Explicit Bayesian Inference to Uncover the Latent Themes of Large Language Models

Raymond Li[†] **, Chuyuan Li**[†] **, Gabriel Murray**[‡] **, Giuseppe Carenini**[†]
[†] University of British Columbia, Vancouver, BC, Canada
[‡] University of Fraser Valley, Abbotsford, BC, Canada
{raymondl, carenini}@cs.ubc.ca, chuyuan.li@ubc.ca
gabriel.murray@ufv.ca

Abstract

Large language models (LLMs) have demonstrated impressive generative capabilities, yet their inner mechanisms remain largely opaque. In this work, we introduce a novel approach to interpret LLMs generation process through the lens of an explicit Bayesian framework by inferring latent topic variables via variational inference. Specifically, we leverage a variational autoencoder-based neural topic model to dynamically approximate the posterior distribution over the high-level latent topic variables at each generation step. By reconstructing the LLM's next-token predictions through these latent topics and maintaining a regularized latent space, our method yields interpretable and diverse topic representations but also has the ability to effectively captures semantic shifts throughout the text. We validate our approach on multiple datasets, showing that our latent topics outperform state-of-the-art topic models on intrinsic measures of coherence and diversity. Furthermore, we demonstrate the utility of our approach in downstream applications by using the inferred topic distributions to retrieve relevant demonstration examples for in-context learning, resulting in significant gains on classification and summarization tasks.

1 Introduction

Large Language Models (LLMs) have achieved impressive performance across a variety of benchmarks (Brown et al., 2020; OpenAI et al., 2024; Touvron et al., 2023a,b; Grattafiori et al., 2024). This can primarily be attributed to the large-scale pre-training on massive text corpora, where they learn to generalize by outputting the appropriate continuation given the task prompts. However, at such large scales, understanding their inner workings becomes very challenging due to the opaque nature of deep neural networks. One promising direction to understand their behavior is through the Bayesian framework (Xie et al., 2022; Wang et al., 2023; Dalal and Misra, 2024), where the model learns to first infer a posterior distribution over some latent variables conditioned on the previous context before predicting the next token. When these latent variables are disentangled (Kingma and Welling, 2014; N et al., 2017; Higgins et al., 2017; Adel et al., 2018), each component can reflect a distinct interpretable feature (e.g., a topic) in the generation process, where the model then uses the inferred distribution over these features to guide its predictions. Thus, the activations of these latent variables may provide a transparent overview of the model's internal structures during the generation process. In this work, we focus on inferring latent topic variables, which are high-level thematic dimensions of the model represented as clusters of related words and concepts.

However, since LLMs do not explicitly learn a latent topic distribution (Xie et al., 2022), supplementary techniques are required to infer the posterior distribution over latent topics from the model predictions. For example, the recent work by Wang et al. (2023) used prompt tuning to learn a set of token embeddings as the optimal latent variable values under the Empirical Bayesian framework. However, their fixed topic embeddings are difficult to interpret and cannot be used to reveal how the model's topical shifts at each step. Instead, we propose a novel framework that uses variational inference to approximate the posterior distribution over the latent topics at each generation step. Specifically, we train a Neural Topic Model (NTM) (Srivastava and Sutton, 2017) based on the variational autoencoder (VAE) architecture (Kingma and Welling, 2014) to dynamically infer a distribution over topic variables at each generation step, where topics are represented as distributions over the vocabulary. At each step of the generation, the VAE constructs the LLM's predicted token probability based on the sampled topic distribution conditioned on the previous context.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21819–21833

We empirically validate the effectiveness of our proposed method by performing an intrinsic and extrinsic evaluation of the latent topics. For **intrinsic evaluation**, we perform experiments to assess the interpretability of the inferred latent topics of an LLM by training our model on a collection of documents from the AG News (Zhang et al., 2015), DBPedia, and GovReports dataset (Huang et al., 2021). Comparison against state-of-the-art topic modeling techniques reveals that our method can consistently learn high-quality topics which can be used directly by humans for analyzing the given text corpus (Section 4). For **extrinsic evaluation**, we conduct experiments to assess the effectiveness of the inferred topics on two popular downstream tasks—text classification and summarization—using in-context learning. In these tasks, the topic distribution serves as a text embedding to identify demonstrations with similar topic distributions (see Section 5).

In summary, our contributions are three fold:

1. We propose a novel technique to infer the latent topic variables of LLMs through VAEs as a form of approximated Explicit Bayesian Inference.

2. We perform an intrinsic evaluation of the topic quality which indicated that the latent topic yielded by our proposal outperformed SOTA topic models.

3. We also perform an extrinsic evaluation of the learned topics showing that they can be effectively used to dynamically retrieve demonstration examples for ICL.

