LLM Reading Tea Leaves: Automatically Evaluating Topic Models with Large Language Models

Xiaohao Yang[◇] **He Zhao**[†*] **Dinh Phung**[◇] **Wray Buntine**[‡] **Lan Du**[◇*]
[◇]Monash University, Melbourne, Australia
{xiaohao.yang,dinh.phung,lan.du}@monash.edu
[†]CSIRO's Data61, Sydney, Australia
he.zhao@data61.csiro.au
[‡]VinUniversity, Hanoi, Vietnam
wray.b@vinuni.edu.vn

Abstract

Topic modeling has been a widely used tool for unsupervised text analysis. However, comprehensive evaluations of a topic model remain challenging. Existing evaluation methods are either less comparable across different models (e.g., perplexity) or focus on only one specific aspect of a model (e.g., topic quality or document representation quality) at a time, which is insufficient to reflect the overall model performance. In this paper, we propose WALM (**W**ord **A**greement with **L**anguage **M**odel), a new evaluation method for topic modeling that considers the semantic quality of document representations and topics in a joint manner, leveraging the power of Large Language Models (LLMs). With extensive experiments involving different types of topic models, WALM is shown to align with human judgment and can serve as a complementary evaluation method to the existing ones, bringing a new perspective to topic modeling. Our software package is available at https://github.com/Xiaohao-Yang/Topic_Model_Evaluation.

1 Introduction

Topic modeling (Blei et al., 2003), a popular unsupervised text analysis technique, has been applied to various domains, including information retrieval (Yi and Allan, 2009), marketing analysis (Reisenbichler and Reutterer, 2019), social media analysis (Laureate et al., 2023), bioinformatics (Liu et al., 2016), and more. A topic model typically learns a set of global topics to interpret a text corpus and the topic proportion of a document as its semantic representation.

Although topic models have been time-tested for two decades, as an unsupervised technique, comprehensive evaluations of a topic model remain challenging (Zhao et al., 2021a). Originally, topic models are implemented as probabilistic graphical models such as Latent Dirichlet Allocation (Blei et al., 2003) and many of its Bayesian extensions (e.g., Blei et al., 2010; Paisley et al., 2015; Gan et al., 2015; Zhou et al., 2016; Zhao et al., 2018a, b). For these models, it has been common practice to measure the log-likelihood or perplexity of a model on held-out test documents. While log-likelihood or perplexity provides a straightforward quantitative comparison between models, several issues still persist. Since topic models are not primarily designed to predict words in documents but rather to learn semantically meaningful topics and interpretable document representations, these metrics fail to capture these aspects. Furthermore, estimating the predictive probability is often intractable for Bayesian models, and different papers may employ different sampling or approximation techniques (Wallach et al., 2009; Buntine, 2009). For recently proposed Neural Topic Models (NTMs) (Zhao et al., 2021a), the computation of log-likelihood is even more inconsistent.

In addition to log-likelihood or perplexity, document representation quality and topic quality are evaluated separately. For document representation quality, downstream task performance is typically used as a metric, such as document classification (Yang et al., 2023), clustering (Zhao et al., 2021a), and retrieval (Larochelle and Lauly, 2012). For topic quality, the ultimate evaluation method is human evaluation, which is time-consuming and expensive. Thus, various automatic metrics have been proposed, such as topic coherence (Lau et al., 2014), which measures how semantically coherent the representative words in a topic are, and topic diversity (Dieng et al., 2020), which measures

*Corresponding authors: Lan Du, He Zhao.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 357–375, 2025. https://doi.org/10.1162/tacl_a_00744
Action Editor: Michael Elhadad. Submission batch: 1/2025; Revision batch: 11/2024; Published 4/2025.

how diverse discovered topics are. To comprehensively evaluate the performance of a topic model, one needs to report multiple metrics on both document representation and topic qualities. However, these metrics can be contradictory, e.g., a topic model with good topic quality may not preserve good quality on document representation, and vice versa. This discrepancy complicates the model selection process for topic models in practice.

In this paper, we aim to develop a new evaluation approach for topic modeling that considers both the semantic quality of document representations and topics in a joint manner, leveraging the power of Large Language Models (LLMs). Our key idea is as follows: After being trained, a topic model can infer a document's distribution over topics and each topic is a distribution over vocabulary words. With these two distributions, a model can generate a set of ''topical'' words given a document, such as by looking at its representative topics and the representative words of each topic. The generation of the topical words takes both the topic distribution of a document and the word distributions of the topics into account, which captures the semantic summary of the document and is expected to align with the keywords identified by humans. Given the high cost of human evaluation, we propose using LLMs as a proxy by employing appropriate prompts to generate keywords for the document, which are then compared with the topical words produced by a topic model. Finally, to quantify the agreement between the words from the topic model and the LLM, a series of WALM (**W**ord **A**greement with **L**anguage **M**odel) metrics are proposed. WALM has the following appealing properties:

- It is a joint metric that evaluates the quality of both document representations and topics.
- It assesses how effectively a topic model captures the semantics of a document, which is a core objective of topic modeling.
- It allows for comparisons across various types of topic models.

To examine WALM series metrics, we conduct extensive experiments using various popular topic models on different datasets, comparing them with other widely used topic model evaluation metrics. Moreover, human evaluation is also conducted to demonstrate the alignment of WALM with human judgment.

2 Related Work

As an unsupervised technique for uncovering hidden themes in text, evaluating topic models remains challenging. Early evaluations of a topic model rely on the log-likelihood or perplexity of held-out documents (Blei et al., 2003), which measures how well the model predicts the words of documents. As the computation of predictive probability is often intractable for conventional Bayesian topic models, various sampling or approximation techniques have been proposed (Wallach et al., 2009; Buntine, 2009). Apart from the inconsistent estimation, held-out likelihood is regarded as not correlated with the interpretability of topics from a human perspective (Chang et al., 2009), prompting the direct evaluation of topics and document representation quality.

As for the evaluation of topics, Chang et al. (2009) design the word and topic intrusion tasks for human annotators, where high-quality topics or document representations are those where annotators can easily identify the intruders. Newman et al. (2010) and Mimno et al. (2011) evaluate topic coherence by direct ratings from human experts. Although human judgment is commonly regarded as the gold standard, it is expensive and impractical for large-scale evaluation. Automated evaluation of topic coherence is more practical, such as Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014), which relies on the co-occurrence of the topic's top words in the reference corpus to measure topic coherence, with the underlying assumption that a large reference corpus such as Wikipedia can capture prevalent language patterns. Although they automate the evaluation of topics and strongly correlate with human judgment (Newman et al., 2010), counting word co-occurrence in a large reference corpus is still relatively expensive. Moreover, coherence metrics can vary depending on the reference corpus, and there is no single ''right'' reference corpus that is suitable for all datasets (Doogan and Buntine, 2021). Recent works propose leveraging word embeddings (Nikolenko, 2016) or contextualized embeddings (Hoover et al., 2021) for efficiently evaluating topic coherence, incorporating semantics from pre-trained embeddings. Due

to common posterior collapse issues (Lucas et al., 2019) in the growing field of neural topic models (Zhao et al., 2021a), recent works also consider topic diversity (Dieng et al., 2020) during evaluation, which measures how distinct the top words of each topic are.

As for the evaluation of document representation, early works focus on how well the topic proportion of a document represents the document content, assessed through a topic intrusion task by human annotators (Chang et al., 2009), which is further extended as automated metrics (Bhatia et al., 2017, 2018). Recent topic models often use the topic proportions as document representations, the quality of which is commonly investigated through downstream tasks, including their use as features for document classification (Nguyen and Luu, 2021), clustering (Zhao et al., 2021b), and retrieval (Larochelle and Lauly, 2012). Recently, the generalization ability of topic models is investigated by evaluating their quality of document representations across different unseen corpora (Yang et al., 2023).

