlete breakfast drink eat habit intake meal obesity sleep

Some might divide the words into two groups identifying Group 1 is "alcoholic"-themed and Group 2 is "healthy"-themed.

	Group 1	Group 2	Group 3	Group 4	Not Related
alcohol	O				
athlete		O			
breakfast		O			
drink		O			
eat		O			
habit		O			
intake		O			
meal		O			
obesity	O				
sleep		O			

In another example given: atom calcium component material reduction temperature titanium typical weight yield

Some might group most of the words as "chemistry"-themed.

	Group 1	Group 2	Group 3	Group 4	Not Related
atom	O				
calcium	O				
component	O				
material	O				
reduction	O				
temperature	O				
titanium	O				
typical					O
weight	O				
yield	O				

If you believe that certain word(s) do not belong in any group, select the "Not Related" option in the last column. There can be multiple words that are not related to each other.

For example: animal bed carrot fungible great osmosis paradise star telcommunication water

	Group 1	Group 2	Group 3	Group 4	Not Related
animal					O
bed					O
carrot					O
fungible					O
great					O
osmosis	O				
paradise					O
star					O
telcommunication					O
water	O				

We want to emphasize that there are no right or wrong answers for the tasks, we wish to capture your beliefs on what you think is "correct." We understand that at times, you might encounter words that belong to multiple groups; however to simplify the tasks, we ask that you be the tiebreaker and assign it to the word-group with the strongest similarity.

Appendix B. Supplementary Tables

This appendix contains all the additional tables for various results:

1. Table B1 presents the full user study results for all three proxy tasks across corpora (see Section 6).
2. Table B2 presents the complete breakdown of Proxy Task I by groups across corpora (see Section 6).

3. Table B3 describes overlapping vocabularies between corpora (see Section 3.2).
4. Tables B4, B5, B6 describes mean correlation scores in corpus hyperparameter search for ArXiv and PubMed respectively (see Section 6.3).
5. Tables B7, B8, B9 describes correlation scores for individual user study groups (see Section 6.5).
6. Table B10 describes inter-metric correlation scores in ArXiv and PubMed (see Section 7.2).
7. Table B11 presents the complete analysis results on the ambiguity gap for each study participant (see Section 8.2).
8. Table B12 presents the complete statistics on the 2-optimality for each study participant (see Section 8.3).
9. Table B13 describes data on ambiguity gap for each study participant on arXiv (see Section 8.5.2).
10. Table B14 describes the raw data used in Figure 10.

Table B1
Average Spearman's ρ between automated coherence metrics and respective proxy measure. The values shown are the mean correlation scores from the 8 study groups with error bars. We omit reporting the lemmatized version of the corpus as its values are similar to the original. $C_{\text{UMass},s}$ and $C_{P,s}$ are omitted as they are almost identical to their o variant. Results from Lim and Lauw (2023b).

(a) Proxy Task I: Density of agreement among study participants.

	ArXiv	PubMed	Wiki
$C_{V,\notin}^{\gamma=1}$	0.319 ± 0.152	0.516 ± 0.067	0.651 ± 0.099
$C_{V,\notin}^{\gamma=2}$	0.356 ± 0.146	0.510 ± 0.095	0.652 ± 0.119
$C_{\text{NPMI},\notin}$	0.366 ± 0.136	0.521 ± 0.064	0.664 ± 0.094
C_{NPMI}	0.304 ± 0.169	0.428 ± 0.111	0.624 ± 0.087
$C_{P,o}$	0.266 ± 0.178	0.459 ± 0.093	0.634 ± 0.091
$C_{\text{UMass},o}$	0.243 ± 0.176	0.183 ± 0.161	0.329 ± 0.066

(b) Proxy Task II: Mean of the maximum coherent group between study participants.

	ArXiv	PubMed	Wiki
$C_{V,\notin}^{\gamma=1}$	0.316 ± 0.159	0.511 ± 0.053	0.643 ± 0.110
$C_{V,\notin}^{\gamma=2}$	0.355 ± 0.153	0.507 ± 0.080	0.648 ± 0.130
$C_{\text{NPMI},\notin}$	0.369 ± 0.135	0.517 ± 0.049	0.654 ± 0.104
C_{NPMI}	0.303 ± 0.175	0.421 ± 0.094	0.615 ± 0.090
$C_{P,o}$	0.260 ± 0.182	0.454 ± 0.081	0.624 ± 0.103
$C_{\text{UMass},o}$	0.232 ± 0.182	0.170 ± 0.152	0.320 ± 0.060

(c) Proxy Task III: Mean of coherent group counts between study participants. For this task, a stronger negative score is better as a completely coherent topic gets $P_3(t) = 1$ while an incoherent topic gets $P_3(t) = 10$. Hence, this proxy measure is inversely related to the automated coherence metric, where a larger score indicates coherence.

	ArXiv	PubMed	Wiki
$C_{V,\notin}^{\gamma=1}$	−0.382 ± 0.164	−0.547 ± 0.109	−0.645 ± 0.085
$C_{V,\notin}^{\gamma=2}$	−0.415 ± 0.168	−0.541 ± 0.135	−0.648 ± 0.100
$C_{\text{NPMI},\notin}$	−0.434 ± 0.171	−0.549 ± 0.118	−0.660 ± 0.084
C_{NPMI}	−0.342 ± 0.195	−0.453 ± 0.118	−0.627 ± 0.085
$C_{P,o}$	−0.320 ± 0.200	−0.484 ± 0.107	−0.631 ± 0.082
$C_{\text{UMass},o}$	−0.277 ± 0.172	−0.202 ± 0.126	−0.354 ± 0.053

Table B2
Detailed breakdown of Proxy Task I, values are Spearman's ρ of density of agreement and coherence scores. $C_{\text{UMass},s}$ and $C_{P,s}$ are omitted as they are almost identical to their o variant. These scores are from Lim and Lauw (2023b), with similar trends for Proxy Task II and III.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D.)
ArXiv									
$C_{V,\notin}^{\gamma=1}$	0.464	0.448	−0.021	0.330	0.399	0.437	0.218	0.281	0.319 ± 0.152
$C_{V,\notin}^{\gamma=2}$	0.503	0.469	0.030	0.281	0.459	0.462	0.344	0.300	0.356 ± 0.146
$C_{\text{NPMI},\notin}$	0.475	0.426	0.073	0.392	0.516	0.470	0.304	0.270	0.366 ± 0.136
C_{NPMI}	0.368	0.490	−0.110	0.309	0.386	0.394	0.251	0.348	0.304 ± 0.169
$C_{P,o}$	0.372	0.455	−0.157	0.285	0.355	0.383	0.208	0.231	0.266 ± 0.178
$C_{\text{UMass},o}$	0.348	0.476	−0.162	0.256	0.309	0.261	0.152	0.305	0.243 ± 0.176
PubMed									
$C_{V,\notin}^{\gamma=1}$	0.609	0.560	0.372	0.550	0.462	0.511	0.526	0.535	0.516 ± 0.067
$C_{V,\notin}^{\gamma=2}$	0.662	0.622	0.356	0.465	0.415	0.543	0.492	0.521	0.510 ± 0.095
$C_{\text{NPMI},\notin}$	0.574	0.605	0.396	0.534	0.453	0.498	0.548	0.560	0.521 ± 0.064
C_{NPMI}	0.479	0.447	0.165	0.531	0.442	0.368	0.453	0.537	0.428 ± 0.111
$C_{P,o}$	0.519	0.511	0.231	0.531	0.482	0.409	0.502	0.488	0.459 ± 0.093
$C_{\text{UMass},o}$	0.252	0.177	−0.115	0.327	0.280	0.043	0.087	0.417	0.183 ± 0.161
Wiki									
$C_{V,\notin}^{\gamma=1}$	0.692	0.715	0.413	0.758	0.607	0.670	0.692	0.664	0.651 ± 0.099
$C_{V,\notin}^{\gamma=2}$	0.719	0.739	0.348	0.727	0.631	0.673	0.702	0.678	0.652 ± 0.119
$C_{\text{NPMI},\notin}$	0.737	0.718	0.445	0.760	0.608	0.670	0.706	0.664	0.664 ± 0.094
C_{NPMI}	0.718	0.679	0.451	0.734	0.556	0.582	0.641	0.630	0.624 ± 0.087
$C_{P,o}$	0.658	0.695	0.422	0.737	0.585	0.671	0.684	0.621	0.634 ± 0.091
$C_{\text{UMass},o}$	0.405	0.322	0.226	0.427	0.381	0.272	0.272	0.326	0.329 ± 0.066

Table B3
Quantity of common vocabularies between corpus. Suffix -l. short form for -lemma. Palmetto was re-constructed using 20K most frequent words excluding stop words.

