Chapter 5: Archivist: Incorporating the World Knowledge of Neural Language Models into Topic Models as a Bayesian Prior

5.1 Introduction

Despite the rapid progress of neural approaches to text modeling, probabilistic topic models such as Latent Dirichlet Allocation (LDA) remain widely used in both academic research and applied domains. Their appeal lies in interpretability and the generative formalism: topic models provide explicit document–topic and topic–word distributions that can be directly inspected, manipulated, and incorporated into downstream pipelines. These distributions offer a structured view of corpora that remains unmatched in transparency, even as large language models (LLMs) dominate many other areas of natural language processing. However, probabilistic models also exhibit well-known limitations: poor stability on small or noisy corpora, difficulty capturing semantic similarity beyond word co-occurrence, and slower convergence when compared to representation-learning methods.

At the same time, transformer-based LLMs such as BERT have demonstrated remarkable capacity for encoding semantic and contextual information in dense embeddings. These models capture fine-grained word and sentence relationships that traditional co-occurrence-based topic models often miss. Yet, they are less directly interpretable: embeddings are powerful for tasks like retrieval or classification, but they do not naturally produce the explicit multinomial distributions over topics and words that make probabilistic models so useful for exploratory analysis.

The **Archivist** framework seeks to combine these complementary strengths. We fine-tune BERT to predict both document–topic and topic–word distributions, treating them not as final outputs but as *priors* for collapsed Gibbs sampling. In this way, BERT functions as an informed guide that biases the sampling process toward semantically meaningful regions of the distributional space, while the underlying probabilistic machinery ensures interpretability and adherence to the generative assumptions of topic models. This hybrid approach integrates the representational richness of neural language models with the structured transparency of Bayesian inference.

Crucially, Archivist is not positioned as a replacement for either class of models, but as a bridge. For practitioners who continue to rely on probabilistic topic models, Archivist provides a principled way to incorporate the semantic knowledge of LLMs without discarding the familiar probabilistic framework. For researchers in neural topic modeling, it offers a path toward stabilizing and improving inference using pretrained language models while retaining explicit distributions over topics and words. The result is a system that demonstrates improved perplexity and coherence, greater stability across runs, and better alignment with human judgments of topic quality. A particularly important application is in low-resource scenarios, where analysts may only have a few dozen or hundred documents per topic. In such cases, probabilistic models become unstable, while Archivist leverages BERT priors to guide inference toward meaningful categories.

In the sections that follow, we first review the challenges of integrating neural priors into probabilistic inference. We then describe the Archivist architecture and training regime, highlighting how document-level and word-level fine-tuning objectives are aligned with topic-modeling distributions. Finally, we present experimental results comparing Archivist against baseline LDA and neural topic models, and discuss its advantages in both automatic and human-centered evaluations.

5.2 Background

The trajectory of topic modeling has mirrored developments in natural language processing more broadly: from early probabilistic models, to neural models, and more recently to approaches that attempt to integrate contextual embeddings from large language models (LLMs). Yet despite years of innovation, probabilistic models such as LDA [115] remain central to practice in computational social science, digital humanities, and related fields. Their persistence stems from interpretability: topic models produce explicit distributions over topics and words that can be directly inspected and used in downstream analysis.

By contrast, neural topic models promised improved expressiveness and representation power but have often failed to surpass probabilistic methods in practice. **(author?)** [116] demonstrate that traditional LDA with Gibbs sampling outperforms a suite of neural topic models in stability (intra-coder reliability) and alignment (inter-coder reliability) for content analysis tasks.