2 Related Work

In subsection 2.1, we describe the Bayesian Frameworks for interpreting LLMs. While in subsection 2.2 and subsection 2.3, we discuss recent works on topic modeling and in-context learning, respectively.

2.1 Bayesian Frameworks for LLMs

Most neural networks including LLMs can be interpreted as probabilistic models that predict a categorical distribution given the context. In contrast, than relying solely on a single set of parameters obtained via maximum likelihood estimation (MLE) or maximum a posteriori (MAP), Bayesian methods maintain a posterior over the parameters or latent variables such as topics. This approach yields the posterior predictive distribution, which integrates over all latent variable (e.g., topics) configurations. The Bayesian generative framework offers more interpretability by requiring an explicit specification of the data-generating process through priors and likelihood functions, which allows for a more transparent view of how predictions are made (Afrabandpey et al., 2020; Mihaljević et al., 2021; Xie et al., 2022).

While LLMs do not explicitly learn a latent variable distribution, many recent studies have attempted to interpret LLMs in the context of a Bayesian inference framework. For example, Xie et al. (2022) interpreted the few-shot inference (through in-context learning) of LLMs as performing implicit Bayesian inference over latent concepts. Specifically, during pretraining, the data exposed to the LLMs can be modeled as being generated from a mixture of Hidden Markov Models (HMMs), each representing a different latent concept θ (e.g., topics or tasks). During inference, when the model is given a prompt consisting of in-context examples, it implicitly infers the shared latent concept underlying these examples to generate the appropriate continuation. This process can be viewed analogously to computing the posterior predictive distribution over the latent concepts $p(y|x, \text{prompt}) = \int p(y|x, \theta)p(\theta|\text{prompt})d\theta$. More recently, Dalal and Misra (2024) provides a broader Bayesian framework for understanding general LLM behavior by interpreting the LLM as performing Bayesian inference over multinomial distributions of next tokens, with prior distributions approximated as mixtures of Dirichlet distributions. Specially, they conceptualize the entirety of possible text as a massive multinomial transition probability matrix, considering the pre-trained model as encapsulating prior knowledge about language in the form of token probability distributions, and the prompt as new evidence that updates this prior. In contrast to their work, we conceptualize generation at a more abstract level where we infer the latent topic variables at each step of the generation. Finally, most similar to our work, Wang et al. (2023) extended the prior work of Xie et al. (2022) under the in-context learning framework by simplifying the assumption so that generated tokens are assumed to be conditionally independent of previous tokens such that $p(y|x) = \int p(y|\theta)p(\theta|x)d\theta$, where the latent topic variables θ are learned as a set of new token embeddings using prompt tuning. However, in contrast with all these previous

studies that rely on implicitly learned or partially structured Bayesian mechanisms, our approach explicitly infers a latent variable distribution during the LLM generative process.

2.2 Topic Modeling

Topic modeling is a suite of algorithms used for discovering the latent themes in a large collection of documents. The topics, often conceptualized as a multidimensional distribution over the vocabulary, provide an interpretable representation of the documents useful for downstream applications including text generation (Tang et al., 2019; Yang et al., 2021; Zhang et al., 2022) and content recommendation (Jin et al., 2018; Esmaeili et al., 2019). While conventional approaches have either embraced probabilistic graphical models (Blei et al., 2003) or non-negative matrix factorization (Steyvers and Griffiths, 2007), neural topic models (NTMs) (Miao et al., 2016, 2017; Srivastava and Sutton, 2017) have become the dominant approach by leveraging deep learning architectures such as the Variational Autoencoder (VAE) (Kingma and Welling, 2014) to more efficiently infer parameters through backpropagation to model the latent topics.

Due to their flexibility and efficiency, various methods have been developed to enhance NTMs including the incorporation of external embeddings (Dieng et al., 2020; Bianchi et al., 2021a,b), knowledge distillation (Hoyle et al., 2020), and modifications to the loss function (Ding et al., 2018; Li et al., 2023). More recent works have directly leveraged the strengths of large language models (LLMs) to uncover latent topics, such as using the rich representations of LLM embeddings (Xu et al., 2023), or using the LLM to directly generate the topics (Pham et al., 2024; Mu et al., 2024). However, these works do not incorporate the rich token-level information available at each generation step. In contrast, our approach integrates the topic inference process directly into each generative step of the LLM, resulting in a tighter coupling between language modeling and topic modeling to capture subtle shifts in semantic context as the text unfolds.

