PRISM: PRIor from corpus Statistics for topic Modeling

Anonymous ACL submission

Abstract

Topic modeling seeks to uncover latent semantic structure in text, with LDA providing a foundational probabilistic framework. While recent methods often incorporate external knowledge (e.g., pre-trained embeddings), such reliance limits applicability in emerging or underexplored domains. We introduce **PRISM**, a corpus-intrinsic method that derives a Dirichlet parameter from word co-occurrence statistics to initialize LDA without altering its generative process. Experiments on text and single cell RNA-seq data show that PRISM improves topic coherence and interpretability, rivaling models that rely on external knowledge. These results underscore the value of corpus-driven initialization for topic modeling in resource-constrained settings.
Code will be released upon acceptance.

1 Introduction

Topic modeling is a cornerstone technique in Natural Language Processing (NLP) for uncovering latent semantic structures in text. It infers the thematic composition of a corpus by representing each topic as a probability distribution over words. The versatility of topic modeling is evidenced by its application across a spectrum of disciplines - from analyzing customer feedback in e-commerce platforms to modeling gene expression patterns in biology by treating genes as "vocabulary" and samples as "documents". Such cross-disciplinary utility underscores topic modeling's broad relevance in both applied and scientific contexts.

Since its inception, among topic modeling techniques, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) remains a foundation model, leveraging Bayesian inference to estimate document-topic and topic-word distributions while inspiring numerous generative extensions. Topic modeling research has since evolved along many directions, which can be broadly categorized into two main paradigms. The first follows LDA's corpus-intrinsic approach, relying solely on statistical patterns within the target corpus through methods including graph-based and neural models that enhance semantic representation and topic coherence. The second paradigm incorporates external knowledge—such as pre-trained embeddings from large-scale language models or domain-specific priors—to guide topic discovery beyond the statistical patterns available in the input corpus alone.

Corpus-intrinsic topic modeling offers clear advantages for knowledge discovery in emerging scientific domains, where foundation models are underdeveloped and existing knowledge is often fragmented. In fields like biology, key regulatory genes or functional proteins may be undiscovered or poorly characterized, limiting the reliability of external knowledge sources. While pre-trained models excel in established domains, they often fall short in data-scarce settings. In contrast, corpus-intrinsic methods enable unbiased pattern discovery directly from domain-specific data, supporting systematic exploration in knowledge-limited environments.

In this work, we introduce **PRISM**—a **PRI**or from corpus **S**tatistics for topic **M**odeling—which enhances LDA solely through corpus-intrinsic initialization (Figure 1). PRISM shapes the model's "initial perspective" by deriving informed topic-word distributions from statistical patterns in the data, serving as prior-like guidance, which leads to higher topic quality. Empirical results across five text corpora and a single-cell RNA sequencing dataset show that PRISM significantly improves topic coherence and interpretability, often matching or exceeding externally guided methods.