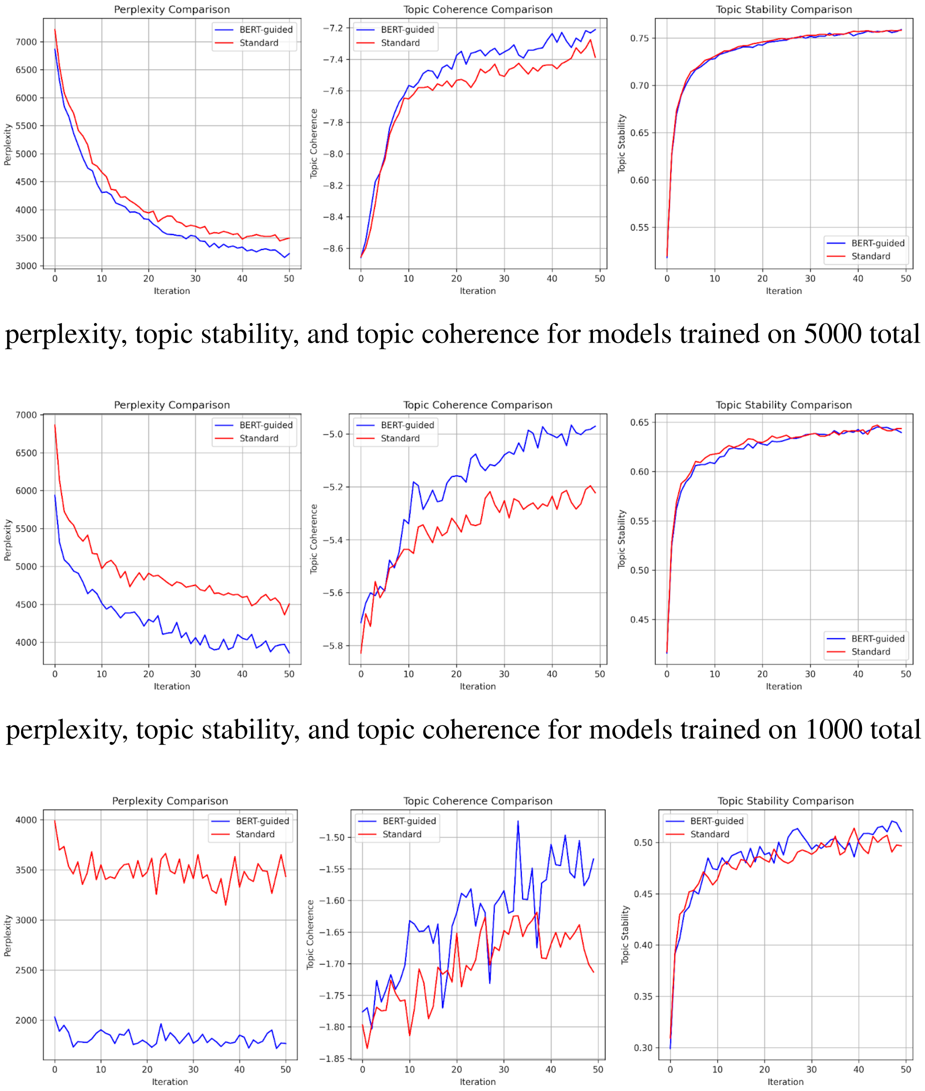
Similarly, **(author?)** [117] show that clustering sentence embeddings can yield more coherent and diverse topics than many dedicated neural topic models. These findings highlight a paradox: while neural models capture richer semantics, they often do not translate into better human-interpretable topics. This discrepancy is partly attributable to evaluation incentives. Automated metrics such as topic coherence have long been the standard for benchmarking, but they are weakly correlated with human judgments of interpretability [96, 118, 119]. Neural topic models frequently achieve superior coherence scores while still producing less interpretable topics for end users. More recent evaluation strategies, such as topic intrusion tasks [120], provide better alignment with human judgment, but they are not yet the norm.

LDA itself exemplifies the probabilistic paradigm: it assumes a generative process where each document is represented as a mixture of latent topics, and each topic is a distribution over words [48, 121]. Its Bayesian framework allows the incorporation of informed priors, which has been leveraged in interactive topic modeling to incorporate dictionaries [122], domain-specific lexicons [52], or organizational constraints [?]. However, these methods largely rely on user-supplied knowledge and do not exploit the massive linguistic and world knowledge encoded in pretrained LLMs. LLMs such as BERT [?] encode deep semantic knowledge that bag-of-words models cannot capture: they disambiguate word senses, recognize paraphrases, and model subtle contextual shifts. Prior work has attempted to exploit these strengths by using contextualized embeddings as inputs to neural topic models [? ?] or by clustering BERT-derived representations [?]. While these methods improve topic discovery, they often discard the transparent probabilistic structure that makes models like LDA appealing.