corpus	ArXiv	ArXiv-l.	PubMed	PubMed-l.	Wiki	Wiki-l.
Total	26,620	22,184	38,829	39,997	40003	40,009
ArXiv	–	19,637	13,138	10,527	12,955	10,230
ArXiv-l	19,637	–	9,636	11,015	9,563	10,504
PubMed	13,138	9,636	–	23,328	15,459	12,565
PubMed-l	10,527	11,015	23,328	–	12,637	14,112
Wiki	12,955	9,563	15,459	12,637	–	31,047
Wiki-l	10,230	10,504	12,565	14,112	31,047	–

Table B4
Results of average correlation scores (Spearman's ρ) on Proxy Task I, the density of agreement on Wiki-based coherence scores, when corpus hyperparameters are adjusted. Hyperparameters adjusted are window size, *wsz*, to determine counts and minimum frequency of word occurrences. Boolean document, *bd*, is where we treat the entire document length as the window size. Underlined values are hyperparameters defined in Röder, Both, and Hinneburg (2015). Results on Proxy Task II and III exhibit similar trends as well. These results exclude outlier group U_3.

wsz	minimum frequency					
$C_{V,\notin}^{\gamma=1}$	0	10	100	400	700	1,100
10	0.690 ± 0.044	0.689 ± 0.045	0.693 ± 0.051	0.691 ± 0.059	0.684 ± 0.060	0.680 ± 0.058
40	0.687 ± 0.042	0.687 ± 0.043	0.687 ± 0.043	0.689 ± 0.045	0.688 ± 0.047	0.690 ± 0.049
70	0.685 ± 0.043	0.685 ± 0.042	0.684 ± 0.043	0.686 ± 0.044	0.688 ± 0.045	0.689 ± 0.045
110	0.682 ± 0.042	0.682 ± 0.042	0.682 ± 0.042	0.682 ± 0.042	0.683 ± 0.042	0.686 ± 0.043
bd	0.587 ± 0.063	0.587 ± 0.063	0.558 ± 0.067	0.412 ± 0.078	0.335 ± 0.089	0.274 ± 0.087
$C_{V,\notin}^{\gamma=2}$	0	10	100	400	700	1,100
10	0.689 ± 0.036	0.699 ± 0.038	0.707 ± 0.041	0.706 ± 0.049	0.707 ± 0.052	0.711 ± 0.052
40	0.690 ± 0.034	0.692 ± 0.034	0.697 ± 0.034	0.702 ± 0.037	0.704 ± 0.038	0.704 ± 0.039
70	0.690 ± 0.032	0.690 ± 0.033	0.693 ± 0.032	0.697 ± 0.034	0.699 ± 0.034	0.702 ± 0.036
110	0.687 ± 0.032	0.688 ± 0.033	0.690 ± 0.033	0.692 ± 0.033	0.694 ± 0.034	0.696 ± 0.035
bd	0.613 ± 0.059	0.613 ± 0.059	0.610 ± 0.058	0.561 ± 0.068	0.505 ± 0.083	0.437 ± 0.082
$C_{\text{NPMI},\notin}$	0	10	100	400	700	1,100
10	0.693 ± 0.043	0.691 ± 0.045	0.695 ± 0.048	0.696 ± 0.057	0.690 ± 0.058	0.687 ± 0.059
40	0.689 ± 0.040	0.689 ± 0.040	0.689 ± 0.041	0.692 ± 0.043	0.691 ± 0.044	0.693 ± 0.046
70	0.688 ± 0.041	0.688 ± 0.041	0.688 ± 0.041	0.689 ± 0.042	0.690 ± 0.043	0.691 ± 0.042
110	0.686 ± 0.040	0.686 ± 0.040	0.686 ± 0.040	0.686 ± 0.040	0.686 ± 0.040	0.688 ± 0.042
bd	0.601 ± 0.064	0.601 ± 0.064	0.583 ± 0.064	0.475 ± 0.078	0.391 ± 0.086	0.318 ± 0.079
C_{NPMI}	0	10	100	400	700	1,100
10	0.694 ± 0.044	0.692 ± 0.045	0.649 ± 0.061	0.546 ± 0.073	0.511 ± 0.068	0.490 ± 0.072
40	0.690 ± 0.041	0.690 ± 0.040	0.691 ± 0.042	0.674 ± 0.047	0.648 ± 0.045	0.626 ± 0.058
70	0.688 ± 0.041	0.689 ± 0.041	0.690 ± 0.042	0.685 ± 0.044	0.675 ± 0.047	0.667 ± 0.045
110	0.686 ± 0.040	0.686 ± 0.040	0.688 ± 0.041	0.689 ± 0.042	0.684 ± 0.044	0.679 ± 0.045
bd	0.601 ± 0.064	0.601 ± 0.063	0.539 ± 0.065	0.339 ± 0.084	0.258 ± 0.094	0.219 ± 0.089
$C_{P,o}$	0	10	100	400	700	1,100
10	0.666 ± 0.039	0.667 ± 0.041	0.639 ± 0.055	0.550 ± 0.074	0.501 ± 0.069	0.468 ± 0.071
40	0.667 ± 0.042	0.667 ± 0.042	0.670 ± 0.043	0.662 ± 0.043	0.648 ± 0.042	0.632 ± 0.060
70	0.665 ± 0.042	0.665 ± 0.043	0.667 ± 0.042	0.669 ± 0.045	0.664 ± 0.046	0.659 ± 0.045
110	0.664 ± 0.041	0.664 ± 0.041	0.665 ± 0.042	0.670 ± 0.042	0.669 ± 0.046	0.667 ± 0.046
bd	0.543 ± 0.051	0.555 ± 0.044	0.511 ± 0.067	0.314 ± 0.082	0.231 ± 0.094	0.197 ± 0.087
$C_{\text{UMass},o}$	0	10	100	400	700	1,100
10	0.582 ± 0.044	0.570 ± 0.045	0.474 ± 0.061	0.389 ± 0.085	0.385 ± 0.084	0.377 ± 0.081
40	0.556 ± 0.044	0.555 ± 0.042	0.547 ± 0.043	0.501 ± 0.037	0.467 ± 0.054	0.424 ± 0.071
70	0.534 ± 0.044	0.535 ± 0.044	0.533 ± 0.043	0.519 ± 0.048	0.498 ± 0.030	0.475 ± 0.034
110	0.512 ± 0.045	0.512 ± 0.044	0.512 ± 0.045	0.508 ± 0.042	0.497 ± 0.045	0.484 ± 0.043
bd	0.342 ± 0.056	0.343 ± 0.058	0.300 ± 0.053	0.215 ± 0.091	0.177 ± 0.092	0.158 ± 0.085

Table B5
Results of average correlation scores (Spearman's ρ) on Proxy Task I, the density of agreement on ArXiv-based coherence scores, when corpus hyperparameters are adjusted. Hyperparameters adjusted are window size *wsz* to register counts and minimum frequency of word occurrences. For Boolean document *bd*, we treat the entire document length as the window size. Results on Proxy Task II and III exhibit similar trends as well. These results exclude outlier group U_3.