In the era of LLMs (Brown et al., 2020; Thoppilan et al., 2022; Touvron et al., 2023a,b; Chowdhery et al., 2024), recent research has begun leveraging LLMs to evaluate topic models, such as using ChatGPT[1] as a proxy for human annotators for word intrusion and topic rating tasks for evaluating topic coherence (Stammbach et al., 2023; Rahimi et al., 2024). The focus of these works is still on topic quality only.

In this work, we propose new evaluation metrics for topic models, differing from previous works in the following ways: (1) Unlike evaluations that focus on only sub-components of a topic model (i.e., topics or document representations), our evaluation metrics offer a joint approach to topic model evaluation, considering both topics and document representations together. (2) Compared with log-likelihood or perplexity, which also evaluate based on documents, our evaluation metrics consider semantics from documents and align with the focus of topic modeling. (3) Different from recent LLM-based evaluations that use LLMs for topic quality evaluation, ours considers both topic quality and document representation quality and our use of LLMs is quite different from previous works.

[1] https://openai.com/index/chatgpt/.

3 Background

Given a document collection $\mathcal{D} := \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$ with V vocabulary words, a topic model is typically trained on their Bag-of-Words (BOWs), e.g., $\boldsymbol{x} \in \mathbb{N}^V$. The topic model can infer a distribution over K topics for each document by running its inference process:

$$\boldsymbol{z} := f_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters of the inference process; $\boldsymbol{z} \in \Delta^K$ (Δ denotes the probability simplex) indicates the proportion of each topic present in the document and is commonly used as its semantic representation. Additionally, the topic model also discovers K global topics for the corpus (i.e., $\mathcal{T} := \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K\}$), where each topic $\boldsymbol{t} \in \Delta^V$ is a distribution over V vocabularies. Ideally, each topic captures a semantic concept that can be interpreted by its top-weighted words. To train a topic model, one often needs to generate or reconstruct the word distribution of the document from $\boldsymbol{z}$ by running its generative process:

$$\boldsymbol{w} := f_{\boldsymbol{\phi}}(\boldsymbol{z}, \mathcal{T}), \tag{2}$$

where $\boldsymbol{\phi}$ are the model parameters of the generative process; $\boldsymbol{w} \in \Delta^V$ is the per-document word distribution from which $\boldsymbol{x}$ is sampled. Let $\mathcal{Z} := \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ be the semantic representations of N test documents and $\mathcal{T}$ be the K learned topics, current evaluation of a topic model is commonly conducted based on either $\mathcal{Z}$ or $\mathcal{T}$ separately.

4 Method

4.1 Motivation

Both topics and document representations are important components of a topic model. To comprehensively evaluate a topic model, it is common practice to report the performance of both parts. This can be done by measuring topic quality using metrics such as NPMI and assessing document representation quality through downstream classification accuracy (ACC) (see section 5.1 for details of metrics calculation). However, a model that prioritizes topic quality (e.g., NPMI) may not perform well in terms of document representations (e.g., ACC), and vice versa, which creates difficulty during model selection, as illustrated in

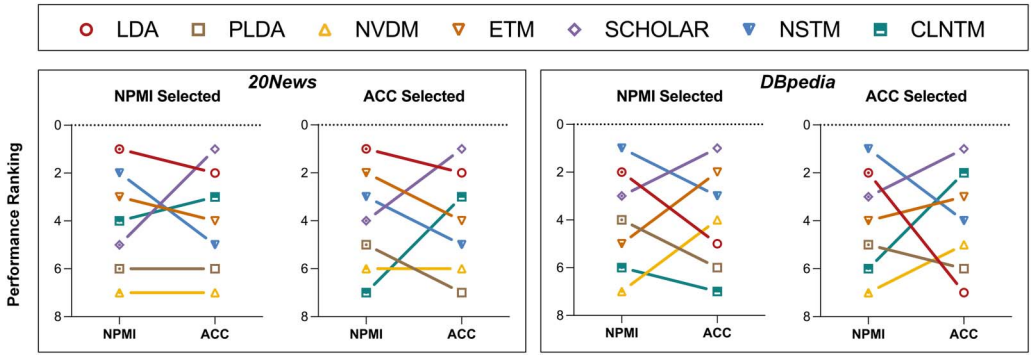

Figure 1: Performance rankings of topic quality (NPMI) and document representation quality (ACC) during model selection. The best model state/checkpoint can be determined using either NPMI or ACC as the selection criterion. However, it can be observed that the rankings for topic quality and document representation quality are inconsistent under the same selection criteria. Experiments are conducted five times, with the number of topics set to 50.

Figure 1. This inconsistency in the performance of the two components is also indicated by Bhatia et al. (2017). Therefore, evaluating a topic model based on sub-components only is insufficient to reveal the entire model's performance. Recent topic models often focus on improving topic quality, such as clustering-based models (Sia et al., 2020; Grootendorst, 2022), but they do not evaluate their effectiveness in representing documents. In this work, we aim to introduce a novel evaluation method for topic modeling that jointly assesses the semantic quality of both topics and document representations, with the help of large language models.

4.2 Key Idea

We propose to conduct the evaluation in a joint manner that considers both document representations and topics, rather than evaluating them separately as in previous works. To do so, we obtain the document-word distribution $\boldsymbol{w}$ for a given document from the topic model by running both its inference and generative process:

$$\boldsymbol{w} := f_{\phi}(f_{\boldsymbol{\theta}}(\boldsymbol{x})). \tag{3}$$

The inference process infers the document representation $\boldsymbol{z}$ for a given document $\boldsymbol{x}$, as in Eq. 1; the generative process[2] generates or reconstructs the word distribution $\boldsymbol{w}$ based on the document representation $\boldsymbol{z}$ and topics $\mathcal{T}$, as in Eq. 2. Therefore, the evaluation based on $\boldsymbol{w}$ involves both document representations and topics.

[2]We omit topics $\mathcal{T}$ in Eq. 3 as they are considered part of the parameters of the generative process ϕ.

Next, we take the top-weighted words $\mathbf{w}$ from the word distribution $\boldsymbol{w}$ generated by the topic model as the ''topical'' words of the document. Those topical words can be regarded as a semantic summary of the document from the target topic model's perspective. To generate high-quality topical words for a document, a topic model should learn good global topics as well as good document representations. Now, the evaluation of a topic model can be reframed as assessing the quality of its topical words. Suppose the true representative words $\mathbf{k}$ of document $\boldsymbol{x}$ are given, then we can formulate our evaluation task as:

$$S(\mathbf{w}, \mathbf{k}), \tag{4}$$

where $S(\cdot, \cdot)$ is a score function (Section 4.4) to quantify the agreement between $\mathbf{w}$ and $\mathbf{k}$ (Section 4.3).

4.3 Word Suggestion by LLM

Keyword Suggestion Following our evaluation task in Eq. 4, the ideal representative words $\mathbf{k}$ are from human summary of the document. However, this is expensive and impractical for large-scale evaluation. With the recent advancements in LLMs, which have demonstrated performance akin to human capabilities in various natural language processing tasks, including text summarization (Wang et al., 2023; Tang et al., 2023; Zhang et al., 2024) and keyphrase extraction (Song et al., 2023; Maragheh et al., 2023; Bai et al., 2024), we propose leveraging LLMs through prompting to generate keyword suggestions for a given document:

$$\mathbf{k} := \text{LLM}(\text{Prompt}(\boldsymbol{d})). \tag{5}$$

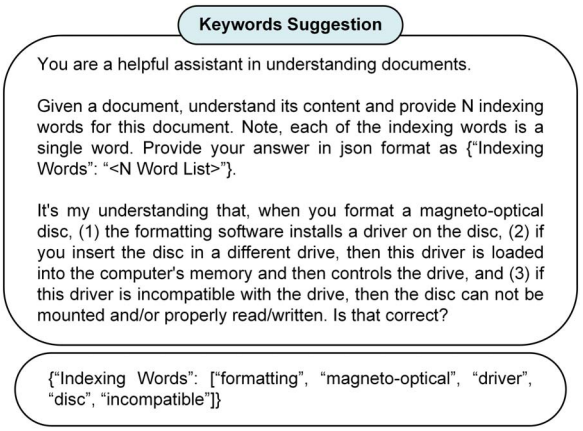

Figure 2: An example prompt and output of keywords suggestion by the LLM. In this example, the number of keywords (i.e., N) is set to 5.