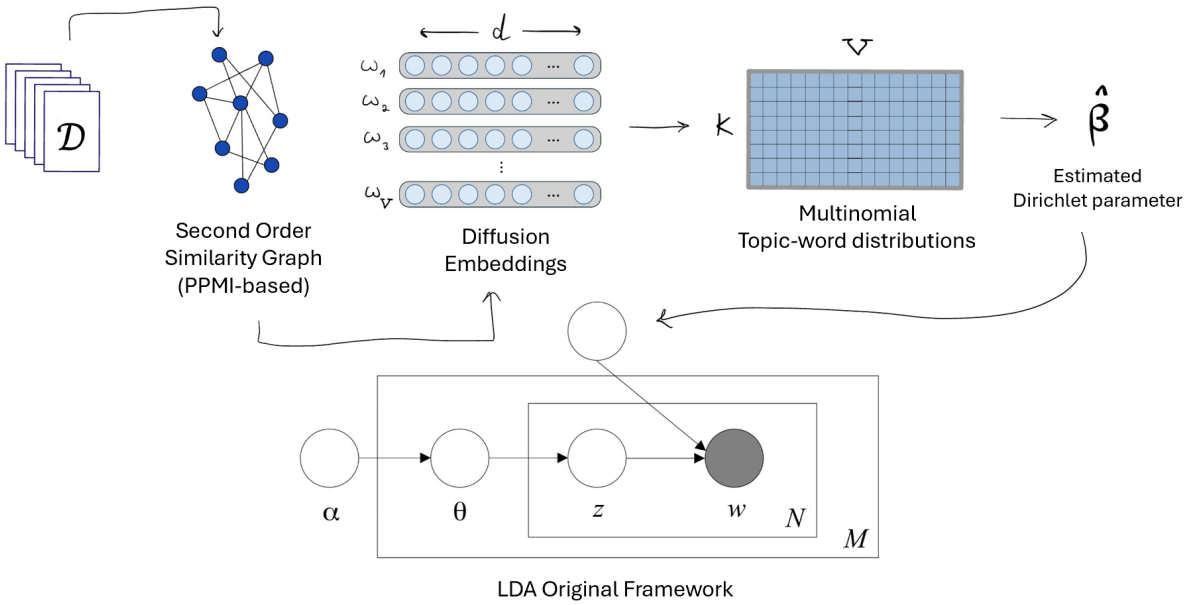

Figure 1: Overview of PRISM framework. Given a corpus $\mathcal{D}$, we construct a second-order word similarity graph using PPMI and cosine similarity. Diffusion maps are applied to obtain low-dimensional word embeddings, which are then soft-clustered into K topics. From the resulting multinomial topic–word distributions, a Dirichlet parameter β is estimated and used to initialize the LDA model without altering it's generative process.

2 Related Work

This section reviews key advancements in topic modeling, focusing on two broad paradigms introduced earlier: (1) methods that incorporate external knowledge, and (2) methods that rely solely on corpus-internal information.

2.1 Methods Using External Knowledge

A prominent line of work enhances topic modeling by integrating external semantic resources. One strategy augments LDA with auxiliary supervision: SeedLDA (Watanabe, 2020) guides topic assignment using predefined seed words, while GHLDA (Yoshida et al., 2023) injects pre-trained word embeddings into the generative process. Another strategy relies on neural architectures built atop contextual embeddings. ETM (Dieng et al., 2020) integrates word2vec-like embeddings into a probabilistic model, whereas CTM (Bianchi et al., 2021) combines transformer-based document representations with variational inference.

Recent approaches, leverage pre-trained semantic representations. While differing in architecture and strategy, all rely on external embeddings to induce topic structure. BERTopic (Grootendorst, 2022) uses sentence-level transformer embeddings (BERT) for document clustering. Top2Vec (Angelov, 2024) embeds both documents and words into a shared semantic space using unsupervised word embeddings. FASTopic (Nguyen et al., 2024) further adopts this paradigm with quantization techniques for scalable inference, while still relying on pre-trained models for semantic guidance.

2.2 Methods Using no External Knowledge

Topic models in this category aim to improve topic modeling while relying exclusively on corpus-internal signals. LDA (Blei et al., 2003), especially when trained with Collapsed Gibbs Sampling (Darling, 2019), remains a strong probabilistic baseline. NeuralLDA (Terragni et al., 2021) and ProdLDA (Srivastava and Sutton, 2017) apply variational inference to LDA, with ProdLDA using a VAE framework for greater scalability and flexibility. Deep NMF (Wang and Zhang, 2021) employs a multi-layer non-negative matrix factorization architecture to capture hierarchical structure in document-term matrices while preserving interpretability. Recently, GINopic (Liu et al., 2024) introduced a graph neural network framework that refines topic assignments by modeling both semantic similarity and document-word co-occurrence within a unified GNN architecture.

These approaches demonstrate the value of corpus-internal signals for enhancing topic models without external supervision. Building on this, PRISM introduces a 'corpus-derived prior' as a

data-driven initialization for LDA—offering a more informed starting point, a prism through which the model better captures semantic structure, while preserving its generative foundations.

3 Preliminaries

This section introduces the key mathematical tools underpinning our approach.