wsz	minimum frequency					
$C_V^{\gamma=1}$	0	10	100	400	700	1,100
10	0.440 ± 0.078	0.434 ± 0.076	0.406 ± 0.081	0.391 ± 0.071	0.369 ± 0.066	0.352 ± 0.084
40	0.456 ± 0.071	0.451 ± 0.072	0.448 ± 0.068	0.427 ± 0.073	0.410 ± 0.078	0.394 ± 0.067
70	0.457 ± 0.076	0.454 ± 0.075	0.448 ± 0.070	0.419 ± 0.080	0.412 ± 0.086	0.399 ± 0.079
110	0.457 ± 0.077	0.452 ± 0.077	0.429 ± 0.076	0.406 ± 0.090	0.382 ± 0.089	0.368 ± 0.087
bd	0.437 ± 0.080	0.421 ± 0.082	0.368 ± 0.071	0.334 ± 0.084	0.282 ± 0.087	0.241 ± 0.087
$C_V^{\gamma=2}$	0	10	100	400	700	1,100
10	0.434 ± 0.110	0.468 ± 0.095	0.439 ± 0.076	0.409 ± 0.083	0.386 ± 0.060	0.369 ± 0.078
40	0.430 ± 0.103	0.442 ± 0.096	0.465 ± 0.087	0.442 ± 0.080	0.429 ± 0.085	0.414 ± 0.085
70	0.426 ± 0.108	0.445 ± 0.100	0.463 ± 0.089	0.434 ± 0.088	0.430 ± 0.086	0.415 ± 0.088
110	0.415 ± 0.102	0.450 ± 0.096	0.450 ± 0.083	0.423 ± 0.094	0.404 ± 0.095	0.403 ± 0.085
bd	0.439 ± 0.107	0.445 ± 0.100	0.409 ± 0.092	0.365 ± 0.095	0.322 ± 0.080	0.289 ± 0.081
$C_{\text{NPMI},\cancel{o}}$	0	10	100	400	700	1,100
10	0.444 ± 0.080	0.438 ± 0.078	0.408 ± 0.085	0.391 ± 0.071	0.374 ± 0.064	0.352 ± 0.085
40	0.460 ± 0.075	0.457 ± 0.078	0.452 ± 0.074	0.430 ± 0.072	0.415 ± 0.074	0.400 ± 0.072
70	0.463 ± 0.080	0.459 ± 0.078	0.451 ± 0.075	0.424 ± 0.081	0.414 ± 0.085	0.405 ± 0.079
110	0.463 ± 0.080	0.456 ± 0.080	0.436 ± 0.076	0.411 ± 0.091	0.385 ± 0.088	0.377 ± 0.081
bd	0.442 ± 0.084	0.427 ± 0.085	0.376 ± 0.075	0.337 ± 0.088	0.288 ± 0.087	0.247 ± 0.089
C_{NPMI}	0	10	100	400	700	1,100
10	0.423 ± 0.073	0.397 ± 0.082	0.364 ± 0.069	0.320 ± 0.086	0.302 ± 0.077	0.270 ± 0.103
40	0.423 ± 0.080	0.416 ± 0.081	0.366 ± 0.082	0.350 ± 0.076	0.344 ± 0.077	0.331 ± 0.070
70	0.422 ± 0.084	0.416 ± 0.084	0.373 ± 0.086	0.347 ± 0.085	0.327 ± 0.082	0.321 ± 0.088
110	0.417 ± 0.086	0.402 ± 0.089	0.344 ± 0.085	0.318 ± 0.091	0.297 ± 0.090	0.279 ± 0.088
bd	0.412 ± 0.089	0.353 ± 0.090	0.308 ± 0.072	0.178 ± 0.081	0.087 ± 0.070	0.035 ± 0.117
$C_{P,o}$	0	10	100	400	700	1,100
10	0.432 ± 0.076	0.411 ± 0.076	0.342 ± 0.069	0.304 ± 0.079	0.287 ± 0.083	0.282 ± 0.097
40	0.434 ± 0.074	0.432 ± 0.072	0.382 ± 0.078	0.341 ± 0.075	0.331 ± 0.077	0.315 ± 0.074
70	0.431 ± 0.079	0.427 ± 0.077	0.389 ± 0.078	0.337 ± 0.083	0.327 ± 0.082	0.311 ± 0.088
110	0.428 ± 0.077	0.411 ± 0.083	0.357 ± 0.082	0.319 ± 0.091	0.291 ± 0.094	0.272 ± 0.092
bd	0.414 ± 0.092	0.356 ± 0.084	0.301 ± 0.072	0.179 ± 0.093	0.105 ± 0.069	0.052 ± 0.121
$C_{\text{UMass},o}$	0	10	100	400	700	1,100
10	0.353 ± 0.084	0.341 ± 0.085	0.298 ± 0.075	0.275 ± 0.076	0.272 ± 0.075	0.250 ± 0.099
40	0.343 ± 0.089	0.341 ± 0.087	0.317 ± 0.091	0.286 ± 0.082	0.263 ± 0.083	0.257 ± 0.076
70	0.336 ± 0.092	0.332 ± 0.090	0.316 ± 0.096	0.271 ± 0.094	0.257 ± 0.089	0.249 ± 0.089
110	0.328 ± 0.090	0.315 ± 0.084	0.270 ± 0.092	0.248 ± 0.088	0.230 ± 0.089	0.219 ± 0.089
bd	0.335 ± 0.087	0.301 ± 0.092	0.248 ± 0.076	0.101 ± 0.078	0.023 ± 0.086	−0.013 ± 0.132

Table B6
Results of average correlation scores (Spearman's ρ) on Proxy Task I, the density of agreement on PubMed-based coherence scores, when corpus hyperparameters are adjusted. Hyperparameters adjusted are window size *wsz* to register counts and minimum frequency of word occurrences. For Boolean document *bd*, we treat the entire document length as the window size. Results on Proxy Task II and III exhibit similar trends as well. These results exclude outlier group U_3.

wsz	minimum frequency					
$C_V^{\gamma=1}$	0	10	100	400	700	1,100
10	0.535 ± 0.043	0.535 ± 0.043	0.537 ± 0.043	0.532 ± 0.040	0.535 ± 0.046	0.526 ± 0.041
40	0.545 ± 0.046	0.545 ± 0.046	0.544 ± 0.047	0.539 ± 0.047	0.539 ± 0.048	0.541 ± 0.045
70	0.546 ± 0.042	0.546 ± 0.042	0.544 ± 0.044	0.543 ± 0.042	0.542 ± 0.046	0.540 ± 0.044
110	0.536 ± 0.040	0.536 ± 0.040	0.536 ± 0.040	0.535 ± 0.040	0.536 ± 0.043	0.536 ± 0.042
bd	0.467 ± 0.049	0.466 ± 0.050	0.464 ± 0.052	0.459 ± 0.064	0.432 ± 0.068	0.413 ± 0.061
$C_V^{\gamma=2}$	0	10	100	400	700	1,100
10	0.553 ± 0.079	0.553 ± 0.080	0.556 ± 0.081	0.565 ± 0.082	0.558 ± 0.085	0.559 ± 0.082
40	0.534 ± 0.090	0.535 ± 0.090	0.539 ± 0.086	0.550 ± 0.082	0.554 ± 0.082	0.556 ± 0.080
70	0.529 ± 0.084	0.529 ± 0.084	0.530 ± 0.085	0.536 ± 0.085	0.538 ± 0.084	0.542 ± 0.080
110	0.528 ± 0.076	0.528 ± 0.076	0.529 ± 0.076	0.528 ± 0.082	0.531 ± 0.080	0.531 ± 0.080
bd	0.499 ± 0.045	0.499 ± 0.044	0.502 ± 0.046	0.509 ± 0.067	0.501 ± 0.058	0.480 ± 0.062
$C_{\text{NPMI},\not\epsilon}$	0	10	100	400	700	1,100
10	0.537 ± 0.046	0.538 ± 0.047	0.539 ± 0.047	0.534 ± 0.047	0.536 ± 0.051	0.532 ± 0.047
40	0.546 ± 0.056	0.546 ± 0.056	0.544 ± 0.056	0.539 ± 0.056	0.538 ± 0.057	0.543 ± 0.050
70	0.547 ± 0.049	0.547 ± 0.049	0.545 ± 0.050	0.543 ± 0.050	0.542 ± 0.053	0.542 ± 0.051
110	0.541 ± 0.048	0.541 ± 0.048	0.541 ± 0.049	0.541 ± 0.048	0.543 ± 0.048	0.543 ± 0.047
bd	0.480 ± 0.047	0.479 ± 0.048	0.477 ± 0.052	0.470 ± 0.067	0.443 ± 0.067	0.423 ± 0.064
C_{NPMI}	0	10	100	400	700	1,100
10	0.468 ± 0.053	0.469 ± 0.053	0.465 ± 0.054	0.464 ± 0.054	0.448 ± 0.065	0.441 ± 0.088
40	0.469 ± 0.044	0.469 ± 0.044	0.469 ± 0.046	0.470 ± 0.047	0.466 ± 0.048	0.468 ± 0.051
70	0.474 ± 0.038	0.474 ± 0.038	0.474 ± 0.038	0.472 ± 0.036	0.469 ± 0.034	0.469 ± 0.035
110	0.474 ± 0.036	0.474 ± 0.036	0.475 ± 0.036	0.477 ± 0.037	0.473 ± 0.036	0.470 ± 0.033
bd	0.431 ± 0.065	0.431 ± 0.065	0.424 ± 0.068	0.361 ± 0.090	0.325 ± 0.092	0.304 ± 0.098
$C_{P,o}$	0	10	100	400	700	1,100
10	0.473 ± 0.047	0.472 ± 0.047	0.470 ± 0.046	0.457 ± 0.051	0.448 ± 0.056	0.437 ± 0.076
40	0.491 ± 0.044	0.491 ± 0.044	0.491 ± 0.044	0.490 ± 0.043	0.485 ± 0.042	0.483 ± 0.043
70	0.493 ± 0.038	0.493 ± 0.038	0.493 ± 0.038	0.492 ± 0.038	0.492 ± 0.037	0.491 ± 0.036
110	0.490 ± 0.036	0.490 ± 0.036	0.490 ± 0.036	0.491 ± 0.037	0.490 ± 0.037	0.488 ± 0.036
bd	0.428 ± 0.064	0.427 ± 0.064	0.425 ± 0.066	0.376 ± 0.071	0.346 ± 0.080	0.320 ± 0.083
$C_{\text{UMass},o}$	0	10	100	400	700	1,100
10	0.387 ± 0.081	0.387 ± 0.081	0.385 ± 0.080	0.380 ± 0.089	0.382 ± 0.088	0.373 ± 0.104
40	0.368 ± 0.087	0.368 ± 0.087	0.368 ± 0.088	0.365 ± 0.085	0.362 ± 0.083	0.367 ± 0.090
70	0.355 ± 0.075	0.355 ± 0.075	0.355 ± 0.075	0.354 ± 0.078	0.352 ± 0.077	0.352 ± 0.076
110	0.341 ± 0.079	0.341 ± 0.079	0.341 ± 0.079	0.341 ± 0.079	0.340 ± 0.079	0.340 ± 0.077
bd	0.225 ± 0.123	0.226 ± 0.123	0.221 ± 0.121	0.208 ± 0.112	0.212 ± 0.105	0.218 ± 0.098