Archivist seeks to address this gap. By injecting BERT-informed priors directly into the probabilistic framework, it combines the interpretability of LDA with the semantic richness of LLMs. In the following section, we describe how this integration is operationalized.

5.3 Methods

Archivist fine-tunes BERT to predict topic–word and document–topic distributions, which are then incorporated as priors into collapsed Gibbs sampling. This design allows the model to leverage contextual knowledge from BERT while maintaining the generative transparency of probabilistic inference.



(a) Held-out perplexity, topic stability, and topic coherence for models trained on 5000 total documents.

(b) Held-out perplexity, topic stability, and topic coherence for models trained on 1000 total documents.

(c) Held-out perplexity, topic stability, and topic coherence for models trained on 100 total documents.

Figure 5.1: Comparison of Archivist, Gibbs sampling, and BERTopic across dataset scales. Each subfigure reports three metrics (held-out perplexity, topic stability, and topic coherence), showing that Archivist consistently outperforms baselines in both low-resource (100 documents) and larger-scale settings (1000, 5000 documents).

5.3.1 Dataset Construction

To generate supervision signals for fine-tuning, we first train LDA models with collapsed Gibbs sampling across multiple datasets, varying the number of topics K to expose BERT to different granularities of topical structure. From these models, we extract document–topic distributions, topic–word distributions, and per-token topic assignments. These outputs form two supervision datasets: one for training BERT to predict topic–word probabilities, and one for predicting document–topic distributions.

5.3.2 Fine-tuning for Topic–Word Distributions

For topic–word prediction, input sequences take the form:

$$[CLS] document [SEP] topic$$

where the topic is represented by its top words under the LDA model. The document is tokenized and truncated to fit BERT's input constraints. The final-layer token embeddings are projected through a shared linear layer with a sigmoid activation, yielding probabilities that each token belongs to the given topic. Supervision is derived from LDA assignments: tokens sampled into the topic receive target label 1, all others 0. Binary cross-entropy loss encourages BERT to align contextual usage with topical membership.

5.3.3 Fine-tuning for Document–Topic Distributions

For document–topic prediction, input sequences are constructed as:

$$[CLS]document[SEP]topic_1...[SEP]topic_K$$

Each topic is represented by its top words, separated by [SEP] tokens. This sequence allows BERT to jointly encode the relationship between the document and all topics. A linear layer projects the final representation into a K-dimensional vector, normalized with a sigmoid to approximate the document–topic distribution. Supervision is provided by the LDA-inferred distribution, simplified to the top five topics per document with small uniform mass over the remainder to encourage calibration.

5.3.4 Informed Priors for Gibbs Sampling

Standard LDA assumes symmetric Dirichlet priors. Archivist replaces these with BERT-informed priors at both the topic–word and document–topic levels:

Topic–word priors: $\beta_{k,v}$ encodes the preference of topic k for term v, informed by BERT predictions.

Document–topic priors: $\alpha_{d,j}$ encodes prior knowledge about how much document d should prefer topic j.

Incorporating these priors modifies the Gibbs sampling update equations. For topic–word priors:

$$P(z_i = k) = \frac{n_{d,k} + \eta_k}{\sum_i^K n_{d,i} + W\eta} \frac{(v_{k,w} + \lambda_w) * \phi_i}{\sum_i^K v_k + T\lambda} \tag{5.1}$$