3.1 Latent Dirichlet Allocation (LDA)

LDA (Blei et al., 2003) is a generative probabilistic model in which each document is a mixture over K latent topics, and each topic is a distribution over words. For each document d, topic proportions θ_d are drawn from a Dirichlet distribution with parameter α. Each word w_{dn} is then generated by first sampling a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$, followed by sampling the word from the corresponding topic-word distribution, drawn from a Dirichlet distribution with parameter β.

Inference is commonly performed using *Collapsed Gibbs Sampling* (Darling, 2019), which samples topic assignments z_{dn} based on current token counts and priors - α and β. This method, implemented efficiently in MALLET (McCallum, 2002), is a widely adopted baseline.

3.2 Pointwise Mutual Information and Variants

Pointwise Mutual Information (PMI) (Church and Hanks, 1990) quantifies the strength of association between two words w_i and w_j by comparing their joint probability to the product of their marginal probabilities. To remove noisy or uninformative associations, Positive PMI (PPMI) retains only non-negative values.

Rather than focusing solely on direct co-occurrence, prior work has shown that applying similarity metrics—such as cosine similarity—over PPMI-based word vectors effectively captures second-order semantic similarity (Bullinaria and Levy, 2012; Schütze, 1998). This approach enables the identification of semantically related words that may not co-occur directly, based on the similarity of their distributional contexts. This approach enables the model to estimate semantic relatedness between words that do not co-occur directly by leveraging overlapping associations. For example, while *surgeon* and *physician* may not co-occur, both co-occur with *patient*, *hospital*, and *diagnosis*.

3.3 Diffusion Maps

Diffusion Maps (Coifman and Lafon, 2006) provides a nonlinear dimensionality reduction technique that captures the intrinsic geometry of data represented as a similarity graph. Given an undirected word-word similarity graph W, the method constructs a Markov transition matrix P according the following formulation: $P = D^{-1}W$, where D is the diagonal degree matrix with entries $D_{ii} = \sum_j W_{ij}$. Then, it performs eigen-decomposition over P to obtain the diffusion embedding. Each word is then embedded as a vector:

$$\Psi_t(w_i) = \left(\lambda_1^t \psi_1(i), \lambda_2^t \psi_2(i), \ldots, \lambda_m^t \psi_m(i)\right),$$

where λ_k and ψ_k are the k-th eigenvalue and eigenvector of the transition matrix, respectively; m is the embedding dimension, and t is the diffusion time controlling the decay.

4 Proposed Method: Effective LDA Initialization

We introduce **PRISM**, a corpus-intrinsic method for improving topic modeling by providing a data-driven initialization for LDA. The approach consists of two main stages: (1) constructing word embeddings based solely on corpus statistics, and (2) estimating a Dirichlet parameter β over topic-word distributions derived from these embeddings. The resulting prior is used to initialize LDA, offering a principled way to guide inference without altering the model's generative process.

4.1 Rationale

While contemporary models benefit significantly from pre-trained word embeddings that capture semantic relationships, we sought to harness similar advantages without relying on external knowledge sources. To achieve this, we propose to derive our own word embeddings directly from the target corpus. Our key insight is to derive dense word embeddings from corpus-internal statistics using a semantic similarity graph constructed from PMI variants, followed by diffusion maps to capture global semantic structure embeddings. These embeddings are then softly clustered to produce a probabilistic topic-word distribution, from which we estimate a Dirichlet parameter that guides LDA initialization.

4.2 Constructing Word Embeddings

4.2.1 Similarity Graph

To capture semantic similarity between words, we construct an undirected graph W based on Positive PMI (PPMI), as described in Section 3.2. The PPMI matrix is computed using document-level co-occurrence, treating each document as a context window. Each word w_i is represented by its corresponding row vector $\mathbf{v}_i$ from the PPMI matrix, encoding its distributional context. We then define pairwise similarity using cosine similarity as $W_{i,j} = \cos(\mathbf{v}_i, \mathbf{v}_j)$, where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the PPMI vectors of words w_i and w_j, respectively. The resulting similarity graph W captures both direct and indirect semantic associations by leveraging the principle that words appearing in similar contexts tend to have similar vector representations.