Table B7
A detailed breakdown of the three Proxy Tasks; values are Spearman's ρ of the density of agreement and Wiki-based coherence scores, with *wsz* = 10 and minimum frequency = 0. We show both means that include and exclude outlier group U_3. Wiki-lemma has similar results.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D.)	
Proxy Task I - density of agreement									Inc. U_3	Ex. U_3
$C_{V,\notin}^{\gamma=1}$	0.718	0.713	0.454	0.750	0.607	0.678	0.711	0.656	0.661 ± 0.088	0.690 ± 0.044
$C_{V,\notin}^{\gamma=2}$	0.727	0.727	0.368	0.696	0.629	0.680	0.719	0.648	0.649 ± 0.112	0.689 ± 0.036
$C_{\text{NPMI},\notin}$	0.723	0.716	0.452	0.749	0.611	0.676	0.713	0.660	0.663 ± 0.089	0.693 ± 0.043
C_{NPMI}	0.721	0.718	0.460	0.752	0.609	0.681	0.712	0.662	0.664 ± 0.088	0.694 ± 0.044
$C_{P,s}$	0.668	0.691	0.456	0.725	0.595	0.662	0.692	0.633	0.640 ± 0.079	0.667 ± 0.039
$C_{P,o}$	0.668	0.691	0.455	0.724	0.595	0.662	0.691	0.633	0.640 ± 0.079	0.666 ± 0.039
$C_{\text{UMass},s}$	0.654	0.579	0.359	0.653	0.494	0.490	0.538	0.553	0.540 ± 0.090	0.566 ± 0.062
$C_{\text{UMass},o}$	0.631	0.599	0.416	0.642	0.583	0.504	0.548	0.568	0.561 ± 0.069	0.582 ± 0.044
Proxy Task II - maximum coherent group size										
$C_{V,\notin}^{\gamma=1}$	0.708	0.698	0.406	0.743	0.624	0.689	0.692	0.652	0.651 ± 0.099	0.687 ± 0.036
$C_{V,\notin}^{\gamma=2}$	0.724	0.711	0.324	0.692	0.652	0.703	0.699	0.650	0.644 ± 0.123	0.690 ± 0.026
$C_{\text{NPMI},\notin}$	0.713	0.700	0.404	0.744	0.628	0.689	0.695	0.655	0.654 ± 0.100	0.689 ± 0.035
C_{NPMI}	0.711	0.702	0.409	0.746	0.623	0.693	0.696	0.659	0.655 ± 0.099	0.690 ± 0.036
$C_{P,s}$	0.655	0.671	0.409	0.722	0.605	0.670	0.671	0.632	0.629 ± 0.089	0.661 ± 0.034
$C_{P,o}$	0.655	0.672	0.408	0.722	0.605	0.670	0.671	0.631	0.629 ± 0.089	0.661 ± 0.034
$C_{\text{UMass},s}$	0.650	0.563	0.339	0.628	0.484	0.498	0.549	0.544	0.532 ± 0.090	0.559 ± 0.057
$C_{\text{UMass},o}$	0.621	0.590	0.393	0.620	0.576	0.513	0.556	0.553	0.553 ± 0.069	0.576 ± 0.036
Proxy Task III - mean coherent group count										
$C_{V,\notin}^{\gamma=1}$	−0.746	−0.661	−0.500	−0.745	−0.565	−0.700	−0.697	−0.635	−0.656 ± 0.081	−0.678 ± 0.059
$C_{V,\notin}^{\gamma=2}$	−0.768	−0.676	−0.428	−0.685	−0.604	−0.703	−0.710	−0.625	−0.650 ± 0.096	−0.682 ± 0.051
$C_{\text{NPMI},\notin}$	−0.750	−0.663	−0.501	−0.740	−0.572	−0.700	−0.697	−0.638	−0.658 ± 0.080	−0.680 ± 0.057
C_{NPMI}	−0.749	−0.666	−0.513	−0.749	−0.575	−0.707	−0.700	−0.644	−0.663 ± 0.078	−0.684 ± 0.057
$C_{P,s}$	−0.700	−0.640	−0.497	−0.730	−0.555	−0.686	−0.683	−0.605	−0.637 ± 0.074	−0.657 ± 0.056
$C_{P,o}$	−0.700	−0.641	−0.498	−0.729	−0.555	−0.686	−0.682	−0.605	−0.637 ± 0.074	−0.657 ± 0.056
$C_{\text{UMass},s}$	−0.668	−0.592	−0.418	−0.703	−0.483	−0.533	−0.577	−0.575	−0.569 ± 0.086	−0.590 ± 0.069
$C_{\text{UMass},o}$	−0.651	−0.620	−0.456	−0.681	−0.572	−0.567	−0.565	−0.607	−0.590 ± 0.064	−0.609 ± 0.042

Table B8
A detailed breakdown of the three Proxy Tasks; values are Spearman's ρ of the density of agreement and ArXiv-based coherence scores, with *wsz* = 10 and minimum frequency = 0. We show both means that include and exclude outlier group U_3. ArXiv-lemma has similar results.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D.)	
Proxy Task I - density of agreement									Inc. U_3	Ex. U_3
$C_V^{\gamma=1}$	0.484	0.429	0.168	0.387	0.584	0.484	0.337	0.374	0.406 ± 0.116	0.440 ± 0.078
$C_V^{\gamma=2}$	0.485	0.412	0.263	0.306	0.638	0.497	0.398	0.299	0.412 ± 0.117	0.434 ± 0.110
$C_{\text{NPMI},\not{\epsilon}}$	0.485	0.436	0.183	0.383	0.599	0.481	0.348	0.374	0.411 ± 0.114	0.444 ± 0.080
C_{NPMI}	0.497	0.432	0.065	0.388	0.530	0.455	0.314	0.346	0.378 ± 0.137	0.423 ± 0.073
$C_{P,s}$	0.496	0.429	0.067	0.420	0.557	0.455	0.321	0.346	0.386 ± 0.140	0.432 ± 0.076
$C_{P,o}$	0.496	0.429	0.067	0.420	0.557	0.455	0.320	0.346	0.386 ± 0.140	0.432 ± 0.076
$C_{\text{UMass},s}$	0.398	0.546	−0.216	0.306	0.293	0.346	0.205	0.382	0.283 ± 0.210	0.354 ± 0.099
$C_{\text{UMass},o}$	0.402	0.505	−0.165	0.302	0.339	0.349	0.208	0.369	0.289 ± 0.189	0.353 ± 0.084
Proxy Task II - maximum coherent group size										
$C_V^{\gamma=1}$	0.526	0.390	0.195	0.368	0.549	0.526	0.322	0.374	0.406 ± 0.114	0.436 ± 0.087
$C_V^{\gamma=2}$	0.528	0.373	0.280	0.294	0.617	0.540	0.392	0.298	0.415 ± 0.121	0.434 ± 0.118
$C_{\text{NPMI},\not{\epsilon}}$	0.526	0.397	0.211	0.363	0.566	0.522	0.331	0.377	0.412 ± 0.112	0.440 ± 0.088
C_{NPMI}	0.531	0.390	0.076	0.356	0.484	0.491	0.313	0.349	0.374 ± 0.134	0.416 ± 0.078
$C_{P,s}$	0.533	0.380	0.086	0.385	0.507	0.493	0.321	0.350	0.382 ± 0.134	0.424 ± 0.078
$C_{P,o}$	0.533	0.380	0.085	0.385	0.507	0.493	0.320	0.350	0.382 ± 0.134	0.424 ± 0.079
$C_{\text{UMass},s}$	0.419	0.536	−0.236	0.275	0.280	0.373	0.216	0.378	0.280 ± 0.216	0.354 ± 0.099
$C_{\text{UMass},o}$	0.414	0.501	−0.170	0.262	0.319	0.376	0.224	0.373	0.287 ± 0.191	0.353 ± 0.087
Proxy Task III - mean coherent group count										
$C_V^{\gamma=1}$	−0.564	−0.605	−0.120	−0.437	−0.600	−0.530	−0.458	−0.474	−0.473 ± 0.146	−0.524 ± 0.064
$C_V^{\gamma=2}$	−0.546	−0.546	−0.227	−0.365	−0.617	−0.545	−0.528	−0.399	−0.472 ± 0.121	−0.507 ± 0.083
$C_{\text{NPMI},\not{\epsilon}}$	−0.562	−0.608	−0.137	−0.434	−0.610	−0.527	−0.466	−0.468	−0.476 ± 0.142	−0.525 ± 0.066
C_{NPMI}	−0.576	−0.583	−0.051	−0.446	−0.595	−0.485	−0.429	−0.461	−0.453 ± 0.164	−0.511 ± 0.066
$C_{P,s}$	−0.578	−0.599	−0.044	−0.476	−0.610	−0.487	−0.440	−0.464	−0.462 ± 0.170	−0.522 ± 0.066
$C_{P,o}$	−0.578	−0.599	−0.045	−0.476	−0.610	−0.487	−0.438	−0.464	−0.462 ± 0.169	−0.522 ± 0.066
$C_{\text{UMass},s}$	−0.447	−0.568	0.174	−0.336	−0.387	−0.318	−0.248	−0.436	−0.321 ± 0.208	−0.392 ± 0.096
$C_{\text{UMass},o}$	−0.443	−0.503	0.128	−0.336	−0.457	−0.323	−0.281	−0.427	−0.330 ± 0.187	−0.396 ± 0.076