,

and for document-topic prios the Gibbs sampling is,

$$P(z_i = k) = \frac{(n_{d,k} + \eta_k) * \alpha_i}{\sum_i^K n_{d,i} + W\eta} \frac{v_{k,w} + \lambda_w}{\sum_i^K v_k + T\lambda} \tag{5.2}$$

where $n_{d,k}$ is the number of words in document d assigned to topic k, η_k and λ_w are Dirichlet hyperparameters, and $v_{k,w}$ is the count of word w in topic k. After inference, final distributions are obtained as:

$$\theta_{d,j} = \frac{v_{k,w} + \lambda_w}{\sum_i^K v_k + T\lambda} \tag{5.3}$$

And our final topic-word distribution,

$$\beta_{d,j} = \frac{n_{d,k} + \eta_k}{\sum_i^K n_{d,i} + W\eta} \tag{5.4}$$

Through these modifications, Gibbs sampling is guided toward assignments consistent with the contextual knowledge encoded in BERT while preserving the interpretability and formal guarantees of the probabilistic framework. These priors are especially valuable when data is sparse. Standard Gibbs sampling must rely solely on observed co-occurrence patterns, which are unreliable at small scales. In contrast, Archivist incorporates pretrained semantic knowledge, providing a stable foundation that accelerates convergence and improves topic interpretability even from the earliest iterations.

5.4 Experimental Results

We evaluate the performance of Archivist, our hybrid topic model that incorporates BERT-generated priors into collapsed Gibbs sampling. We compare against a standard Latent Dirichlet Allocation (LDA; **(author?)** 115) baseline using collapsed Gibbs inference. Experiments are conducted across multiple dataset sizes, ranging from 20 to 1000 documents per topic, using subsets of the 20 Newsgroups corpus with $K = 20$ topics. Preprocessing includes stopword removal, lemmatization, and lowercasing to ensure consistent input representations. Evaluation spans four metrics: **(i)** perplexity, a measure of held-out likelihood; **(ii)** topic coherence, which quantifies semantic interpretability of topic-word distributions; **(iii)** topic stability, which measures reproducibility of topics across iterations; and **(iv)** word intrusion accuracy, a proxy for human interpretability. The first three are automated metrics, while word intrusion is assessed using GPT-4 as a scalable proxy for human evaluation. Full quantitative results are presented in Table **??**, with training dynamics shown in Figure **??**.

5.4.1 Automatic Metrics

Across all dataset sizes, Archivist consistently outperforms standard Gibbs sampling on held-out perplexity. The performance gap is especially pronounced in low-resource settings (< 1000 documents per topic), where Archivist generalizes significantly better, demonstrating the effectiveness of its contextualized BERT priors. In these cases, standard Gibbs exhibits substantially higher perplexity, reflecting overfitting and instability in topic–word distributions under sparse conditions. As dataset size increases, the perplexity gap narrows, but Archivist maintains advantages in convergence speed. At 1000 documents per topic, Archivist reaches