4.2.2 From Graph to Embeddings

To obtain dense word representations from the similarity graph, we apply diffusion maps, as defined in Section 3.3. This spectral embedding technique captures high-order semantic relationships by modeling multi-step transitions over the graph. Intuitively, if a diffusion process is initiated at two semantically similar words, their transition probabilities over time will be similar, reflecting shared contextual neighborhoods.

We embed each word using the top m diffusion components—i.e., the leading eigenvectors of the transition matrix scaled by their corresponding eigenvalues. Empirically, we find that selecting between 80 and 130 components yields the best semantic representation, with the optimal m chosen based on the highest topic coherence score for each dataset and topic configuration.

The resulting vectors serve as our corpus-specific word embeddings, encoding semantic structure without relying on any external resources.

4.3 Estimating Dirichlet parameter β

4.3.1 Topic-Word Distributions

To compute empirical topic-word distributions from the learned word embeddings, we apply a Gaussian Mixture Model (GMM) (Reynolds, 2009) with K components to softly cluster the word representations. The GMM yields the posterior probability $p(z \mid w)$ of each word w belonging to topic z, along with the topic priors $p(z)$.

However, our goal is to obtain $p(w \mid z)$—the probability of a word given a topic—as required to estimate the Dirichlet parameter β over topic-word distributions. To do so, we apply Bayes' Rule:

$$p(w \mid z) = \frac{p(z \mid w)p(w)}{p(z)},$$

where $p(z \mid w)$ is given by the GMM, $p(z)$ is the mixture weight for component z, and $p(w)$ is taken from unigram distribution, computed as the frequency of word w in the corpus divided by the total number of tokens.

This yields multinomial topic-word distributions matrix $\boldsymbol{X} \in \mathbb{R}^{K \times V}$, where each row corresponds to $p(w \mid z)$ for a given topic. The matrix is grounded entirely in corpus-derived signals—capturing both contextual similarity and token-level frequency—used to estimate the Dirichlet parameter β.

4.3.2 Parameter $\boldsymbol{\beta}$ Estimation

To estimate a Dirichlet parameter $\boldsymbol{\beta} \in \mathbb{R}^V$ over the vocabulary, we apply the method of moments (McCallum, 2002), a classical statistical technique used for parameter estimation. The core idea is to match the theoretical moments of the Dirichlet distribution to empirical moments computed from data. Let $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)} \in \Delta^{V-1}$ be k observed samples from a $Dirichlet(\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_V)$ distribution over the $(V-1)-$ probability simplex. The method of moments estimator for β is given by

$$\hat{\beta}_i = \mathbb{E}[X_i] \left(\frac{\mathbb{E}[X_j](1 - \mathbb{E}[X_j])}{\mathbb{V}[X_j]} - 1 \right),$$

where

$$\mathbb{E}[X_i] \approx \frac{1}{k} \sum_{\ell=1}^{k} X_i^{(\ell)},$$

and

$$\mathbb{V}[X_i] \approx \frac{1}{k-1} \sum_{\ell=1}^{k} \left(X_i^{(\ell)} - \mathbb{E}[X_i] \right)^2.$$

4.3.3 Initializing LDA with $\hat{\boldsymbol{\beta}}$

The estimated vector $\hat{\boldsymbol{\beta}}$ is then used to initialize the LDA model, replacing the standard uniform or fixed scalar prior. We modified the MALLET implementation to support a vector valued $\hat{\beta}$ parameter, enabling topic-word distributions to reflect corpus-specific semantic structure from the outset. Apart from this change, the rest of the LDA inference pipeline in MALLET remains unaltered.

5 Experiments

We evaluate the effectiveness of our corpus-informed Dirichlet parameter by assessing its impact on standard LDA. Our goal is to determine whether data-driven initialization can substantially improve topic quality and bring LDA closer to, or even surpass, state-of-the-art topic modeling methods. Experiments are conducted on five diverse text corpora, using three complementary metrics.

Datasets. To ensure a fair and objective evaluation, we use pre-processed datasets from the OCTIS framework (Terragni et al., 2021), thereby avoiding model-specific tuning of the preprocessing pipeline. Specifically, we experiment with four diverse OCTIS datasets: 20NEWSGROUP, BBC NEWS, M10, and DBLP, covering a range of domains and document styles.