Table B9
A detailed breakdown of the three Proxy Tasks; values are Spearman's ρ of the density of agreement and PubMed-based coherence scores, with *wsz* = 10 and minimum frequency = 0. We show both means that include and exclude outlier group U_3. PubMed-lemma has similar results.

Groups	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	Mean (S.D.)	
Proxy Task I - density of agreement									Inc. U_3	Ex. U_3
$C_V^{\gamma=1}$	0.554	0.590	0.400	0.542	0.453	0.493	0.543	0.568	0.518 ± 0.060	0.535 ± 0.043
$C_V^{\gamma=2}$	0.646	0.650	0.398	0.537	0.409	0.514	0.516	0.595	0.533 ± 0.090	0.553 ± 0.079
$C_{\text{NPMI},\not\epsilon}$	0.565	0.600	0.389	0.530	0.453	0.493	0.550	0.569	0.519 ± 0.065	0.537 ± 0.046
C_{NPMI}	0.482	0.448	0.163	0.539	0.447	0.372	0.457	0.535	0.430 ± 0.113	0.468 ± 0.053
$C_{P,s}$	0.472	0.480	0.189	0.524	0.461	0.371	0.487	0.517	0.438 ± 0.104	0.473 ± 0.047
$C_{P,o}$	0.472	0.480	0.189	0.524	0.461	0.370	0.485	0.517	0.437 ± 0.104	0.473 ± 0.047
$C_{\text{UMass},s}$	0.401	0.318	0.029	0.498	0.407	0.264	0.310	0.473	0.338 ± 0.139	0.382 ± 0.081
$C_{\text{UMass},o}$	0.436	0.327	0.023	0.487	0.405	0.262	0.312	0.480	0.342 ± 0.142	0.387 ± 0.081
Proxy Task II - maximum coherent group size										
$C_V^{\gamma=1}$	0.559	0.560	0.433	0.510	0.451	0.511	0.528	0.550	0.513 ± 0.045	0.524 ± 0.035
$C_V^{\gamma=2}$	0.651	0.604	0.432	0.498	0.422	0.530	0.524	0.582	0.530 ± 0.075	0.544 ± 0.070
$C_{\text{NPMI},\not\epsilon}$	0.569	0.566	0.425	0.498	0.454	0.510	0.537	0.553	0.514 ± 0.049	0.527 ± 0.039
C_{NPMI}	0.471	0.447	0.189	0.490	0.457	0.378	0.458	0.496	0.423 ± 0.095	0.457 ± 0.036
$C_{P,s}$	0.463	0.475	0.224	0.477	0.472	0.378	0.483	0.480	0.431 ± 0.085	0.461 ± 0.034
$C_{P,o}$	0.463	0.475	0.224	0.477	0.472	0.378	0.481	0.480	0.431 ± 0.085	0.461 ± 0.034
$C_{\text{UMass},s}$	0.385	0.313	0.050	0.455	0.413	0.265	0.319	0.422	0.328 ± 0.121	0.367 ± 0.064
$C_{\text{UMass},o}$	0.419	0.324	0.042	0.449	0.412	0.262	0.324	0.428	0.333 ± 0.125	0.374 ± 0.065
Proxy Task III - mean coherent group count										
$C_V^{\gamma=1}$	−0.575	−0.689	−0.293	−0.616	−0.461	−0.575	−0.579	−0.573	−0.545 ± 0.112	−0.581 ± 0.063
$C_V^{\gamma=2}$	−0.642	−0.732	−0.280	−0.593	−0.435	−0.553	−0.591	−0.613	−0.555 ± 0.130	−0.594 ± 0.083
$C_{\text{NPMI},\not\epsilon}$	−0.590	−0.703	−0.278	−0.610	−0.473	−0.574	−0.589	−0.568	−0.548 ± 0.118	−0.587 ± 0.063
C_{NPMI}	−0.505	−0.457	−0.175	0.600	−0.422	−0.435	−0.487	−0.556	−0.455 ± 0.120	−0.495 ± 0.060
$C_{P,s}$	−0.490	−0.504	−0.179	−0.594	−0.427	−0.444	−0.509	−0.532	−0.460 ± 0.117	−0.500 ± 0.051
$C_{P,o}$	−0.490	−0.504	−0.179	−0.594	−0.427	−0.443	−0.508	−0.532	−0.460 ± 0.117	−0.500 ± 0.052
$C_{\text{UMass},s}$	−0.444	−0.296	−0.101	−0.527	−0.389	−0.310	−0.326	−0.493	−0.361 ± 0.127	−0.398 ± 0.085
$C_{\text{UMass},o}$	−0.479	−0.310	−0.101	−0.513	−0.394	−0.317	−0.324	−0.510	−0.368 ± 0.129	−0.407 ± 0.086

Table B10
Comparing correlations (mean of 5 independently sampled sets of topic representations) between selected ArXiv and PubMed-based coherence metrics. Error bars omitted as S.D ≤ 0.02.

(a) Correlation scores of metrics measured on ArXiv.

	$C^{\gamma=1}_{V,\not\ell}$	$C^{\gamma=2}_{V,\not\ell}$	$C_{\text{NPMI},\not\ell}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
$C^{\gamma=1}_{V,\not\ell}$	–	0.58	0.99	0.43	0.61	0.61	0.14	0.18
$C^{\gamma=2}_{V,\not\ell}$	0.58	–	0.65	0.47	0.46	0.46	0.17	0.18
$C_{\text{NPMI},\not\ell}$	0.99	0.65	–	0.44	0.60	0.61	0.15	0.18
C_{NPMI}	0.43	0.47	0.44	–	0.97	0.97	0.89	0.90
$C_{P,s}$	0.61	0.46	0.60	0.97	–	1.00	0.82	0.84
$C_{P,o}$	0.61	0.46	0.61	0.97	1.00	–	0.82	0.84
$C_{\text{UMass},s}$	0.14	0.17	0.15	0.89	0.82	0.82	–	0.99
$C_{\text{UMass},o}$	0.18	0.18	0.18	0.90	0.84	0.84	0.99	–

(b) Correlation scores of metrics on subsection of data used in Table B10a where $C_{\text{NPMI}} > 0$.

	$C^{\gamma=1}_{V,\not\ell}$	$C^{\gamma=2}_{V,\not\ell}$	$C_{\text{NPMI},\not\ell}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
$C^{\gamma=1}_{V,\not\ell}$	–	0.89	0.99	0.97	0.98	0.98	0.15	0.22
$C^{\gamma=2}_{V,\not\ell}$	0.89	–	0.93	0.90	0.83	0.83	0.02	0.09
$C_{\text{NPMI},\not\ell}$	0.99	0.93	–	0.98	0.96	0.96	0.18	0.25
C_{NPMI}	0.97	0.90	0.98	–	0.97	0.97	0.25	0.30
$C_{P,s}$	0.98	0.83	0.96	0.97	–	1.00	0.19	0.24
$C_{P,o}$	0.98	0.83	0.96	0.97	1.00	–	0.19	0.24
$C_{\text{UMass},s}$	0.15	0.02	0.18	0.25	0.19	0.19	–	0.97
$C_{\text{UMass},o}$	0.22	0.09	0.25	0.30	0.24	0.24	0.97	–

(c) Correlation scores of metrics measured on PubMed.