Specifically, we query keywords $\mathbf{k}$ for a given document $\boldsymbol{d}$ from an LLM by proposing the prompt shown in Figure 2. The prompt consists of a task instruction and the queried document.

Topic-Aware Keywords Suggestion Analogous to the generation of topical words in topic modeling—where global topics of the document collections are identified first, followed by associated keywords for each given document—we propose prompting the LLM in a similar manner, ensuring it considers collection-level topics when providing keywords suggestion for each document, written as:

$$\mathbf{k} := \text{LLM}(\text{Prompt}(\boldsymbol{d}, \mathcal{T})), \tag{6}$$

where $\mathcal{T}$ denotes the set of topics of the text corpus. To obtain the collection-level topics $\mathcal{T}$ for the corpus by the LLM, we follow the topic generation approach by Pham et al. (2024). Briefly, it leverages an LLM to iteratively identify new topics from each document. A subsequent refinement process then merges similar topics and removes those with low frequency. For further details on the topic generation process, we refer readers to Pham et al. (2024) (section 3.1).

Using these corpus-level topics, we prompt the LLM to generate keywords for each document in a two-stage process, considering the overarching themes of the collection. In the first stage, the LLM selects relevant topics for the target document from the corpus-level topics. In the second stage, we prompt the LLM to generate indexing words for the document based on each selected topic.

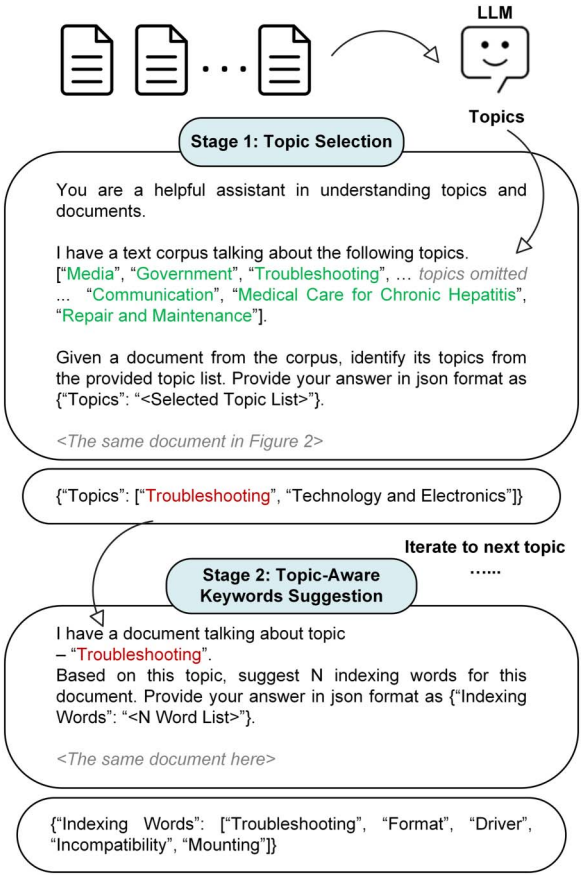

Figure 3: An illustration of topic-aware keywords suggestion pipeline. The words highlighted in green represent collection-level topics generated by the LLM. Each topic selected in stage 1 is used in the stage 2 prompt to generate topic-aware keywords.

The final set of keywords is obtained by merging the words generated for each selected topic. An example prompt and output for topic-aware keywords suggestion is shown in Figure 3.

4.4 Choices of the Score Function

For the score function $S(\cdot, \cdot)$ in Eq. 4, we propose different ways to calculate it: (1) Overlap-based, which computes the number of overlapping words between $\mathbf{w}$ and $\mathbf{k}$, and (2) Embedding-based, which calculates the overall semantic similarity between the two word sets using pre-trained word embeddings.

Word Overlap A straightforward choice of the score function is directly counting the overlaps between $\mathbf{w}$ and $\mathbf{k}$. Considering the potential variant in forms of the same word, we convert each word to its root form before counting, formulated as:

$$S_{\text{overlap}} := C(f_{\text{root}}(\mathbf{w}) \cap f_{\text{root}}(\mathbf{k})) \times f_{\text{n}}(\mathbf{w}, \mathbf{k}), \tag{7}$$

where $C(\cdot)$ and $f_{\text{root}}(\cdot)$ are the counting and rooting (e.g., stemming or lemmatization) operation, respectively; $f_{\text{n}}(\mathbf{w}, \mathbf{k}) := 1/(N+M)$ returns the normalising factor based on two input word sets, where N and M are the number of words in $\mathbf{w}$ and $\mathbf{k}$, respectively.

Synset Overlap Considering the case that different words may describe the same or similar concept (e.g., ''puppy'' and ''dog''), we leverage WordNet (Miller, 1995) synsets to determine word overlaps: if the synsets of two words intersect, they are considered to overlap. Then, we define the synset overlap score as:

$$S_{\text{synset}} := (\sum_{i=1}^{N}\sum_{j=1}^{M} \mathbb{1}(C(f_{\text{synset}}(\text{w}_i) \cap f_{\text{synset}}(\text{k}_j)) > 0)) \times f_{\text{n}}(\mathbf{w}, \mathbf{k}), \quad (8)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function; $f_{\text{synset}}(\cdot)$ is a function that returns the synset for a given word. Intuitively, the synset-based score builds on the idea of word overlap, considering two words as overlapping if their synsets intersect, rather than requiring an exact match.

Word Optimal Assignment We consider another choice of S, which measures the overall semantic similarity between two sets of words with pre-trained word embeddings (Mikolov et al., 2013; Pennington et al., 2014). Since the alignments between words from $\mathbf{w}$ and $\mathbf{k}$ are unknown, directly measuring similarity between word embeddings is not feasible. To automatically find the alignment for each word of $\mathbf{w}$ to each word of $\mathbf{k}$, we formulate it as the following Optimal Assignment (OA) problem and solve it using the Hungarian algorithm (Kuhn, 1955): Given a word set $\mathbf{w}$ that has N words: $\mathbf{w} := \{\text{w}_1, \text{w}_2, \ldots, \text{w}_N\}$ and their embedding vectors $\boldsymbol{E}^{\mathbf{w}} := \{\boldsymbol{e}^{\text{w}_1}, \boldsymbol{e}^{\text{w}_2}, \ldots, \boldsymbol{e}^{\text{w}_N}\}$; and another word set $\mathbf{k}$ that has M words: $\mathbf{k} := \{\text{k}_1, \text{k}_2, \ldots, \text{k}_M\}$ with related embedding vectors $\boldsymbol{E}^{\mathbf{k}} := \{\boldsymbol{e}^{\text{k}_1}, \boldsymbol{e}^{\text{k}_2}, \ldots, \boldsymbol{e}^{\text{k}_M}\}$. Define a cost matrix $\boldsymbol{C} \in \mathbb{R}_{\geq 0}^{N \times M}$ whose entry $C_{i,j} := \text{CosD}(\boldsymbol{e}^{\text{w}_i}, \boldsymbol{e}^{\text{k}_j})$, where $\text{CosD}(\cdot,\cdot)$ denotes the cosine distance function; and a binary matrix $\boldsymbol{A} \in \{0,1\}^{N \times M}$ whose entry $A_{i,j} = 1$ if word w_i is assigned to word k_j, and 0 otherwise. The goal is to solve the following optimal assignment problem:

$$\min_{\boldsymbol{A}} \sum_{i=1}^{N}\sum_{j=1}^{M} C_{i,j} \times A_{i,j}, \quad (9)$$

subject to $\sum_{j=1}^{M} A_{i,j} = 1$ and $\sum_{i=1}^{N} A_{i,j} = 1$. By finding the optimal binary matrix $\boldsymbol{A}^*$, we obtain the distance between $\mathbf{w}$ and $\mathbf{k}$ by:

$$S_{\text{oa}} := D_{\text{oa}}(\mathbf{w}, \mathbf{k}) := \sum_{i=1}^{N}\sum_{j=1}^{M} C_{ij} \times A^*_{ij}. \quad (10)$$