In addition, following BERTopic (Grootendorst, 2022), we include the TRUMPTWEETS (TT) dataset to test model performance on informal, short-form text. Since this dataset is not included in OCTIS, we apply OCTIS's preprocessing module with standard, commonly used filtering (see Appendix A for details). This setup allows us to evaluate our method both under standardized preprocessing and in a setting closer to real-world social media text. The statistical details can be seen in Appendix A.

Baselines. We evaluate our approach against a diverse set of topic models. For classical and neural baselines implemented in OCTIS (Terrone et al., 2021), we include LDA (Blei et al., 2003), NMF (Zhao et al., 2017), ProdLDA (Srivastava and Sutton, 2017), NeuralLDA (Srivastava and Sutton, 2017), and the ETM (Dieng et al., 2020). These models rely solely on corpus-internal signals and do not incorporate any external knowledge.

We further compare against recent embedding-based methods that leverage pretrained representations: BERTopic (Grootendorst, 2022), FASTopic (Nguyen et al., 2024), and Top2Vec (Angelov, 2024). These models utilize external knowledge, typically through pretrained sentence embeddings such as MiniLM or the Universal Sentence Encoder. We follow each method's official implementation and recommended configuration as provided in their respective GitHub repositories.

Additionally, we include MALLET (McCallum, 2002), a highly optimized Gibbs-sampling-based LDA implementation, and our proposed model, PRISM. Both were run using the MALLET framework with default hyperparameters, enabling internal parameter optimization (e.g., `optimizeInterval`). For PRISM, we further supplied the model with estimated $\hat{\boldsymbol{\beta}}$ as described in Section 4.

Metrics. We evaluate topic models using statistical and human-aligned metrics that capture different dimensions of quality: c_v *Coherence*, normalized pointwise mutual information (*NPMI*) and the word intrusion detection (*WID*) task. Formal definitions appear in Appendix B. c_v *Coherence* (Röder et al., 2015) and *NPMI* (Bouma, 2009) measure the semantic relatedness of top words within a topic, based on co-occurrence statistics. While both are widely used, we tend to favor c_v due to its empirically stronger correlation with human judgments. *WID* (Chang et al., 2009) evaluates topic interpretability via the identification of an out-of-place word among a topic's top words. To scale this human-centric task, we follow Garg et al. (2023) and use large language models (LLMs) as automated judges. WID implementation details are provided in Appendix B.2.

Setup. To enable fair and robust comparison, we evaluate all models under a unified protocol. For each dataset, we select three values of K near the reference number of ground-truth topics, along with slightly larger values to support finer-grained topic discovery. For TRUMPTWEETS, which lacks labels, we use comparable K values to those in labeled datasets (See Table 4). This protocol reflects a weakly-supervised topic modeling paradigm, in which coarse supervision guides model behavior without enforcing strict topic-label alignment, following the principles outlined by Zhao et al. (2017). All models are evaluated using the top 10 words per topic, and each configuration is run 10 times to account for stochastic variation; final scores are averaged across runs and topic counts.
Models with adaptive topic selection are evaluated accordingly. For BERTopic, we report the best result across runs with and without a fixed K; for Top2Vec, which infers K automatically, we evaluate its output as-is. This Further implementation details appear in Appendix C.

Our Results. We evaluate PRISM across five benchmark datasets, focusing on two key questions: (1) whether it provides consistent improvements over classical LDA as implemented in MALLET,