	$C^{\gamma=1}_{V,\not\ell}$	$C^{\gamma=2}_{V,\not\ell}$	$C_{\text{NPMI},\not\ell}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
$C^{\gamma=1}_{V,\not\ell}$	–	0.64	0.98	0.50	0.62	0.62	0.18	0.19
$C^{\gamma=2}_{V,\not\ell}$	0.64	–	0.78	0.60	0.56	0.56	0.36	0.37
$C_{\text{NPMI},\not\ell}$	0.98	0.78	–	0.54	0.62	0.62	0.20	0.22
C_{NPMI}	0.50	0.60	0.54	–	0.95	0.95	0.92	0.93
$C_{P,s}$	0.62	0.56	0.62	0.95	–	1.00	0.87	0.87
$C_{P,o}$	0.62	0.56	0.62	0.95	1.00	–	0.87	0.87
$C_{\text{UMass},s}$	0.18	0.36	0.20	0.92	0.87	0.87	–	1.00
$C_{\text{UMass},o}$	0.19	0.37	0.22	0.93	0.87	0.87	1.00	–

(d) Correlation scores of metrics on subsection of data used in Table B10c where $C_{\text{NPMI}} > 0$.

	$C^{\gamma=1}_{V,\not\ell}$	$C^{\gamma=2}_{V,\not\ell}$	$C_{\text{NPMI},\not\ell}$	C_{NPMI}	$C_{P,s}$	$C_{P,o}$	$C_{\text{UMass},s}$	$C_{\text{UMass},o}$
$C^{\gamma=1}_{V,\not\ell}$	–	0.89	0.97	0.94	0.96	0.96	0.47	0.55
$C^{\gamma=2}_{V,\not\ell}$	0.89	–	0.97	0.93	0.78	0.78	0.46	0.53
$C_{\text{NPMI},\not\ell}$	0.97	0.97	–	0.97	0.90	0.90	0.50	0.58
C_{NPMI}	0.94	0.93	0.97	–	0.94	0.94	0.66	0.73
$C_{P,s}$	0.96	0.78	0.90	0.94	–	1.00	0.61	0.67
$C_{P,o}$	0.96	0.78	0.90	0.94	1.00	–	0.61	0.67
$C_{\text{UMass},s}$	0.47	0.46	0.50	0.66	0.61	0.61	–	0.94
$C_{\text{UMass},o}$	0.55	0.53	0.58	0.73	0.67	0.67	0.94	–

Table B11
Individual ambiguity gap results. Quantile gap is defined as $[Q_1(\nu^{\min}_{u,i}), Q_3(\nu^{\max}_{u,i})]$. Mean gap defined as $[\bar{\nu}^{\min}_{u,i}, \bar{\nu}^{\max}_{u,i}]$. Mean gap difference is defined as $\bar{\nu}^{\max}_{u,i} - \bar{\nu}^{\min}_{u,i}$. Values scored using Wiki C_{NPMI}.

User Group		Quantile Gap	Mean Gap		
Group	User	Gap	Gap	Difference	Mean Group Difference
U_1	u_1	0.014–0.159	0.074–0.110	0.036	0.082
	u_2	−0.004–0.194	0.067–0.139	0.072	
	u_3	−0.009–0.171	0.058–0.120	0.062	
	u_4	0.003–0.169	0.056–0.120	0.063	
	u_5	−0.032–0.187	0.011–0.126	0.116	
U_2	u_1	−0.020–0.185	0.046–0.131	0.084	0.088
	u_2	0.018–0.198	0.090–0.138	0.048	
	u_3	0.003–0.160	0.055–0.115	0.060	
	u_4	0.009–0.185	0.063–0.126	0.063	
	u_5	−0.029–0.213	0.023–0.148	0.125	
U_3	u_1	0.000–0.243	0.054–0.187	0.133	0.111
	u_2	0.006–0.203	0.080–0.157	0.076	
	u_3	0.018–0.215	0.066–0.159	0.094	
	u_4	−0.001–0.248	0.053–0.181	0.128	
	u_5	−0.001–0.195	0.052–0.149	0.097	
U_4	u_1	−0.035–0.180	0.005–0.124	0.118	0.078
	u_2	0.031–0.190	0.081–0.139	0.057	
	u_3	−0.003–0.163	0.039–0.116	0.077	
	u_4	−0.071–0.110	−0.031–0.081	0.112	
	u_5	0.014–0.177	0.072–0.127	0.055	
U_5	u_1	0.002–0.209	0.071–0.158	0.087	0.095
	u_2	0.008–0.201	0.081–0.145	0.063	
	u_3	−0.017–0.191	0.046–0.135	0.089	
	u_4	0.011–0.252	0.078–0.163	0.085	
	u_5	−0.032–0.163	0.020–0.112	0.091	
U_6	u_1	0.015–0.164	0.076–0.125	0.049	0.087
	u_2	−0.055–0.199	−0.015–0.154	0.169	
	u_3	−0.018–0.169	0.025–0.125	0.100	
	u_4	0.017–0.174	0.085–0.133	0.049	
	u_5	−0.017–0.150	0.023–0.114	0.091	
U_7	u_1	−0.013–0.198	0.048–0.135	0.088	0.078
	u_2	−0.013–0.154	0.032–0.103	0.072	
	u_3	0.014–0.168	0.079–0.123	0.044	
	u_4	−0.062–0.198	−0.017–0.120	0.137	
	u_5	−0.015–0.146	0.034–0.106	0.071	
U_8	u_1	−0.054–0.160	−0.004–0.118	0.122	0.084
	u_2	−0.048–0.130	−0.007–0.090	0.097	
	u_3	0.030–0.209	0.109–0.151	0.042	
	u_4	−0.030–0.210	0.033–0.143	0.110	
	u_5	−0.002–0.170	0.069–0.125	0.056	

Table B12
We examine 2-optimality in individual user study responses, where a swap is an action that transfers a word from one group to another. We define four kinds of swaps: cluster-to-cluster, outlier-to-cluster, cluster-to-outlier, and outlier-to-outlier. % better denotes the percentage of swaps that improve the overall system, where the sum of the change in scores for both groups is better than the selected threshold. We use $\bar{v}^{\max}_{u,i}$ as the threshold tailored to each study participant.