2.3 In-Context Learning

In-context learning (ICL) was discovered as an emergent ability of the LLM to improve task performance by adding demonstration examples to the input prompt (Wei et al., 2022). Since model performance can vary widely depending on the choice of in-context examples (Liu et al., 2022), a key challenge is determining an effective strategy to select demonstration examples for any given task. While early studies often investigated how to find the optimal set of examples for a given task (Wei et al., 2022; Lu et al., 2022; Min et al., 2022), a more effective strategy is to dynamically retrieve demonstration examples based on the current task input (Luo et al., 2024). Popular methods include using off-the-shelf encoders (Reimers and Gurevych, 2019; Izacard et al., 2022), or fine-tuning the retrieval model based on the model performance (Shi et al., 2024). More recently, studies have demonstrated the effectiveness of using LLMs' own hidden states for retrieving similar examples (Wang et al., 2024; Li et al., 2025). This is comparable to our approach, where we use the inferred topic distribution of the LLM as an interpretable representation for retrieving similar examples.

3 Our Explicit Bayesian Approach

3.1 Formulation

Motivated by prior works (Xie et al., 2022; Wang et al., 2023), our formulation is also based on the posterior predictive distribution of the LLM generative process from a Bayesian perspective, where the next word w_t is generated by marginalizing over latent topic variables θ (Equation 1). Specifically, at each generation step, the model first infers a posterior over the latent topic θ based on the previous context $w_{1:t-1}$ before predicting the next word w_t by sampling from a mixture of conditional distributions $p(w_t|\theta)$.

$$P(w_t|w_{1:t-1}) = \int_\theta p(w_t|\theta)p(\theta|w_{1:t-1})d\theta \quad (1)$$

In this formulation, the latent topic variables θ serve as an approximate sufficient statistic for the information in the context. By conditioning on θ, the model summarizes all relevant contextual signals into a single latent representation, allowing the model to generate the next token $p(w_t|\theta)$ without needing the context $w_{1:t-1}$.

3.2 Explicit Bayesian Inference

Since LLMs do not explicitly learn an interpretable latent topic distribution, we approximate the posterior using a variational distribution $q_\phi(\theta|w_{1:t-1})$ through the VAE. In this setup, the *encoder* network takes the input context $w_{1:t-1}$ and estimates a distribution $q_\phi(\theta \mid w_{1:t-1})$ over the latent topics,

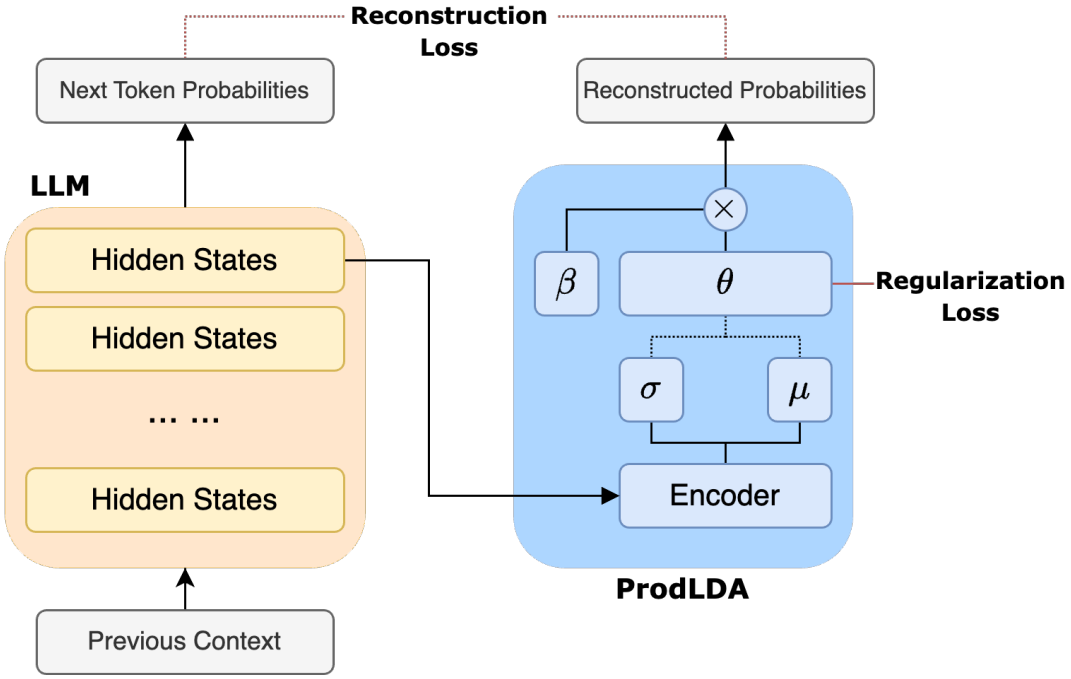

Figure 1: Overview of our proposed method. At each generation step, we use the LLM hidden state representation of the previous context as input to the neural topic model (i.e., ProdLDA) (Srivastava and Sutton, 2017), where the model samples θ based on the distribution parameterized by μ and σ. The decoder then reconstructs the LLM probabilities by computing the weighted mixture of topic distribution θ and the topic-word matrix β.