Model	20NG		BBC News		M10		DBLP		TrumpTweets	
	CV	NPMI	CV	NPMI	CV	NPMI	CV	NPMI	CV	NPMI
ProdLDA	.5976 (.0216)	.0569 (.0057)	.6427 (.0059)	-.0046 (.0110)	.4218 (.0336)	-.0951 (.0954)	.4733 (.0196)	.0072 (.0464)	.5373 (.0099)	.0758 (.0029)
NeuralLDA	.5339 (.0049)	.0425 (.0010)	.5720 (.0234)	-.0532 (.0274)	.4179 (.0183)	-.1950 (.0459)	.3833 (.1022)	-.0207 (.0768)	.4252 (.0159)	-.0249 (.0037)
LDA	.5321 (.0078)	.0457 (.0033)	.4924 (.0150)	-.0409 (.0153)	.3833 (.0116)	-.1602 (.0455)	.3647 (.0148)	-.0273 (.0262)	.4188 (.0196)	-.0109 (.0041)
ETM	.5242 (.0079)	.0433 (.0049)	.4963 (.0062)	-.0176 (.0103)	.3659 (.0150)	-.1255 (.0384)	.2828 (.0581)	-.0389 (.0152)	.4133 (.0338)	.0203 (.0136)
NMF	.5497 (.0087)	.0550 (.0029)	.4564 (.0369)	-.0037 (.0110)	.3536 (.0019)	-.0943 (.0487)	.3663 (.0213)	.0102 (.0188)	.4358 (.0103)	.0326 (.0120)
BERTopic	.5557 (.0155)	.0887 (.0113)	.7124 (.0073)	.1694 (.0022)	.4030 (.0131)	.1310 (.0095)	.3862 (.0075)	-.0039 (.0051)	.4461 (.0025)	-.0287 (.0043)
Top2Vec	.5632 (.0327)	.0682 (.0152)	.5108 (.0484)	-.5750 (.0262)	.3543 (.0033)	-.1675 (.0036)	.3185 (.0005)	-.1328 (.0019)	.3653 (.0028)	-.2349 (.0045)
FASTopic	.5744 (.0258)	.0223 (.0078)	.6572 (.0314)	.0236 (.0136)	.4473 (.0081)	-.2065 (.0172)	.3239 (.0027)	.0012 (.0102)	.3804 (.0243)	-.1319 (.0380)
Mallet	.6381 (.0015)	.1106 (.0081)	.6371 (.0047)	.1128 (.0104)	.4639 (.0080)	.0718 (.0026)	.4394 (.0175)	.0443 (.0082)	.5044 (.0122)	.0721 (.0057)
PRISM (Ours)	**.6581** (.0073)	**.1169** (.0050)	**.6763** (.0149)	**.1309** (.0186)	**.5269** (.0021)	**.0831** (.0128)	**.4638** (.0156)	**.0736** (.0110)	**.5557** (.0046)	**.0952** (.0104)

Table 1: Evaluation of models performance across datasets. Each value shows the avg score over three topic settings. Yellow and orange shadings denote c_v and *NPMI* scores, respectively. Colored values indicate the best-performing model per metric, while lighter shades highlight the second-best. Bolded values indicate cases where PRISM outperforms MALLET.

Model	20NewsGroup	BBC	M10	DBLP	TrumpTweets
ProdLDA	.4561 (.0165)	.4556 (.0681)	.2232 (.0322)	.2528 (.0638)	.2759 (.0722)
NeuralLDA	.2533 (.0518)	.3259 (.0060)	.1194 (.0211)	.1667 (.0667)	.1296 (.0414)
LDA	.1472 (.0202)	.1333 (.3615)	.1139 (.4885)	.0833 (.9211)	.1611 (.5625)
ETM	.3422 (.0301)	.4592 (.0726)	.0889 (.3713)	.1194 (.6224)	.1130 (.9177)
NMF	.1411 (.0196)	.0630 (.3148)	.0010 (.8161)	.0010 (.8161)	.2426 (.3977)
BERTopic	.3811 (.0452)	.5537 (.0092)	.4907 (.0227)	.3584 (.0118)	.3139 (.0103)
Top2Vec	.6022 (.0309)	.5915 (.0618)	.3875 (.0188)	.4928 (.0356)	.3554 (.1016)
FASTopic	.5844 (.0265)	.6315 (.0500)	.3379 (.0442)	.2722 (.0278)	.2352 (.0466)
Mallet	.5744 (.0305)	.4667 (.0330)	.3750 (.0276)	.3111 (.0788)	.2741 (.0226)
PRISM (Ours)	**.6078** (.0374)	**.6137** (.0453)	**.3915** (.0176)	**.3361** (.0742)	**.3028** (.0410)

Table 2: Word Intrusion Detection accuracy across five datasets. Each value reflects the avg accuracy over three topic settings. Yellow shading indicates the best-performing model per dataset, and lighter yellow indicates the second-best. Bolded values highlight cases where PRISM outperforms MALLET.

and (2) how it compares to recent topic models, including both corpus-intrinsic approaches and those that leverage external semantic knowledge. The following results address both aspects. Models above the dashed line are direct baselines; those below use external embeddings and represent an upper bound (Table 1, Table 2).