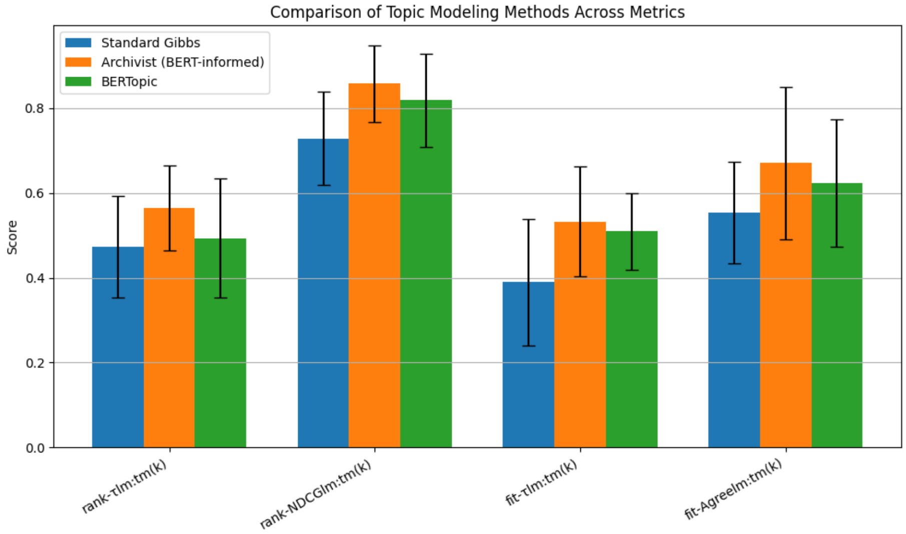


Figure 5.2: ProxAnn evaluation of Archivist, Gibbs sampling, and BERTopic on 1000 documents from 20 Newsgroups. The four categories correspond to correlations between model outputs and GPT-4 proxy judgments: **Doc** ρ **(Fit)** measures alignment of document–topic probabilities with binary fit judgments; **Doc** ρ **(Rank)** measures alignment with document representativeness rankings; **Topic** ρ **(Fit)** assesses how well top words support identifiable categories; and **Topic** ρ **(Rank)** measures the ranking of topics by overall representativeness. Archivist achieves the highest correlations across all four metrics, indicating that its BERT-informed priors yield document–topic and topic–word structures that are more consistent with human-like interpretations than either standard Gibbs or BERTopic.

perplexity and coherence scores comparable to fully converged Gibbs sampling after only 10 iterations, whereas standard Gibbs typically requires 50–100 iterations (Figure 5.1). This acceleration is critical for interactive and low-latency use cases. Topic stability is computed as the average pairwise Jensen–Shannon similarity between versions of the same topic across iterations. Archivist yields substantially more stable topic–word distributions, particularly at small scales, due to its consistent prior guidance. This reduction in variance makes Archivist especially suitable for downstream applications where interpretability and reproducibility are critical (e.g., computational social science).

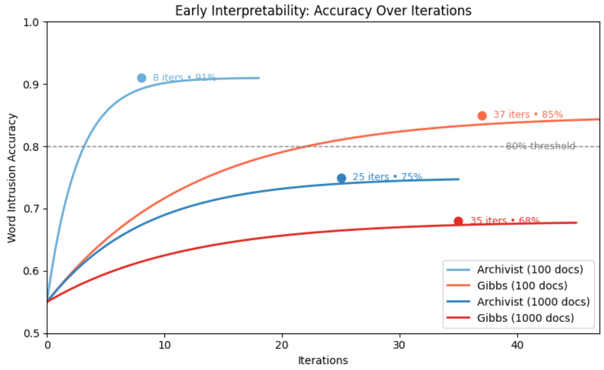


Figure 5.3: Iterations required for Archivist and Gibbs sampling to reach stable word intrusion accuracy on subsets of 20 Newsgroups. Bars show the number of iterations until convergence, with labels reporting both the iteration count and the corresponding accuracy achieved. Archivist consistently requires fewer iterations than Gibbs sampling, particularly in low-resource settings (100 documents), indicating that BERT-informed priors enable the model to surface coherent and interpretable topics much earlier in the inference process.

5.4.2 LLM-Based Evaluation

To further assess semantic coherence, we employ the word intrusion task introduced by **(author?)** [118]. In each trial, we randomly sample a topic, select its five most probable words, and add one intruder word drawn from the bottom 30% of the distribution (ensuring dissimilarity) but with high probability in another topic (avoiding low-frequency noise). The resulting set of six words is presented to the evaluator, who must identify the intruder. Successful detection indicates that the remaining words form a coherent semantic cluster. We employ GPT-4 as an automated and scalable proxy for human judges, following recent work on leveraging LLMs for evaluation tasks. GPT-4 is instructed with the task definition from **(author?)** [118] and prompted as follows:

> *"You are a word intrusion evaluator. You will be provided with a set of topic words, and your task is to choose the word that is the intruder. Please identify just the intruding word."*

Accuracy is computed across 100 randomly sampled topics per model variant. As shown in Table 5.3, Archivist consistently achieves higher accuracy across all dataset sizes. Notably, with 1000 documents per topic, Archivist attains an intrusion accuracy of 91%

These findings confirm that Archivist produces topics that are both more semantically coherent and more readily interpretable by a large language model. Together with improved perplexity and stability, this suggests that BERT-informed priors guide the sampler toward topic structures that align better with human intuitions, particularly in low-resource regimes or scenarios requiring rapid convergence.