Word Optimal Transport Optimal Transport (OT) has recently been used as a powerful geometric tool to measure the distance between distributions, with rich applications in machine learning and related areas (Ge et al., 2021; Zhao et al., 2021c; Nguyen et al., 2021; Wang et al., 2022; Guo et al., 2022; Bui et al., 2022; Vuong et al., 2023; Zhao et al., 2023; Ye et al., 2024; Vo et al., 2024; Gao et al., 2024). Considering that both $\mathbf{w}$ and $\mathbf{k}$ are top words of probability distributions, where each word essentially retains a portion of probability mass. Our previous calculations ignore the probability mass of words and treat each word in the set as equal. Now we include the probability mass and formulate the similarity calculation between $\mathbf{w}$ and $\mathbf{k}$ as an OT problem: Given two discrete distributions $\mu(\mathbf{w}, \boldsymbol{w})$ and $\mu(\mathbf{k}, \boldsymbol{k})$, where $\mathbf{w} := \{\text{w}_1, \text{w}_2, \ldots, \text{w}_N\}$ and $\mathbf{k} := \{\text{k}_1, \text{k}_2, \ldots, \text{k}_M\}$ are the supports of those two distributions; $\boldsymbol{w} \in \Delta^N$ and $\boldsymbol{k} \in \Delta^M$ are their related probability vectors[3]; following the same construction of cost matrix $\boldsymbol{C}$ in the previous OA problem, the OT problem between $\mu(\mathbf{w}, \boldsymbol{w})$ and $\mu(\mathbf{k}, \boldsymbol{k})$ is defined as:

$$\min_{\boldsymbol{P}} \sum_{i=1}^{N}\sum_{j=1}^{M} C_{i,j} \times P_{i,j}, \quad (11)$$

subject to $\sum_{j=1}^{M} P_{i,j} = w_i$ and $\sum_{i=1}^{N} P_{i,j} = k_j$; $\boldsymbol{P} \in \mathbb{R}_{\geq 0}^{N \times M}$ is the transport plan, whose entry $P_{i,j}$ indicates the amount of probability mass moving from w_i to k_j. Similarly, by finding the optimal transport plan $\boldsymbol{P}^*$ using solvers such as those in Flamary et al. (2021), the OT distance between $\mu(\mathbf{w}, \boldsymbol{w})$ and $\mu(\mathbf{k}, \boldsymbol{k})$ is obtained by:

$$S_{\text{ot}} := D_{\text{ot}}(\mu(\mathbf{w}, \boldsymbol{w}), \mu(\mathbf{k}, \boldsymbol{k})) := \sum_{i=1}^{N}\sum_{j=1}^{M} C_{ij} \times P^*_{ij}. \quad (12)$$

Compared with our OA and OT formulations for WALM, they are similar in that they both

[3]We assume $\mu(\mathbf{k}, \boldsymbol{k})$ is a uniform distribution over the keywords $\mathbf{k}$ from the LLM. Thus, $\boldsymbol{k}$ is a uniform probability vector.

construct the cost matrix $\boldsymbol{C}$ using cosine distance between pre-trained word embeddings. However, they differ in the following ways: (1) OA treats words in the set as equal, while OT considers probability mass of each word. (2) OA can be viewed as a ''hard'' assignment problem between two word sets because the entries of $\boldsymbol{A}$ are binary. In contrast, OT can be regarded as a ''soft'' assignment because of the spread of probability mass in $\boldsymbol{P}$.

5 Experiments

5.1 Experimental Setup

Datasets Two widely used datasets, 20Newsgroup (Lang, 1995) (**20News**), which contains long documents, and **DBpedia** (Auer et al., 2007), which includes short documents, are used for our experiments. Our pre-processed datasets are available in the Github repository.

Evaluated Models We conduct experiments on 7 popular topic models from traditional probabilistic to recent neural topic models. (1) Latent Dirichlet Allocation (**LDA**) (Blei et al., 2003), the most popular probabilistic topic model that assumes a document is generated by a mixture of topics. (2) LDA with Products of Experts (**PLDA**) (Srivastava and Sutton, 2017), an early NTM that applies the product of experts instead of the mixture of multinomials in LDA. (3) Neural Variational Document Model (**NVDM**) (Miao et al., 2017), a pioneer NTM that uses a Gaussian as the prior distribution of topic proportions of documents. (4) Embedded Topic Model (**ETM**) (Dieng et al., 2020), an NTM that involves word and topic embeddings in the generative process. (5) Neural Topic Model with Covariates, Supervision, and Sparsity (**SCHOLAR**) (Card et al., 2018), an NTM that applies a logistic normal prior on topic proportions and leverages extra information from metadata. (6) Neural Sinkhorn Topic Model (**NSTM**) (Zhao et al., 2021b), a recent NTM based on an optimal transport framework. (7) Contrastive Learning Neural Topic Model (**CLNTM**) (Nguyen and Luu, 2021), a recent NTM that uses contrastive learning to regularize document representations. We keep all these models' default settings as suggested in their implementations. All experiments are conducted 5 times with different model random seeds; mean and standard deviation values are reported.

Settings of WALM For the WALM settings, we use GloVe word embeddings pre-trained on Wikipedia (Pennington et al., 2014)[4] in our embedding-based metrics. For the LLM generation settings, we use LLAMA3-8B-Instruct[5] for our main experiments. We employ greedy decoding during LLM generation to ensure deterministic outputs, setting the maximum number of generated tokens to 300. When prompting the LLM, we limit the number of generated keywords to 5. For topical words from the topic model, we select the top 10 weighted words from the document-word distribution for each given document.

Settings of Existing Metrics We also evaluate topic models with existing commonly used metrics to compare with ours. (1) Topic Coherence and Diversity: We apply **NPMI** to evaluate topic coherence using Wikipedia as the reference corpus, done by the Palmetto package[6] (Röder et al., 2015). Following standard protocol, we consider the top 10 words of each topic and obtain the average NPMI score of topics by selecting the top 50% coherent topics. As for Topic Diversity (**TD**), we compute the percentage of unique words in the top 25 words of all topics, as defined in Dieng et al. (2020). (2) Document Representation Quality: We conduct document classification and clustering to evaluate the representation capability of topic models. As for classification, we train a Random Forest classifier based on the training documents' representation and test the accuracy (**ACC**) in the testing documents, as in previous works such as Nguyen and Luu (2021). As for clustering, we conduct K-Means clustering based on test documents' representation and report the Purity (**KM-Purity**) and Normalized Mutual Information (**KM-NMI**), as in previous works such as Zhao et al. (2021b). (3) Perplexity: We use document completion perplexity (Wallach et al., 2009) to evaluate the predictive ability of topic models. We split each test document into two equal-length folds randomly. Then we compute the Document Completion Perplexity (**DC-PPL**) on the second fold of documents based on the topic proportion inferred from the first fold, as in previous works such as Dieng et al. (2020).

[4] https://nlp.stanford.edu/projects/glove/.

[5] https://huggingface.co/meta-llamaMeta-/Llama-3-8B-Instruct.

[6] https://github.com/dice-group/Palmetto.