Quantitative Results. As shown in Table 1, PRISM consistently outperforms the original MALLET implementation, achieving substantial gains in both c_v and *NPMI*. Beyond improving upon MALLET, PRISM frequently closes the gap to, or even surpasses, recent embedding-based methods. This is particularly evident on the BBC News, M10, and TrumpTweets datasets, where PRISM not only significantly outperforms MALLET but also achieves the best or second-best scores across both metrics—demonstrating competitiveness with SOTA methods. PRISM obtains the best c_v score on three out of five datasets (20NG, M10, TrumpTweets), and ranks second on BBC News and DBLP. Given that c_v is widely regarded as more reflective of human topic judgments, these results suggest that PRISM produces more interpretable and semantically coherent topics across diverse domains. While *NPMI* improvements are somewhat more modest, PRISM still achieves the best scores on 20NewsGroup, M10 and TrumpTweets and remains competitive throughout. To further assess topic interpretability, we employ the Word Intrusion Detection (WID) task (detailed in Appendix B.2). As shown in Table 2, PRISM ranks among the top two models on three out of five datasets and remains highly competitive on the remaining two. Among corpus-intrinsic models (above the dashed line), PRISM consistently achieves the highest accuracy—outperforming MALLET and all other classical baselines. It also competes strongly with embedding-based methods (below the dashed line), outperforming all of them on 20NG and several on BBC and M10. Overall, these results show that PRISM not only dominates traditional models in both coherence and interpretability but also matches—and at times exceeds—the performance of state-of-the-art models that rely on external knowledge.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
government	film	win	game	virus
election	good	play	phone	mail
party	award	player	mobile	site
labour	music	match	technology	software
plan	win	game	device	program
tory	show	final	service	security
rise	star	club	music	user
country	include	good	video	attack
tax	actor	back	search	computer
economy	band	team	gadget	net

(a) **BERTopic** Top 10 words.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
mobile	category	bug	rise	election
game	victory	yesterday	quarter	government
phone	aviator	shrink	interest	plan
player	album	navigate	investment	party
technology	film	keynote	rate	labour
firm	goal	fate	growth	issue
service	performance	declaration	analyst	tory
user	great	finnish	figure	leader
music	ball	excitement	price	conservative
system	band	inevitable	stock	public

(b) **ProdLDA** Top 10 words.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
government	film	company	win	technology
election	good	market	game	computer
party	award	firm	play	phone
labour	music	rise	player	mobile
plan	win	sale	good	service
tory	show	price	back	user
law	include	economy	match	game
issue	star	share	team	firm
public	top	growth	club	net
minister	actor	month	final	music

(c) **PRISM** Top 10 words.

Figure 2: Top 10 words per topic over the **BBC dataset** inferred by three models - BERTopic (a), ProdLDA (b), and PRISM (c) - with $K = 5$. Each column represents a distinct topic. Colors denote manually interpreted topic categories: politics, entertainment, business, sports, and technology. Lighter shades indicate weaker relevance to the topic, while white denotes no clear association.

Topic
gene
expression
protein
datum
sequence
cluster
dna
microarray
model
fold

(a) **BERTopic**

Topic	Topic
gene	protein
expression	structure
datum	sequence
microarray	dna
analysis	model
cluster	motif
regulatory	fold
model	prediction
method	algorithm
classification	bind

(b) **PRISM**

Topic	Topic
gene	protein
expression	sequence
datum	model
model	structure
network	dna
cluster	motif
analysis	base
microarray	fold
regulatory	prediction
paper	bind

(c) **MALLET**

Figure 3: Top 10 words in M10 dataset, extracted by (a) **BERTopic**, (b) **PRISM**, and (c) **MALLET**. The topics pertain to biology, with faded shading indicating lower relative importance. BERTopic merges gene- and protein-related terms into a single topic, while PRISM and MALLET separate them into two distinct topics.