Type of Swaps:		clus. → clus.		out. → clus.		clus. → out.		out. → out.		Total Swaps	
Group	User	Num.	% better	Num.	% better	Num.	% better	Num.	% better	Num.	% better
U_1	u_1	979	1.4%	302	0.0%	799	1.3%	426	3.8%	2,506	1.6%
	u_2	339	0.9%	245	0.0%	576	0.2%	1,403	8.9%	2,563	5.0%
	u_3	359	0.6%	279	0.4%	722	0.3%	733	8.2%	2,093	3.1%
	u_4	516	0.4%	262	0.0%	702	0.6%	846	8.7%	2,326	3.4%
	u_5	943	0.5%	109	0.0%	948	0.2%	22	18.2%	2,022	0.5%
U_2	u_1	460	1.5%	390	0.5%	690	0.6%	740	8.2%	2,280	3.2%
	u_2	833	1.0%	515	0.0%	642	0.2%	898	9.9%	2,888	3.4%
	u_3	706	1.1%	409	0.2%	735	0.5%	557	6.8%	2,407	2.1%
	u_4	1,390	1.2%	361	0.0%	849	0.7%	134	6.7%	2,734	1.1%
	u_5	1,012	0.2%	265	0.0%	863	0.1%	92	9.8%	2,232	0.5%
U_3	u_1	1,436	0.2%	234	0.0%	900	0.3%	33	15.2%	2,603	0.4%
	u_2	441	0.7%	439	0.0%	660	0.3%	673	7.4%	2,213	2.5%
	u_3	1,651	0.7%	294	0.0%	885	0.5%	85	5.9%	2,915	0.7%
	u_4	1,528	0.4%	145	0.0%	937	0.3%	26	19.2%	2,636	0.5%
	u_5	1,074	0.6%	203	0.0%	873	0.5%	217	8.8%	2,367	1.2%
U_4	u_1	521	1.0%	219	0.0%	850	0.2%	222	10.8%	1,812	1.7%
	u_2	546	1.1%	432	0.2%	572	1.2%	1,208	8.2%	2,758	4.1%
	u_3	1,383	2.0%	222	0.0%	905	2.1%	60	8.3%	2,570	2.0%
	u_4	359	3.3%	92	0.0%	929	1.0%	90	5.6%	1,470	1.8%
	u_5	881	1.9%	378	0.0%	731	1.5%	581	6.5%	2,571	2.6%
U_5	u_1	1,440	0.6%	357	0.0%	853	0.5%	139	6.5%	2,789	0.8%
	u_2	1,088	1.2%	476	0.0%	736	1.4%	453	8.6%	2,753	2.3%
	u_3	1,501	0.9%	170	0.0%	889	0.6%	204	5.4%	2,764	1.0%
	u_4	1,824	0.4%	351	0.0%	875	0.3%	70	7.1%	3,120	0.5%
	u_5	627	1.1%	213	0.0%	850	0.6%	249	3.2%	1,939	1.0%
U_6	u_1	1,018	1.1%	480	0.0%	742	0.5%	436	7.8%	2,676	1.8%
	u_2	804	0.1%	63	0.0%	963	0.0%	4	0.0%	1,834	0.1%
	u_3	1,335	1.0%	241	0.4%	894	1.0%	71	8.5%	2,541	1.2%
	u_4	506	0.6%	428	0.2%	556	0.5%	1,382	10.2%	2,872	5.2%
	u_5	568	1.8%	270	0.0%	782	0.5%	537	5.8%	2,157	2.1%
U_7	u_1	1,219	1.0%	419	0.0%	802	0.2%	246	5.3%	2,686	1.0%
	u_2	964	1.0%	338	0.0%	818	1.5%	256	5.1%	2,376	1.5%
	u_3	876	1.4%	551	0.2%	683	1.3%	668	6.0%	2,778	2.2%
	u_4	851	0.4%	46	0.0%	973	0.4%	8	12.5%	1,878	0.4%
	u_5	1,294	1.2%	271	0.0%	865	1.5%	178	6.7%	2,608	1.6%
U_8	u_1	321	0.0%	180	0.0%	749	0.1%	730	7.7%	1,980	2.9%
	u_2	710	1.4%	140	0.0%	930	1.0%	52	5.8%	1,832	1.2%
	u_3	765	0.5%	549	0.0%	556	0.4%	1,274	8.2%	3,144	3.5%
	u_4	1,324	1.1%	181	0.6%	885	0.8%	168	2.4%	2,558	1.0%
	u_5	1,240	0.9%	378	0.3%	792	0.9%	365	4.1%	2,775	1.2%

Table B13
Ambiguity gap (using ArXiv C_{NPMI}) of each study participant u in study group U (see Section 8.2), and correlation between density of U's response (P_1) to C_{NPMI} (see Section 5.3). Bolded values indicate that we classify u as having some expertise with domains commonly found in ArXiv. Underlined values indicate that u has the largest ambiguity gap within U, implying that they have the greatest disagreement with the corpus statistics.

	u_1	u_2	u_3	u_4	u_5	$Corr(P_1, C_{\text{NPMI}})$
U_1	**0.102**	**0.155**	**0.168**	**0.160**	0.192	0.497
U_2	**0.113**	0.078	**0.113**	0.114	0.154	0.432
U_3	0.167	0.177	0.141	0.181	0.167	0.065
U_4	0.207	0.137	0.185	**0.245**	0.137	0.388
U_5	0.142	0.162	0.156	0.149	**0.217**	0.530
U_6	0.097	0.187	0.149	0.086	**0.089**	0.455
U_7	0.131	0.144	**0.111**	0.237	**0.146**	0.314
U_8	0.149	0.134	0.095	0.110	**0.071**	0.346

Table B14
We select four quantifiable metrics to describe individual study participant's responses. We observe that each study participant responded differently, in some manner, when compared with other study participants within the same study group. Figure 10 visualizes these results.

(a) Self-rated English Proficiency.

	u_1	u_2	u_3	u_4	u_5
U_1	100	80	85	100	100
U_2	60	90	90	70	72
U_3	75	78	80	60	65
U_4	100	80	89	100	70
U_5	50	95	90	92	100
U_6	100	81	100	85	40
U_7	65	95	90	94	80
U_8	96	100	85	80	100

(b) Total number of outliers selected.

	u_1	u_2	u_3	u_4	u_5
U_1	201	424	278	298	52
U_2	310	358	265	151	137
U_3	100	340	115	63	127
U_4	150	428	95	71	269
U_5	147	264	111	125	150
U_6	258	37	106	444	218
U_7	198	182	317	27	135
U_8	251	70	444	115	208

(c) Total number of coherent groups selected.

	u_1	u_2	u_3	u_4	u_5
U_1	208	116	136	148	200
U_2	154	199	185	260	214
U_3	257	154	283	261	215
U_4	159	155	251	138	199
U_5	265	230	256	305	169
U_6	224	183	247	149	162
U_7	244	212	211	187	243
U_8	125	178	187	239	241

(d) Percentage of words in coherent groups where its maximum edge (Wiki C_{NPMI}) is present.

	u_1	u_2	u_3	u_4	u_5
U_1	63.8%	74.7%	78.1%	71.7%	67.1%
U_2	69.0%	57.8%	65.9%	50.8%	61.3%
U_3	52.2%	70.6%	44.1%	48.0%	57.2%
U_4	74.6%	57.5%	52.6%	83.7%	59.5%
U_5	51.1%	52.0%	49.4%	46.5%	71.8%
U_6	57.5%	68.1%	51.9%	65.5%	73.4%
U_7	45.9%	55.3%	54.2%	66.6%	52.7%
U_8	81.0%	74.1%	49.8%	54.1%	54.4%

Appendix C. Supplementary Figure

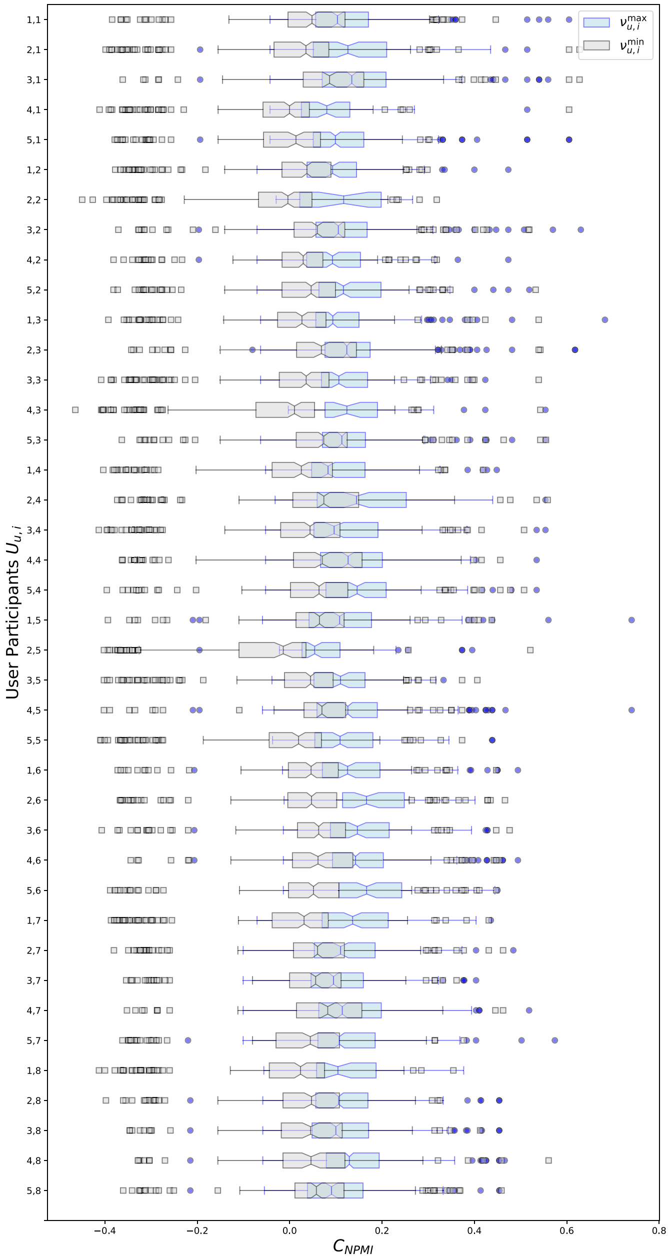


Figure C1
Box and whisker visualization of the distribution reported in Table B11. The box plots in light blue are values from $v^{max}_{u,i}$, and the box plots in light gray are values from $v^{min}_{u,i}$. The boxes denote the interquartile range with its median notched. This visualization ignores outliers (plotted points).

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG award no: AISG3-PhD-2023-08-055T). We extend our gratitude to our user study participants for their efforts, as well as our action editor and our reviewers for their kind feedback and suggestions.

References

Aletras, Nikolaos and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22.