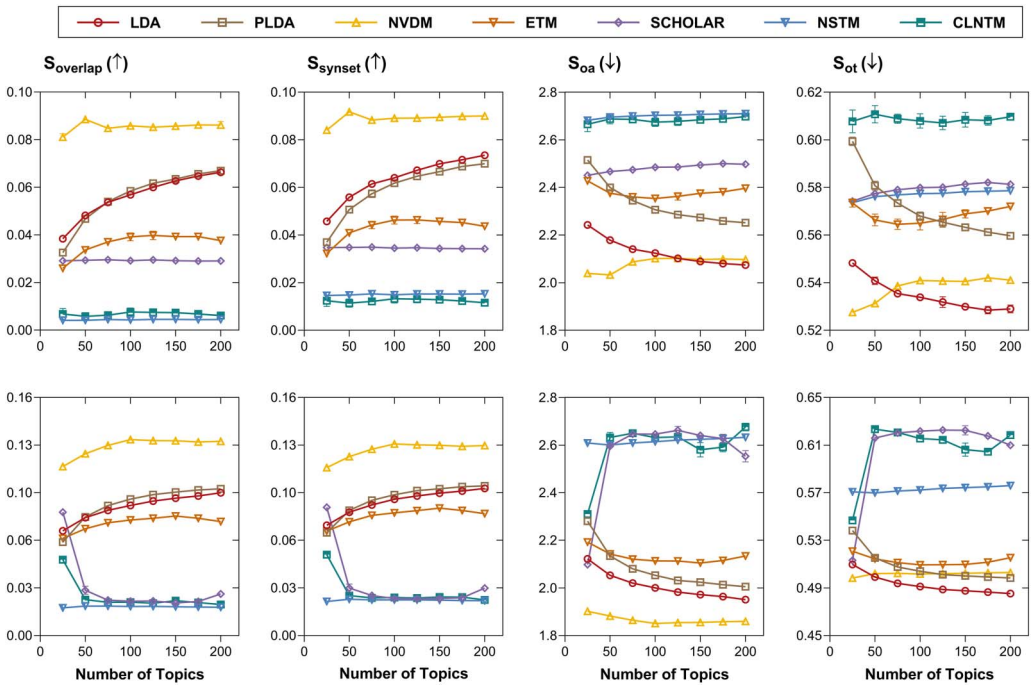

Figure 4: Topic models' performance in terms of WALM with **keywords suggestion** by the LLM on **20News** (top row) and **DBpedia** (bottom row). Error bars represent the standard deviation (omitted for values smaller than the symbol size).

5.2 Results and Analysis

Topic Model Evaluation with WALM We assess topic models' performance based on our evaluation metrics on both 20News and DBpedia. We have the following observations based on our results illustrated in Figure 4: (1) The WALM values of most models on DBpedia show better performance than 20News, which indicates that it is easier for topic models to generate informative topical words for short documents than long documents. (2) The performance ranking indicated by overlap-based metrics (e.g., S_{overlap} and S_{synset}) and embedding-based metrics (e.g., S_{oa} and S_{ot}) is slightly different. The reason is that embedding-based metrics consider the semantic distance among words, which can be more flexible than the exact match in overlap-based metrics. (3) It can be observed that there is little improvement from recent NTMs over LDA and NVDM in terms of our joint metrics. The potential reason is that most contemporary NTMs primarily focus on enhancing topic coherence while neglecting the generation of documents, thus showing weak performance in generating topical words of documents as indicated by WALM. (4) When topic-aware keyword suggestion is applied in WALM (Figure 5), the performance ranking of LDA surpasses that of NVDM as the number of topics increases in the long-document dataset (i.e., 20News). This suggests that LDA benefits more from an increased number of topics when generating topic-aware keywords for documents compared to NVDM.

Learning Curves of WALM In Figure 6, we illustrate the learning curves of topic models in terms of WALM, clearly showing how each metric changes throughout the training process. We observe that most topic models improve with training and eventually converge to a stable state. However, NVDM exhibits overfitting in the later stages of training, as indicated by its WALM scores. Additionally, WALM approaches based on keyword suggestions and topic-aware keyword suggestions exhibit slightly different trends in their learning curves. For instance, LDA surpasses NVDM in the later training stages when topic-aware keywords are used. This suggests that NVDM prioritizes document-level generation while LDA shows stronger awareness of collection-level topics.

Qualitative Analysis on Topical Words for Documents We qualitatively investigate the topical words of documents by topic models at different stages in Table 1, where we randomly sample

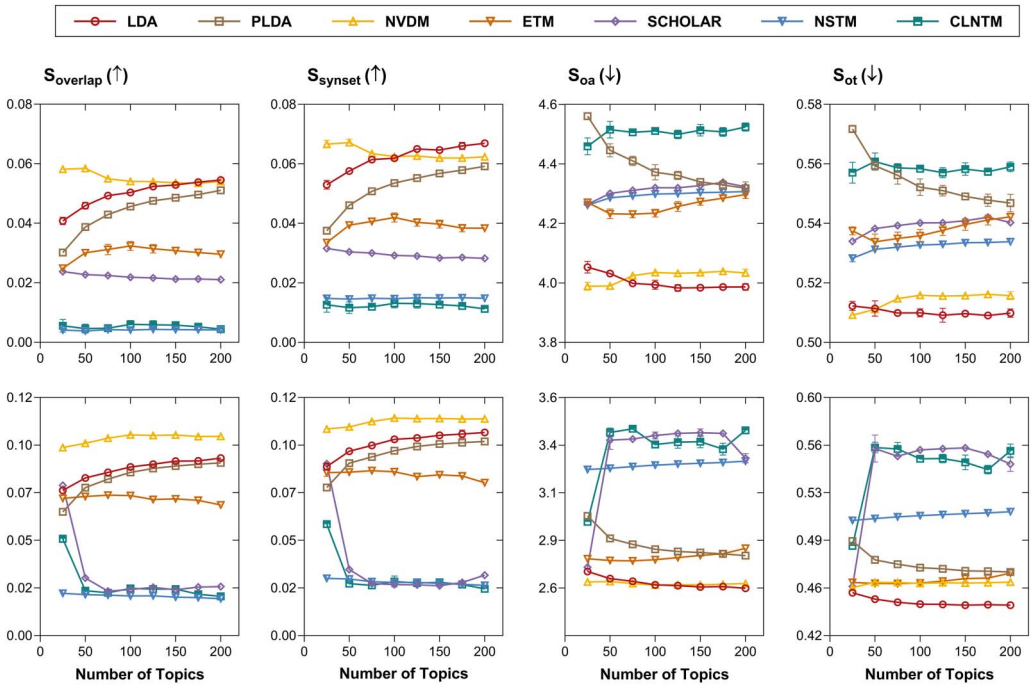

Figure 5: Topic models' performance in terms of WALM with **topic-aware keywords suggestion** by the LLM on **20News** (top row) and **DBpedia** (bottom row). Error bars represent the standard deviation (omitted for values smaller than the symbol size).