while the *decoder* reconstructs the next token probabilities of the LLM conditioned on the sampled θ. This can be interpreted as a form of approximated *Explicit Bayesian Inference* over the latent topics θ, where the VAE infers the parameters of the posterior distribution over the topic variable θ from the observed context and marginalizes out θ when predicting the next token (Equation 1).

$$p(w_t \mid w_{1:t-1}) \approx \int_\theta p_\psi(w_t \mid \theta) q_\phi(\theta \mid w_{1:t-1}) d\theta \tag{2}$$

The log-likelihood of the observed data $\log p_\psi(w_{1:t})$ can be optimized using the Evidence Lower BOund (ELBO) (Equation 3).

$$\log p_\psi(w_{1:t}) \geq \mathbb{E}_{q_\phi(\theta|w_{1:t})}\left[\log p_\psi(w_{1:t} \mid \theta)\right] - D_{\mathrm{KL}}\big(q_\phi(\theta \mid w_{1:t}) \,\big\|\, p(\theta)\big) \tag{3}$$

3.3 Our Method

In this work, we adopt the ProdLDA neural topic model (Srivastava and Sutton, 2017), which replaces the standard Gaussian prior of the VAE (Kingma and Welling, 2014) with a Laplace approximation of the Dirichlet prior. Under this architecture, the encoder network outputs a mean μ and standard deviation σ, which parameterize a Gaussian distribution in the latent space. A sample from the distribution is mapped through a softmax function to obtain the topic distribution θ. Finally, the learned word-topic distribution matrix β, which defines the distribution over the vocabulary for each topic, is used to construct the LLM probabilities weighted by θ. The overview of our proposed method is illustrated in Figure 1.

To represent the previous context, we use the hidden state embedding of the LLM as input to the encoder. In addition, rather than reconstructing the bag-of-words (BoW) representation of the document, we define the decoder's reconstruction loss to be the negative log-likelihood (NLL) of the next token predictions to match the output from the LLM. This leads to the final loss presented in Equation 4.

$$\mathcal{L} = -\ \mathbb{E}_{w \sim p_{\mathrm{LLM}}(w_t|w_{1:t-1})}\left[\log\, p_\psi(w_t \mid \theta)\right] + \lambda\, D_{\mathrm{KL}}\Big(q_\phi(\theta \mid w_{1:t-1}) \,\Big\|\, p(\theta)\Big) \tag{4}$$

This procedure yields an approximate Bayesian topic model that learns to reconstruct the exact output of an LLM, providing interpretable latent representations θ while maintaining faithfulness to the LLM predictions for the given training corpus.

In practice, due to the large difference in representation capacity between the LLM and our model,

we limit the vocabulary of the topic model to be a much smaller size than the vocabulary of the LLM to ensure stability during training. Since most modern language models adopt the Byte-Pair Encoding (BPE) tokenizer (Sennrich et al., 2016), words are sometimes tokenized into multiple sub-word tokens. To combine the probability of sub-word tokens, we compute the whole-word probability by accumulating the probabilities of all sub-word tokens that form each word through products. This is done by using beam search to keep track of partial word candidates, and maintaining the cumulative probability for each partial sequence.

4 Intrinsic Evaluation on Topic Modeling

Since our method infers the underlying topics during LLM generations, these latent topics can capture the underlying semantic patterns providing some transparency regarding the themes leveraged by the model at each generation step. When trained on a collection of documents, these topics can be interpreted as LLM-enriched thematic structures of the corpus analogous to those found in traditional topic models. In this section, we assess the intrinsic quality of the inferred topics by comparing our method with existing topic modeling techniques.

4.1 Experiment Settings

In our experiments, we use Llama3.2-1B (Dubey et al., 2024) as the base LLM for inferring latent topics. We train and evaluate our models on three publicly available datasets of different domain and number of documents, namely, AG News (Zhang et al., 2015), DBPedia, and GovReports (Huang et al., 2021). For all three datasets, we use a fixed vocabulary size of the top 2000 most frequent words, which is roughly equivalent to the dimension of the LLM hidden states (2048). The statistics of the datasets are presented in Table 1[1].

[1]Hyperparameter settings are described in Appendix Appendix A

Dataset	Domain	Words	Size
AG News	News	31	7,600
DBPedia	Wiki	46	76,000
GovReports	Reports	571	973

Table 1: Statistics of the three datasets used in our experiments.