Arakelyan, Erik, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13448–13464. https://doi.org/10.18653/v1/2023.acl-long.752

Attardi, Giusepppe. 2015. Wikiextractor. https://github.com/attardi/wikiextractor

Bar-Noy, Amotz, Reuven Bar-Yehuda, Ari Freund, Joseph (Seffi) Naor, and Baruch Schieber. 2001. A unified approach to approximating resource allocation and scheduling. *Journal of the ACM*, 48(5):1069–1090. https://doi.org/10.1145/502102.502107

Beglar, David. 2010. A Rasch-based validation of the vocabulary size test. *Language Testing*, 27(1):101–118. https://doi.org/10.1177/0265532209340194

Bellman, Richard and Robert Kalaba. 1959. A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, 45(8):1288–1290. https://doi.org/10.1073/pnas.45.8.1288, PubMed: 16590506

Belur, Jyoti, Lisa Tompson, Amy Thornton, and Miranda Simon. 2021. Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research*, 50(2):837–865. https://doi.org/10.1177/0049124118799372

Bianchi, Federico, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683. https://doi.org/10.18653/v1/2021.eacl-main.143

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

Bo, Wenjin Vikki, Mingchen Fu, and Wei Ying Lim. 2023. Revisiting English language proficiency and its impact on the academic performance of domestic university students in Singapore. *Language Testing*, 40(1):133–152. https://doi.org/10.1177/02655322211064629

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2787–2795.

Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*, pages 31–40.

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html

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

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 288–296.

Chiba, Norishige and Takao Nishizeki. 1985. Arboricity and subgraph listing algorithms. *SIAM Journal on Computing*, 14:210–223. https://doi.org/10.1137/0214017

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Croes, G. A. 1958. A method for solving traveling-salesman problems. *Operations Research*, 6(6):791–812. https://doi.org/10.1287/opre.6.6.791

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models.

Danisch, Maximilien, Oana Balalau, and Mauro Sozio. 2018. Listing k-cliques in sparse real-world graphs*. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 589–598, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. https://doi.org/10s.1145/3178876.3186125

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453. https://doi.org/10.1162/tacl_a_00325

Doogan, Caitlin and Wray Buntine. 2021. Topic model or topic twaddle? Re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848. https://doi.org/10.18653/v1/2021.naacl-main.300

Elhage, Nelson, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Fitelson, Branden. 2003. A probabilistic theory of coherence. *Analysis*, 63(3):194–199. https://doi.org/10.1093/analys/63.3.194

Garimella, Aparna, Carmen Banea, and Rada Mihalcea. 2023. Reflection of demographic background on word usage. *Computational Linguistics*, 49(2):373–394. https://doi.org/10.1162/coli_a_00475

Geva, Mor, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45. https://doi.org/10.18653/v1/2022.emnlp-main.3

Griffiths, Thomas and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl 1:5228–35. https://doi.org/10.1073/pnas.0307752101, PubMed: 14872004

Hoffman, Matthew D., David M. Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pages 856–864.

Hoyle, Alexander, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. In *Neural Information Processing Systems*, 35:Art. 155.

Hoyle, Alexander, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344. https://doi.org/10.18653/v1/2022.findings-emnlp.390

Jaccard, P. 1912. The distribution of the flora in the alpine zone 1. *New Phytologist*, 11:37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Ji, Shaoxiong, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514. https://doi.org/10.1109/TNNLS.2021.3070843, PubMed: 33900922

Kingma, Diederik P. and Max Welling. 2014. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.

Krippendorff, K. 2011. Computing krippendorff's alpha-reliability. https://api.semanticscholar.org/CorpusID:59901023

Lau, Jey Han, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365. https://doi.org/10.18653/v1/P17-1033

Lau, Jey Han, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539. https://doi.org/10.3115/v1/E14-1056

Lim, Jia Peng and Hady Lauw. 2023a. Disentangling transformer language models as superposed topic models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8646–8666. https://doi.org/10.18653/v1/2023.emnlp-main.534

Lim, Jia Peng and Hady Lauw. 2023b. Large-scale correlation analysis of automated metrics for topic models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13874–13898. https://doi.org/10.18653/v1/2023.acl-long.776

Mann, H. B. and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60. https://doi.org/10.1214/aoms/1177730491

Meng, Yu, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 3143–3152. https://doi.org/10.1145/3485447.3512034

Meng, Yu, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 1908–1917. https://doi.org/10.1145/3394486.3403242

Miao, Yishu, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736.

Miller, George A. 1995. Wordnet: A lexical database for English. *Communications of ACM*, 38(11):39–41. https://doi.org/10.1145/219717.219748

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011a. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011b. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. https://doi.org/10.18653/v1/S16-1003

Nation, I. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1):59–82. https://doi.org/10.3138/cmlr.63.1.59

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. https://doi.org/10.23915/distill.00024.001

Passonneau, Rebecca. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710. https://doi.org/10.1145/2623330.2623732

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *WSDM*, pages 399–408. https://doi.org/10.1145/2684822.2685324

Rosner, Frank, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. 2014. Evaluating topic coherence measures. *ArXiv preprint arXiv 1403.6397*.

Schofield, Alexandra and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300. https://doi.org/10.1162/tacl_a_00099

Shapiro, S. S. and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shen, Dazhong, Chuan Qin, Chao Wang, Zheng Dong, Hengshu Zhu, and Hui Xiong. 2021. Topic modeling revisited: A document graph-based neural network perspective. In *Advances in Neural Information Processing Systems 34 – 35th Conference on Neural Information Processing Systems, NeurIPS 2021*, pages 14681–14693.

Srivastava, Akash and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR (Poster)*, OpenReview.net.

Stammbach, Dominik, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357. https://doi.org/10.18653/v1/2023.emnlp-main.581

Takahashi, Ayumi. 2009. Self-perception of English ability: Is it related to proficiency and/or class performance? In 新潟大学言語文化研究 (14):39–48.

Thielmann, Anton, Arik Reuter, Quentin Seifert, Elisabeth Bergherr, and Benjamin Säfken. 2024. Topics in the haystack: Enhancing topic quality through corpus expansion. *Computational Linguistics*, pages 1–36. https://doi.org/10.1162/coli_a_00506

Wang, Quan, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

Wang, Wenlin, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177. https://doi.org/10.18653/v1/N19-1015

Wang, Zhengjue, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497. https://doi.org/10.18653/v1/2020.emnlp-main.35

White, Halbert. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838. https://doi.org/10.2307/1912934

Wilcoxon, Frank. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83. https://doi.org/10.2307/3001968

Wu, Shengqiong, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023. Information screening whilst exploiting! Multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751. https://doi.org/10.18653/v1/2023.acl-long.823

Xing, Linzi and Michael J. Paul. 2018. Diagnosing and improving topic models by analyzing posterior variability. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pages 6005–6012. https://doi.org/10.1609/aaai.v32i1.12033

Xu, Chunpu, Jing Li, Piji Li, and Min Yang. 2023. Topic-guided self-introduction generation for social media users. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11387–11402. https://doi.org/10.18653/v1/2023.findings-acl.722

Xu, Ming. 1991. The impact of English-language proficiency on international graduate students' perceived academic difficulty. *Research in Higher Education*, 32(5):557–570. https://doi.org/10.1007/BF00992628

Yang, Liang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph attention topic modeling network. In *Proceedings of the Web Conference 2020*, WWW '20, pages 144–154. https://doi.org/10.1145/3366423.3380102

Yuan, Zhirong, You Peng, Peng Cheng, Li Han, Xuemin Lin, Lei Chen, and Wenjie Zhang. 2022. Efficient *k*-clique listing with set intersection speedup. In *2022 IEEE 38th*

International Conference on Data Engineering (ICDE), pages 1955–1968. `https://doi.org/10.1109/ICDE53745.2022.00192`

Zhang, Ce and Hady W. Lauw. 2020. Topic modeling on document networks with adjacent-encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6737–6745. `https://doi.org/10.1609/aaai.v34i04.6152`

Zhang, Delvin Ce and Hady W. Lauw. 2022. Variational graph author topic modeling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2429–2438. `https://doi.org/10.1145/3534678.3539310`

Zhang, Yu, Yu Meng, Xuan Wang, Sheng Wang, and Jiawei Han. 2022. Seed-guided topic discovery with out-of-vocabulary seeds. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 279–290. `https://doi.org/10.18653/v1/2022.naacl-main.21`

Zhao, He, Lan Du, Wray Buntine, and Gang Liu. 2017. Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 635–644. `https://doi.org/10.1109/ICDM.2017.73`

Zhao, Renbo, Vincent Tan, and Huan Xu. 2017. Online nonnegative matrix factorization with general divergences. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 37–45.

Zheng, Changmeng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. `https://doi.org/10.1109/ICME51207.2021.9428274`