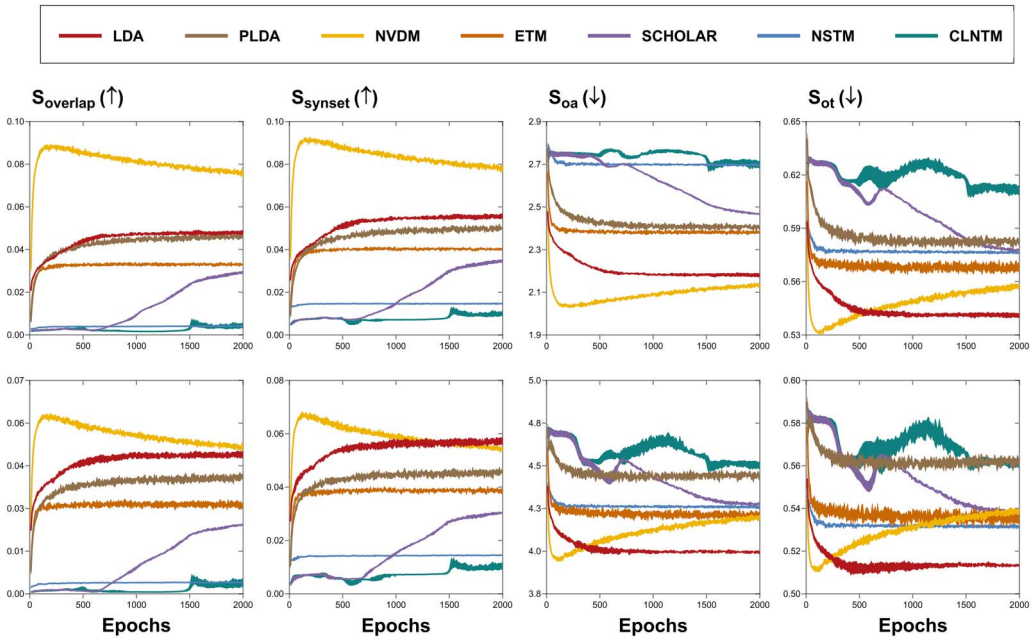

Figure 6: Learning curves of topic models in terms of WALM with **keyword suggestions** (top row) and **topic-aware keyword suggestions** (bottom row) from the LLM on the 20News test set, with the number of topics set to 50. The area within the error bands represents the standard deviation.

one document for 20News and DBpedia, respectively. We have the following observations based on our results: (1) The topical words at the beginning phase contain less semantically related words about the documents than those at convergence, which aligns with the learning status (as in Figure 6) indicated by WALM. (2) The topical words of NVDM include more words that reveal the documents' main messages than LDA, which aligns with the ranking (as in Figure 5) suggested by WALM. (3) The keywords generated by the LLM are similar to those provided

Document	Model	Topical Words
It's my understanding that, when you format a magneto-optical disc, (1) the formatting software installs a driver on the disc, (2) if you insert the disc in a different drive, then this driver is loaded into the computer's memory and then controls the drive, and (3) if this driver is incompatible with the drive, then the disc can not be mounted and/or properly read/written. Is that correct?	LDA_B	drive, disk, card, controller, hard, mb, file, scsi, bios, power
	LDA_C	drive, disk, scsi, hard, card, controller, mb, floppy, ide, sale
	NVDM_B	driver, drive, problem, card, time, file, thanks, need, email, work
	NVDM_C	drive, driver, hard, scsi, window, cd, mb, floppy, disc, work
	LLM	formatting, magneto-optical, driver, disc, incompatible
	LLM (Topic-Aware)	troubleshooting, formatting, incompatibility, magneto-optical, driver, disc, mounting
	Human	driver, disc, computer, hardware, software, memory, formatting, incompatible
Wrong World. Wrong World is a 1985 Australian film directed by Ian Pringle. It was filmed in Nhill and Melbourne in Victoria Australia.	LDA_B	film, american, released, directed, football, album, summer, played, team, hospital
	LDA_C	film, played, directed, baseball, league, australian, major, drama, football, award
	NVDM_B	specie, album, school, known, located, north, film, directed, american, released
	NVDM_C	film, album, released, second, south, new, directed, american, australian, known
	LLM	world, film, australian, directed, victoria
	LLM (Topic-Aware)	film, industry, production, cinema, entertainment
	Human	film, movie, directed, director, australian, melbourne, victoria

Table 1: Documents' topical words from topic models at the beginning phase (e.g., NVDM_B, LDA_B) and convergence phase (e.g., NVDM_C, LDA_C) according to WALM, where the number of topics is set to 50.

by human annotators for the example documents. (4) By using topic-aware keywords suggestion in WALM, the LLM tends to provide keywords that convey the high-level concepts of the topics. For instance, ''troubleshooting'' is identified for the first example document, and ''entertainment'' for the second, which offers higher-level information from topics besides individual document.

Correlation to Other Metrics We compute Pearson's correlation coefficients among existing and WALM series metrics, similar to the correlation analysis in previous works such as Doogan and Buntine (2021) and Rahimi et al. (2024). Pearson's correlation coefficients among the metrics are plotted in a heatmap in Figure 7. Based on the results, we observe that: (1) WALM variants are highly correlated with each other since they originate from the same mechanism. (2) Compared with perplexity, which also evaluates the entire model based on documents, WALM shows weak correlations, suggesting a new family of evaluation metrics. (3) Compared with other types of evaluations, WALM has moderate correlations with document representation metrics (e.g., KM-Purity, KM-NMI, and ACC), and weak correlations with topic quality metrics (e.g., NPMI and TD). This indicates that our joint evaluation metrics take both components into account without relying solely on either one. These observations suggest that WALM can serve as a complementary evaluation method to existing approaches.

5.3 Contextualized Embeddings for WALM

Obtaining Contextualized Embeddings Recall that in Eq. 10 and Eq. 12, the cost matrix

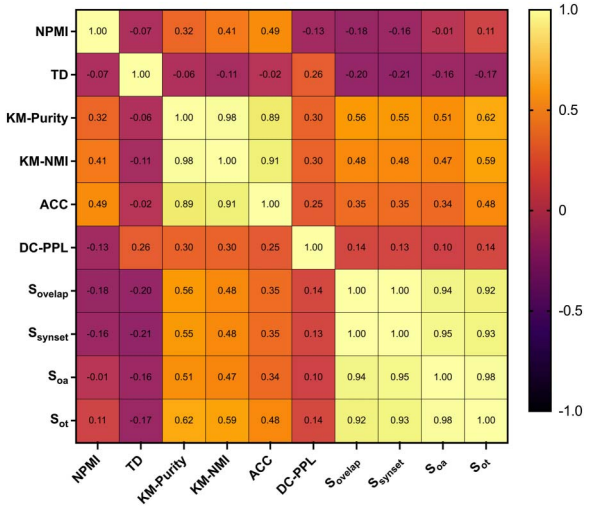

Figure 7: Pearson's correlation coefficient among evaluation metrics.

$\boldsymbol{C}$ is constructed using cosine distances between word embeddings. Here, we change our construction of $\boldsymbol{C}$ from using static word embeddings from GloVe (Pennington et al., 2014) to the contextualized word embeddings from the LLM, considering that the same word may have different semantic meanings in different contexts. We obtain the contextualized embeddings of a word given a document differently in two cases: (1) When the target word appears in the context document, we take the average embeddings of each occurrence as the contextualized embedding. (2) When there is no occurrence of the target word in the given document, we add an auxiliary sentence to the document in the following format:

> *''<Given Document>. This document is talking about <Target Word>.''*

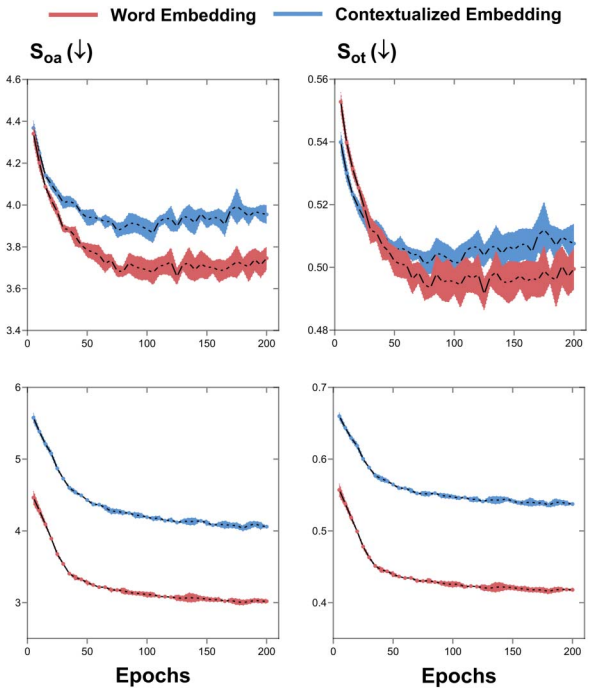

Figure 8: Learning curves of NVDM in terms of embedding-based metrics and their contextualized variants on 20News (top row) and DBpedia (bottom row). The area within the error bands represents the standard deviation.