In our experiments, we use automatic topic evaluation metrics to efficiently assess two important facets of the topics, namely, *coherence* and *diversity*, which are strongly correlated with the intrinsic interpretability of the topics (Dieng et al., 2020). Topic *coherence* measures the extent to which the words within a topic are related to each other in a meaningful way. To assess the *coherence* of the topics, we use the Word Embedding (WE) metric (Fang et al., 2016), which measures the pairwise embedding similarity between words within a topic, and the recently proposed LLM score (Stammbach et al., 2023) which utilizes an LLM to rate word relatedness on a 1-3 scale and has demonstrated the highest correlation with human judgment. In our experiments, we use Word2Vec embeddings[2] (Mikolov et al., 2013) to measure word embedding similarity, and `gpt-4o-2024-08-06` model to rate the topics. To measure the *diversity* between topics, we used Inversed Rank-Biased Overlap (I-RBO) (Terragni et al., 2021b; Bianchi et al., 2021a) measuring the rank-aware difference between all combinations of topic pairs.

4.2 Baselines

We compare our proposed method with both traditional and recent topic modeling techniques, including: (1) LDA (Blei et al., 2003), (2) ProdLDA (Srivastava and Sutton, 2017), (3) ETM (Dieng et al., 2020), (4) CombinedTM (Bianchi et al., 2021a), and (5) ZeroshotTM (Bianchi et al., 2021b).

4.3 Results

From the results displayed in Table 2, our approach consistently outperforms all baselines across all three datasets. In particular, we find that our method simultaneously improves upon the coherence scores while maintaining or even surpassing the best baselines in terms of topic diversity. Arguably, our proposed method may be benefitting from two possible advantages. The first advantage is the increased number of training examples. While traditional topic models reconstructs the bag-of-words (BoW) representation of the document, our approach allows the model to learn from each generation step of the LLM, which significantly increase the number of training examples that our model can learn from. The second advantage is that since our model learns from the LLM predicted next-token probabilities, it benefits from the

[2]We use `word2vec-google-news-300` from the Gensim library (Řehůřek and Sojka, 2010).

	AGNews			DBPedia			GovReport		
Metrics	LLM	WE	I-RBO	LLM	WE	I-RBO	LLM	WE	I-RBO
$K = 50$									
LDA	1.86	.108	.983	2.10	.123	.952	2.62	.150	.762
ProdLDA	2.30	.147	.984	2.46	.171	**.991**	2.32	.141	.986
ETM	2.38	.193	.940	2.76	.252	.944	2.70	.227	.961
CombineTM	2.40	.189	.972	2.72	.243	.979	2.54	.144	.988
ZeroshotTM	2.44	.162	.903	2.66	.204	.916	2.62	.151	.967
GenerativeTM (Ours)	**2.74**	**.269**	**.991**	**2.78**	**.297**	.989	**2.80**	**.254**	**.993**
$K = 100$									
LDA	1.86	.105	.987	1.99	.120	.959	2.40	.148	.807
ProdLDA	2.24	.136	.982	2.49	.170	.989	2.29	.135	.988
ETM	2.36	.194	.945	2.65	.253	.945	2.61	.232	.966
CombineTM	2.06	.134	.896	2.75	.236	.971	2.58	.140	.981
ZeroshotTM	2.42	.172	.917	**2.78**	.270	.897	2.60	.148	.984
GenerativeTM (Ours)	**2.69**	**.254**	**.989**	2.76	**.301**	**.990**	**2.77**	**.250**	**.989**

Table 2: Topic modeling results for number of topics $K \in \{50, 100\}$ on AGNews, DBPedia, and GovReport. Each result is computed by averaging over 5 random seeds.

rich semantic information embedded in the LLM's learned distributions. In other words, because the LLM tends to assign higher probabilities to terms that are contextually and semantically related, our method naturally clusters these related words together, leading to more coherent and thematically consistent topics. To verify the role played by these two advantages, we perform additional experiments.

4.4 Follow-up Experiments

Input Representation For the results presented in Table 2, we used the final hidden layer (i.e., layer 16) representations of the last context token as the input embedding to our model since they are directly used for predicting the probabilities of the next token. To study the effects of using different input representations on the quality of the topics, we tested the performance of our method when we use different layers of the model as well as other types of embeddings (i.e., SBERT) for representing the context.

From the results shown in Table 3, we see that while using the final hidden layer of the LLM achieve the best topic quality, there is a gradual decrease in performance using the earlier layers of the model. This is expected since the last hidden layer is directly used by the LLM to output the next-token distribution. In addition, we find that although using the SBERT embeddings cannot match the performance achieved by our approach, it still performs better than all the baselines, demonstrating the advantage of leverage the rich semantic information from the LLM predictions.

Input	LLM	WE	I-RBO
SBERT	2.45	.232	.984
Layer 2	2.64	.252	.991
Layer 4	2.66	.255	.990
Layer 6	2.70	.262	**.992**
Layer 8	2.64	.258	.992
Layer 10	2.66	.254	.991
Layer 12	2.66	**.270**	.990
Layer 14	**2.78**	.270	.989
Layer 16	2.74	.269	.991

Table 3: Topic modeling performance on AG News ($K = 50$) using the SBERT embeddings and the hidden states from different layers of the LLM as input.