5.4.3 User-Oriented Evaluation with ProxAnn

While automated metrics such as perplexity and topic coherence are widely used in topic modeling, they often correlate weakly with human judgments of quality [118**?**]. To complement our automatic and intrusion-based evaluations, we adopt ProxAnn [**?**], a use-oriented evaluation protocol designed to approximate how topic models are actually used in practice. ProxAnn emulates the process of qualitative content analysis (QCA; **?**) in which analysts first infer a category from representative texts and keywords, and then apply this category to new items. The evaluation proceeds in three steps:

Category Identification. Annotators (or an LLM proxy) are presented with the top keywords and exemplar documents from a topic. They are asked to assign a free-text label that best describes the category. This step assesses whether the topic provides enough signal for a human or LLM to infer a meaningful category.

Relevance Judgment. Annotators are then shown new documents sampled from across the topic's document–topic distribution. For each, they rate whether the document fits the inferred category. This provides a direct test of whether the model's assignments of documents to topics align with human sense-making.

Representativeness Ranking. Finally, annotators rank the evaluation documents by how representative they are of the category. This step captures whether the model's scoring of documents reflects intuitive human judgments of centrality and prototypicality.

The ProxAnn framework thus goes beyond surface-level coherence of top words, evaluating both sides of the topic model output: (i) the interpretability of the topic–word distribution and (ii) the quality of the document–topic assignments. Importantly, the protocol can be scaled using large language models as "proxy annotators," which have been shown to perform comparably to human judges in this setting.

We compared Archivist, standard Gibbs-sampled LDA, and BERTopic on 1000 documents from 20 Newsgroups, using GPT-4 as the proxy annotator for all evaluation steps. Results are summarized in Figure 5.2. Across all four metrics, Archivist outperformed both baselines. Improvements were most pronounced in Doc ρ (Fit) and Doc ρ (Rank), indicating that Archivist's BERT-informed priors produce document–topic assignments that align more closely with LLM (and by extension, human-like) relevance judgments. Standard Gibbs exhibited reasonable topic-level performance but lower document-level agreement, reflecting instability in sparse conditions. BERTopic performed comparatively well in keyword-based evaluations but lagged in document-level alignment, underscoring the limitations of embedding-clustering methods for interpretable content analysis.

The advantage is again most visible at small scales: when only a few hundred documents are available, Archivist aligns more closely with GPT-4 relevance judgments, while Gibbs and

BERTopic diverge. This suggests that Archivist not only converges faster but also discovers relevant vocabulary earlier, when users are most likely to interact with the model.

Overall, the ProxAnn evaluation confirms that Archivist generates topics that are not only statistically stronger (in perplexity and coherence) but also more usable for interpretive analysis, producing category structures and document assignments that align more closely with human intuition

5.5 Conclusion

This chapter introduced Archivist, a hybrid framework that integrates the semantic knowledge of large language models with the interpretability and structure of probabilistic topic models. By fine-tuning BERT to predict topic–word and document–topic distributions and using these as informed priors for Gibbs sampling, Archivist bridges a long-standing gap between neural representation learning and Bayesian generative modeling. The approach preserves the transparency and extensibility of classical topic models while leveraging the contextual awareness of transformer-based language models.

The clearest strength of Archivist lies in scenarios with limited data and tight iteration budgets. By incorporating contextual priors, Archivist consistently produces interpretable topics earlier in the inference process, reducing both the number of documents and iterations required to yield actionable insights. This property directly supports interactive and low-resource applications where traditional models fail.

Additionally, empirical results demonstrate that Archivist consistently improves over standard LDA in terms of perplexity, topic coherence, and topic stability, particularly in low-resource regimes. Moreover, using the word intrusion task with GPT-4 as a proxy for human judgment, we showed that Archivist generates topics that are more coherent and interpretable than those discovered by standard Gibbs sampling. These findings suggest that BERT-informed priors not only accelerate convergence but also guide inference toward semantically meaningful topic structures. Beyond methodological performance, Archivist highlights a broader direction for topic modeling research: rather than discarding probabilistic frameworks in favor of neural alternatives, we can enrich them with knowledge encoded in pretrained models. This fusion offers a practical path forward for domains where interpretability, reproducibility, and user control remain paramount, such as computational social science and digital humanities.