Then, we obtain the contextualized embedding of the target word given the document with the auxiliary sentence. By replacing the global word embeddings with contextualized word embeddings, we have new variants of our embedding-based WALM (i.e., S_{oa} and S_{ot}), i.e., $\mathbf{S_{oa\_c}}$ and $\mathbf{S_{ot\_c}}$.

Observations Since the cost of obtaining contextualized embeddings is high for LLMs, we compute $S_{\text{oa\_c}}$ and $S_{\text{ot\_c}}$ in a case study, where we test NVDM on 100 documents randomly sampled from the test sets of 20News and DBpedia, respectively. We plot the learning curves on test documents in Figure 8. We observe that using word embeddings or contextualized embeddings in our embedding-based scores exhibits similar trends but with different values on both datasets.

5.4 Sensitivity Study

Here, we examine two factors that can influence the WALM scores: the number of keywords generated by the LLM and the choice of the LLM. To investigate the effect of the number of keywords, we vary the number from 3 to 10 and plot the performance ranking of topic models in Figure 9 (top row). We observe that, although the values of WALM metrics can vary with different numbers of keywords, the overall performance ranking of the topic models remains largely unaffected by these changes, especially for the overlap-based metrics. To investigate the effect of LLMs, we use different latest LLMs for keyword generation apart from LLAMA3-8B-Instruct, including Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Phi-3-Mini-128K-Instruct (Abdin et al., 2024) and Yi-1.5-9B-Chat (Young et al., 2024). From the results illustrated in Figure 9 (bottom row), we observe that overlap-based metrics show minimal variation with different choices of LLMs, and the performance ranking of the topic models is unaffected in most cases. These observations suggest that the overlap-based metrics are less sensitive to the number of words and the choice of LLMs.

5.5 Comparisons with Human Annotation

Evaluation Gap with Human Annotation WALM computes the difference between documents' topical words generated by topic models and an LLM, treating the words from the LLM as the ground truth. Here, we investigate the gap between using LLM and human judgment as the true topical words in WALM. To quantify this gap, we use the following calculation:

$$G := \frac{|S(\text{LLM as Truth}) - S(\text{Human as Truth})|}{S(\text{Human as Truth})}, \tag{13}$$

where S is the WLAM scores we propose in Section 4.4 (and the contextualized variants in Section 5.3). Intuitively, the gap function measures the difference between using ground truth (e.g., keywords) from the LLM and those from human annotators. We empirically observe that topical words from topic models consistently differ from those identified by humans, so the denominator in Eq. 13 will not be zero.

As human annotation is expensive for large-scale investigation, we randomly sample 200 test documents from 20News and DBpedia as a case study. We engaged three English speakers as annotators, trained with a few examples, to provide keywords that capture the main points of each document. Then, given a trained topic model, we compute the gap between using the words from the LLM and human in our metrics using Eq. 13.

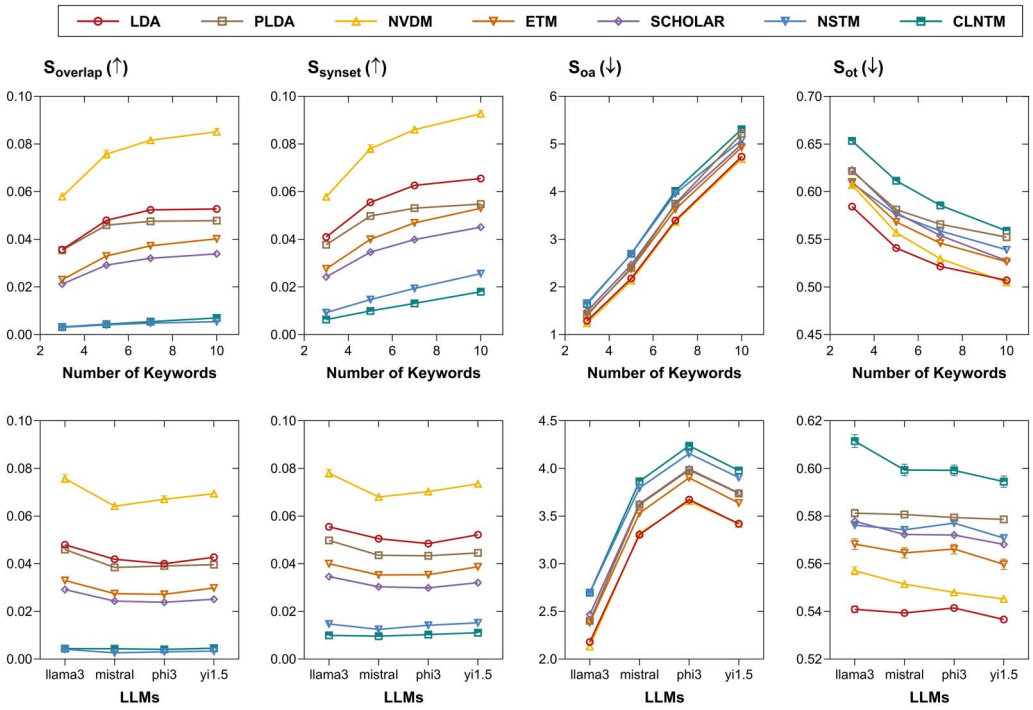

Figure 9: Sensitivity study. **Top row**: Performance of topic models in terms of WALM with varying numbers of keywords. **Bottom row**: Performance of topic models in terms of WALM with different LLMs. Experiments are conducted on the 20News dataset with the number of topics set to 50. Error bars represent the standard deviation (omitted for values smaller than the symbol size).

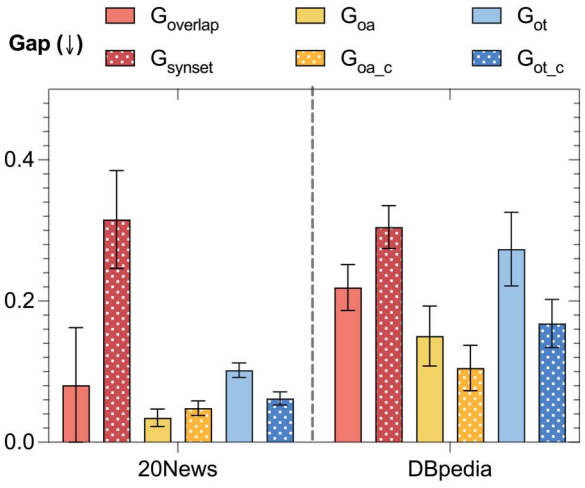

Figure 10: Evaluation gap between using the LLM and human judgment as the ‘‘true’’ topical words. Error bars represent the standard deviation.

The results are illustrated in Figure 10, where the evaluated model is NVDM with $K = 50$ trained on 20News and DBpedia, respectively. We have the following observations based on the results: (1) Comparing the datasets, the gap between using human judgment and the LLM in 20News is lower than in DBpedia in most cases. This indicates that for long documents such as those in 20News, the topical words generated by the LLM are closer to human judgment than in short documents in DBpedia. (2) Comparing the metrics, S_{oa} exhibits the lowest gap among WALM metrics, with a gap value of 0.03 and 0.15 on 20News and DBpedia, respectively. This shows the effectiveness of using the LLM as a proxy for human judgment when applied in S_{oa}. (3) Comparing the embeddings, using contextualized embeddings from the LLM can further narrow the evaluation gap for S_{oa} and S_{ot} on short documents.