Number of Training Examples We also perform experiments to measure the effects of the number of training examples on the quality of the inferred topic. In particular, we perform experiments to evaluate how varying the training set size affects the performance of our model. From Table 5, we see

Label	ProdLDA	ZeroshotTM	GenerativeTM
Company	operator, base, network, company, operations, distribution, internet, content, subsidiary, computer	business, firm, investment, bank, countries, offices, funds, banks, companies, businesses	retailer, brand, company, distributor, label, shop, manufacturer, firm, supplier, clothing
Film	comedy, role, director, action, films, movie, roles, cinema, film, feature	life, play, drama, film, screenplay, woman, man, adaptation, plot, title	crime, drama, thriller, romance, comedy, suspense, horror, noir, mystery, fantasy
Mean of Transportation	model, car, brand, generation, manufacturer, line, luxury, store, production, accessories	car, model, railway, unit, type, operator, bus, train, rail, units	sedan, chassis, car, prototype, cars, engine, cockpit, kit, vehicle, model
Natural Place	point, level, island, range, mountain, views, elevation, border, peak, hill	point, range, mountain, peak, border, views, end, pass, level, trail	forests, forest, habitat, habitats, woodland, scrub, destruction, vegetation, soils, grass
Written Work	volume, issue, authors, magazine, science, anthology, trade, edition, aspects, review	field, aspects, research, journal, editor, behalf, chief, access, review, peer	book, novel, novels, chapter, books, tale, trilogy, memoir, poem, manga

Table 4: Top-10 words for each topic generated by ZeroshotTM, ProdLDA, and our GenerativeTM model.

that our model performs on-par with the baseline (zeroshotTM) even with a fraction of the training documents (1000). This advantage can be seen from the number of LLM prediction steps that our model learns from. Since topic models often struggle with sparse targets often present in shorter texts (Qiang et al., 2022), our approach can yield a significant advantage by increasing the number of training instances and leveraging the rich semantic information of LLMs.

Documents	Steps	LLM	WE	I-RBO
1000	6,466	2.44	.225	.982
2000	12,026	2.58	.244	.988
3000	17,704	2.58	.244	.984
4000	23,340	2.64	.256	.990
5000	28,769	2.72	.266	.986
6000	34,310	2.68	.255	.990
7600	43,117	**2.74**	**269**	**.991**

Table 5: Topic modeling performance on AG News ($K = 50$) using the different number of training documents.

4.5 Qualitative Comparison

For a qualitative comparison, we choose three neural topic models: ProdLDA, ZeroshotTM, and our GenerativeTM, all sharing the same architecture and training configurations as the standard ProdLDA. In order to compare between topics, we choose the DBPedia dataset with ground-truth label and use OpenAI's `GPT-4o` to assign each topic to one of the 14 ground truth classes using the following prompt.

> `You will be given a topic represented as a plain list of words.`
>
> `Choose` **`exactly one`** `class from the DBpedia-14 ontology that best matches the topic.`
>
> `Allowed classes:` <DBPedia classes>
>
> `Topic:` <List of Topic Words>

To select the topics for visualization, we display the top-10 words by selecting 5 classes and taking the most coherent topic according to LLM and WE score. In Table 4, we see that the topics from GenerativeTM consists of semantic similar words that can often be used interchangeably. This is expected behavior since our method uses the next-token distribution as the reconstruction target, where words that are likely to occur under similar context. On the other hand, the topics from the standard ProdLDA and ZeroshotTM are learned from reconstructing BoW document representation, which captures surface-level patterns and can be misled by corpus-level noises. For example, for the topic inferred by ZeroshotTM, "*aspect*" and "*behalf*" might be commonly co-occurring with other topical words in documents from the "Written Work" class (last row), they are poor semantic descriptors of the topic, making the topic less interpretable overall. We leave more sophisticated

strategies for learning topics from the deeper meanings of text as an exciting venue for future work.

5 Extrinsic Evaluation on In-Context Learning

While the experiments in section 4 provided an intrinsic evaluation of the topic quality, we now perform an extrinsic evaluation by assessing the usefulness of the topics for in-context learning (ICL). In particular, during inference on the test set, we use the topic distribution of the input documents to dynamically retrieve the examples with the most similar topic distribution in the training set.

5.1 Experiment Setting

We perform experiments on two popular NLP tasks, classification and summarization. For classification, we use the DBPedia-14 dataset previously used for topic modeling, where each document is classified into one of the 14 non-overlapping categories, e.g., *Company*, *Animal*, *Plant*. For summarization, we use the XSum dataset (Narayan et al., 2018), which consists of $204K$ training and $11K$ test examples. Each example in XSum contains a news article paired with a single one-sentence summary extracted from the title.