Correlation with Human Annotation We use an existing annotated dataset, 500N-KPCrowd (Marujo et al., 2012) for the keyphrase extraction task (Hasan and Ng, 2014), where each test document is paired with labeled keywords. We run LDA on the training documents and infer the topical words for the test documents, then compute the Pearson’s correlation coefficient between the WALM scores using the LLM-generated keywords and the test labels as the ground truth. The results are illustrated in Table 2. We observe that (1) using keywords from the LLM in WALM scores correlates with using the labeled keyphrases, and (2) the correlation can potentially

	S_{overlap}	S_{synset}	S_{oa}	S_{ot}
5-word suggestion	0.55	0.50	0.57	0.63
10-word suggestion	0.52	0.58	0.56	0.68

Table 2: Pearson's correlation coefficient between WALM using LLM-generated keywords and human annotations as the ground truth on the 500N-KPCrowd dataset.

improve when more keywords are included in the LLM's suggestions.

6 Conclusion

In this work, we propose WALM for topic model evaluation, which takes both topic and document representation quality into account jointly. WALM measures the agreement between the topical words generated by topic models and those from the LLM for given documents. The topical words from the LLM are obtained through keyword prompting or topic-aware keyword prompting, with the latter tending to capture higher-level information. To quantify the agreement between word sets, we propose different calculations, including overlap-based and embedding-based metrics. Our experiments demonstrate that the WALM series effectively reflect the capability of topic models to provide semantic summaries of documents. We show that WALM metrics align with human judgment and can serve as an informative complementary method for topic model evaluation. We suggest that overlap-based metrics demonstrate better sensitivity handling, while embedding-based metrics show a smaller evaluation gap. A potential risk of using WALM is that models chasing this metric only may be affected by the bias of LLMs. To mitigate the risk, we suggest using WALM with other metrics together.

Acknowledgments

We thank the anonymous reviewers and the action editor, Michael Elhadad, for their valuable feedback, which has significantly strengthened this work.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *International Semantic Web Conference*, pages 722–735. Springer. https://doi.org/10.1007/978-3-540-76298-0_52

Xiao Bai, Xue Wu, Ivan Stojkovic, and Kostas Tsioutsiouliklis. 2024. Leveraging large language models for improving keyphrase generation for contextual targeting. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM'24, pages 4349–4357, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3627673.3680093

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2017. An automatic approach for document-level topic model evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 206–215, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-1022

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2018. Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1098

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2). https://doi.org/10.1145/1667053.1667056

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Anh Tuan Bui, Trung Le, Quan Hung Tran, He Zhao, and Dinh Phung. 2022. A unified Wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*.

Wray Buntine. 2009. Estimating likelihoods for topic models. In *Asian Conference on Machine Learning*, pages 51–64. Springer. https://doi.org/10.1007/978-3-642-05224-8_6

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1189

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean,

Slav Petrov, and Noah Fiedel. 2024. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(1).

Adji B. Dieng, 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

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

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Zhe Gan, R. Henao, D. Carlson, and Lawrence Carin. 2015. Learning deep sigmoid belief networks with data augmentation. In *AISTATS*, pages 268–276.

Jintong Gao, He Zhao, Dan dan Guo, and Hongyuan Zha. 2024. Distribution alignment optimization through neural collapse for long-tailed classification. In *Forty-first International Conference on Machine Learning*.

Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. 2021. OTA: Optimal transport assignment for object detection. In *CVPR*, pages 303–312. https://doi.org/10.1109/CVPR46437.2021.00037

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Dandan Guo, Long Tian, Minghe Zhang, Mingyuan Zhou, and Hongyuan Zha. 2022. Learning prototype-oriented set representations for meta-learning. In *International Conference on Learning Representations*.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273.

Jacob Louis Hoover, Wenyu Du, Alessandro Sordoni, and Timothy J. O'Donnell. 2021. Linguistic dependencies and statistical dependence. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2963, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.234

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97. https://doi.org/10.1002/nav.3800020109

Ken Lang. 1995. Newsweeder: Learning to filter netnews, *Machine Learning Proceedings 1995*, pages 331–339. Elsevier. https://doi.org/10.1016/B978-1-55860-377-6.50048-7

Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. *Advances in Neural Information Processing Systems*, 25.

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

Caitlin Doogan Poet Laureate, Wray Buntine, and Henry Linger. 2023. A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, pages 1–33.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic

modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22. https://doi.org/10.1186/s40064-016-3252-8, PubMed: 27652181

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. 2019. Don't blame the elbo! A linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32.

Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Llm-take: theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4318–4324. IEEE. https://doi.org/10.1109/BigData59044.2023.10386476

Luís Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, and Joaćfo P. Neto. 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey. European Language Resources Association (ELRA).

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, pages 2410–2419. PMLR.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

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

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *Advances in Neural Information Processing Systems*, 34:11974–11986.

Tuan Nguyen, Trung Le, He Zhao, Quan Hung Tran, Truyen Nguyen, and Dinh Phung. 2021. Most: Multi-source domain adaptation via optimal transport for student-teacher learning. In *UAI*, pages 225–235.

Sergey I. Nikolenko. 2016. Topic quality metrics based on distributed word representations. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1029–1032. https://doi.org/10.1145/2911451.2914720

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2015. Nested hierarchical Dirichlet processes. *TPAMI*, 37(2):256–270. https://doi.org/10.1109/TPAMI.2014.2318728, PubMed: 26353240

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.164

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

Martin Reisenbichler and Thomas Reutterer. 2019. Topic modeling in marketing: Recent advances and research opportunities. *Journal of Business Economics*, 89(3):327–356. https://doi.org/10.1007/s11573-018-0915-7

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408. https://doi.org/10.1145/2684822.2685324

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Mingyang Song, Xuelian Geng, Songfang Yao, Shilong Lu, Yi Feng, and Liping Jing. 2023. Large language models as zero-shot keyphrase extractor: A preliminary empirical study. *arXiv preprint arXiv:2312.15156*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations*.

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

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, 6(1):158. https://doi.org/10.1038/s41746-023-00896-7, PubMed: 37620423

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian

Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vy Vo, He Zhao, Trung Le, Edwin V. Bonilla, and Dinh Phung. 2024. Optimal transport for structure learning under missing data. In *International Conference on Machine Learning*.

Long Tung Vuong, Trung Le, He Zhao, Chuanxia Zheng, Mehrtash Harandi, Jianfei Cai, and Dinh Phung. 2023. Vector quantized wasserstein auto-encoder. In *International Conference on Machine Learning*, pages 35223–35242. PMLR.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. https://doi.org/10.1145/1553374.1553515

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. In *International Conference on Learning Representations*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.newsum-1.2

Xiaohao Yang, He Zhao, Dinh Phung, and Lan Du. 2023. Towards generalising neural topical representations. *arXiv preprint arXiv:2307.12564*.

Hangting Ye, Wei Fan, Xiaozhuang Song, Shun Zheng, He Zhao, Dan dan Guo, and Yi Chang. 2024. Ptarl: Prototype-based tabular representation learning via space calibration. In *International Conference on Learning Representations*.

Xing Yi and James Allan. 2009. A comparative study of utilizing topic models for information retrieval. In *Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6–9, 2009. Proceedings 31*, pages 29–41. Springer. https://doi.org/10.1007/978-3-642-00958-7_6

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01. AI. *arXiv preprint arXiv:2403.04652*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57. https://doi.org/10.1162/tacl_a_00632

He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. 2018a. Dirichlet belief networks for topic structure learning. In *NeurIPS*, pages 7966–7977. https://doi.org/10.24963/ijcai.2021/638

He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. 2018b. Inter and intra topic structure learning with word embeddings. In *ICML*, pages 5887–5896.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021a. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4713–4720. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2021c. Neural topic model via

optimal transport. In *International Conference on Learning Representations*.

He Zhao, Ke Sun, Amir Dezfouli, and Edwin V. Bonilla. 2023. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*, pages 42159–42186. PMLR.

Mingyuan Zhou, Yulai Cong, and Bo Chen. 2016. Augmentable gamma belief networks. *JMLR*, 17(163):1–44.