For both tasks, we use the instruction-tuned Llama3.2-1B-Instruct, where the base model is fine-tuned on curated datasets of instruction–response pairs. In particular, we use the entire document as input context and use the mean topic distribution θ as retrieval representation of the example. Since our neural topic model is trained to reconstruct the probabilities of the next word in our vocab, we use a manually designed prompt that instructs the model to summarize the entire document in one word (Table 6).

In our experiments, we first compute the retrieval representation of all examples in the training and test set using the respective prompts in Table 6. During inference, we select the top training examples with the most similar topic distribution (measured by KL divergence) and use the input-target pairs as few-shot demonstrations. For DBPedia-14, we use 5-shot inference and evaluate the performance with classification accuracy. For XSum, we use the ROUGE metric (ROUGE-1, 2, LSum) (Lin, 2004) for summarization evaluation, which computes the token overlap between predicted and reference summary, where we perform 1-shot learning with only a single demonstration example since we find that adding more examples resulted in diminished performance.

DBPedia-14
Retrieval Prompt: Using a single word, output the most likely topic of the Wiki article. **Task Prompt**: Classify the document into one of the following categories by outputting the category name ONLY: Company EducationalInstitution ...
XSum
Retrieval Prompt: Summarize the news article in a single word. **Task Prompt**: Generate a single sentence summary for the topic of the news article.

Table 6: The *Retrieval Prompt* for creating the retrieval representation of the document and the *Task Prompt* for task inference.

5.2 Baselines

For baselines, we include zero-shot inference without any demonstrations, as well as the LLM hidden state representation (input to our NTM) which has shown to be a strong baseline for retrieving demonstrations (Wang et al., 2024; Li et al., 2025). Lastly, we also include the next token probabilities of the LLM, which is also significantly more computationally intensive compared to our approach due to the large vocab size of the LLM. We use cosine similarity and KL divergence as the similarity metric for the LLM hidden states and next token probabilities respectively.

5.3 Results

From the results in Table 7, we see that our approach significantly outperforms all baselines on the DBPedia-14 dataset for classification. Since for classification tasks, the demonstration examples often bias the prediction to output the most frequent class (Min et al., 2022), our method was able to consistently retrieve examples with the same ground-truth label. While the next-token probability distribution of the LLM provides a more granular representation of the document topic (instructed by the retrieval prompt), the extremely large vocab

	DBPedia-14	XSum		
	Accuray	ROUGE-1	ROUGE-2	ROUGE-LSum
Zeroshot	52.64	16.93	3.96	14.50
Hidden State	53.14	17.47	4.01	14.73
Probabilities	66.07	17.78	**4.24**	**15.09**
Topic (K = 50)	73.21	17.69	4.08	14.90
Topic (K = 100)	**74.00**	**18.00**	4.12	14.85

Table 7: In-context learning results using different demonstration retrieval methods on the DBPedia-14 (classification) and XSum (summarization) dataset. The results using our approach (Topic) are averaged over 5 random seeds.

size can cause the distribution to be dominated by other unimportant tokens. These unimportant and rarely used tokens still contribute to the computation in noisy or unhelpful ways, not to mention to the added computational overhead of computing similarity between large matrices. In contrast, the topic distribution can be interpreted as a coarse summary of the next-token distribution by semantically partitioning the vocab space into more meaningful concepts. This allows our approach to retrieve examples focusing on the high-level themes of the document while mitigating noise from low-frequency tokens. For the results on the XSum summarization dataset, we see that using the next-token probabilities performs similarly to our topic distribution approach. In contrast to the classification task, where the categories can be often be encoded in a few salient token, summarization requires the model to generate a broader and contextually consistent text. Therefore, infrequent tokens have less sway in determining the overall meaning of the generated summary and have a less impact on the retrieval representation for ICL. Summaries are generated based on the broader context rather than any individual token, so occasional rare tokens introduce far less noise in identifying and retrieving relevant examples for the summarization task.

6 Conclusion

In this work, we propose a novel approach for inferring the latent topics of Large Language Models under explicit Bayesian framwork. By training a variational autoencoder (VAE) to reconstruct the LLM's predicted next-token distribution, we obtain interpretable topic variables that capture high-level thematic patterns at each generation step. Our experiments demonstrate that this topic-based representation provides both intrinsic benefits—yielding more coherent and diverse topics than standard topic models—and extrinsic advantages by improving retrieval-based in-context learning (ICL), particularly in classification tasks where rare-token noise can be detrimental. For abstractive summarization, the coarse topic distribution performs comparably to the next-token distribution approach but at a fraction of the computational cost.

These results suggest that modeling generation with a learned topic distribution can serve as a powerful lens for understanding LLM behavior, bridging probabilistic generative models and large-scale neural networks. In future work, we plan to extend our approach by designing more capable probabilistic models that capture a large proportion of the LLM output space, and investigate ways to additional methods to utilize the latent topics in downstream applications (e.g., hallucination detection, uncertainty quantification, etc.). We hope this work inspires further research on incorporating explicitly modeled latent structures into large language models for enhanced transparency, and task performance.

Limitations

While our method provides an interpretable latent topic space and demonstrates advantages for downstream tasks, it also faces several limitations. First, our approach restricts the vocabulary size in the neural topic model to a much smaller subset than the LLM's full vocabulary. This constraint helps maintain stable training and manage computational requirements but may overlook nuances encoded in less frequent or domain-specific terms. Second, our experiments focus mainly on English text. Given that multilingual LLMs handle linguistic features such as morphology, syntax, and vocabulary differently across languages, directly extending our approach to non-English contexts may require language-specific adaptations, especially for languages that exhibit significant subword overlap or complex morphological structures.

References

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. 2018. Discovering interpretable representations for both deep generative and discriminative models. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 50–59. PMLR.

Homayun Afrabandpey, Tomi Peltola, Juho Piironen, Aki Vehtari, and Samuel Kaski. 2020. A decision-theoretic approach for model interpretability in bayesian framework. *Machine learning*, 109(9):1855–1876.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. 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, Online. Association for Computational Linguistics.

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

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Siddhartha Dalal and Vishal Misra. 2024. Beyond the black box: A statistical model for llm reasoning and inference. *arXiv preprint arXiv:2402.03175*.

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.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836, Brussels, Belgium. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Babak Esmaeili, Hongyi Huang, Byron Wallace, and Jan-Willem van de Meent. 2019. Structured neural topic models for reviews. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3429–3439. PMLR.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, page 1057–1060, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh,

Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary

DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving Neural Topic Models using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1752–1771, Online. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Mingmin Jin, Xin Luo, Huiling Zhu, and Hankz Hankui Zhuo. 2018. Combining deep learning and topic modeling for review understanding in context-aware recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1605–1614, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Raymond Li, Yuxi Feng, Zhenan Fan, Giuseppe Carenini, Weiwei Zhang, Mohammadreza Pourreza, and Yong Zhang. 2025. DeTriever: Decoder-representation-based retriever for improving NL2SQL in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8173–8183, Abu Dhabi, UAE. Association for Computational Linguistics.

Raymond Li, Felipe Gonzalez-Pizarro, Linzi Xing, Gabriel Murray, and Giuseppe Carenini. 2023. Diversity-aware coherence loss for improving neural topic models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1710–1722, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR.

Yishu Miao, 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, New York, New York, USA. PMLR.

Bojan Mihaljević, Concha Bielza, and Pedro Larrañaga. 2021. Bayesian networks for interpretable machine learning and optimization. *Neurocomputing*, 456:648–665.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large language models offer an alternative to the traditional approach of topic modelling. In *Proceedings of the 2024 Joint International*

Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 10160–10171, Torino, Italia. ELRA and ICCL.

Siddharth N, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. 2017. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2022. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–

50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. 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 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*, pages 439–460. Psychology Press.

Hongyin Tang, Miao Li, and Beihong Jin. 2019. A topic augmented text generation model: Joint learning of semantics and structural features. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5090–5099, Hong Kong, China. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021b. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*, pages 33–45. Springer.

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. *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, Vedanuj Goswami, 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. *Preprint*, arXiv:2307.09288.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 15614–15638. Curran Associates, Inc.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Weijie Xu, Wenxiang Hu, Fanyou Wu, and Srinivasan Sengamedu. 2023. DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9040–9057, Singapore. Association for Computational Linguistics.

Yazheng Yang, Boyuan Pan, Deng Cai, and Huan Sun. 2021. Topnet: Learning from neural topic model to generate long stories. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1997–2005, New York, NY, USA. Association for Computing Machinery.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yuxiang Zhang, Tao Jiang, Tianyu Yang, Xiaoli Li, and Suge Wang. 2022. Htkg: Deep keyphrase generation with neural hierarchical topic guidance. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1044–1054, New York, NY, USA. Association for Computing Machinery.

A Hyperparameters

For all our experiments, we keep the hyperparameter values constant without performing any searches. We use a two layer 200 dimensional MLP as the encoder with a batch size of 64 and a learning rate of 2e-3. For all baselines, we use the implementation from the OCTIS library (Terragni et al., 2021a), where we train all baselines for a total of 50 epochs, while training our model for 20 epochs due to the increased number of training examples.

All experiments are performed on single A100 GPU with 40GB VRAM. All topic models take less than 10 minutes to train.