Topic
soil
water
crop
yield
climate
irrigation
remote
change
agriculture
production

(a) **BERTopic**

Topic
soil
water
crop
yield
change
area
climate
effect
production
remote

(b) **PRISM**

Topic
water
soil
diesel
change
engine
model
climate
effect
production
fuel

(c) **MALLET**

Figure 4: Top 10 words in M10 dataset, extracted by (a) **BERTopic**, (b) **PRISM**, and (c) **MALLET**. The topic appears to relate to climate and agriculture, with faded shading indicating lower relative importance.

Qualitative Analysis. Over BBC dataset, we show a comparison of PRISM to BERTopic, which achieved the highest c_v and *NPMI* scores, and to ProdLDA, the strongest corpus-intrinsic baseline in c_v (Table 1) over BBC. The dataset contains five ground-truth topic labels. As shown in Figure 2, PRISM successfully recovers all five topics—politics, business, sports, technology, and entertainment—with minimal off-topic words. In contrast, BERTopic fails to recover the business category and exhibits redundancy across two overlapping technology topics. ProdLDA clearly captures politics and business, while entertainment and technology are only partially distinguishable, and sports is entirely missing. These observations align with the WID results (Table 2), where PRISM ranks second overall, while BERTopic and ProdLDA rank fourth and seventh, respectively. This suggests that PRISM produces more coherent and uniquely identifiable topics. A supplementary example is included in Appendix E.

On M10, we compare PRISM to BERTopic and MALLET. PRISM achieves the highest c_v and second-best WID score, while BERTopic leads in WID and *NPMI*, and MALLET ranks second in c_v (Table 1, Table 2). As shown in Figure 3, all models capture biologically meaningful themes, though MALLET includes a few less relevant terms (e.g., "paper," "network"). BERTopic merges gene-related and protein-related concepts into a single broad topic, whereas both PRISM and MALLET separate them into two distinct but related topics, one focused on gene expression and microarray, the other on protein structure and binding. This finer-grained separation reflects PRISM's stronger topical coherence and is also evident in its ranking of more meaningful terms (e.g., "microarray," "regulatory," "structure") with higher probability than MALLET. Interestingly, BERTopic's broader topic structure may contribute to its higher WID score: in the WID task, intruder words are sampled from other high-probability topics (details in Appendix B.2), and when topics overlap semantically, as with genes and proteins, intruders may feel topically adjacent rather than clearly out of place. PRISM's more specific topics make intruder

detection more challenging, which may explain its slightly lower WID despite better topic distinctiveness. Figure 4 presents another M10 topic, likely related to climate and agriculture. PRISM and BERTopic show high overlap in top-ranked words ("soil," "water," "crop," "yield," etc.), with strong topic focus. In contrast, MALLET's output contains generic or loosely related terms ("engine," "fuel," "model"), leading to reduced coherence. These results support the quantitative findings, where PRISM outperforms MALLET across all metrics and approaches the interpretability of BERTopic without using external knowledge.

6 Biological Experiments

Motivation and Analogy. We investigate the applicability of PRISM to biological data, aiming to uncover latent Biological Processes (BPs) from single-cell RNA sequencing (scRNA-seq) data. This task naturally parallels topic modeling: cells correspond to documents, genes to words, and BPs to topics. As in text, where documents often span multiple topics and words can take on different meanings depending on context, each cell may be involved in multiple BPs, and individual genes may participate in several biological functions, reflecting the many-to-many relationships captured by topic models. Furthermore, scRNA-seq data is organized as a count matrix, where each entry denotes the expression level of a gene in a cell, directly analogous to the word-document count matrix in LDA.

Dataset. We evaluate on a scRNA-seq dataset of human breast cancer tissue, generously shared with us by a collaborating research lab in pre-processed form[1].

Baselines. As a proof-of-concept, we compare PRISM to original MALLET, to assess whether our corpus-intrinsic initialization can enhance biological processes interpretability in a biological context.

Evaluation Metric. To assess the biological plausibility of discovered topics, we adopt a GPT-4-based evaluation method inspired by Hu et al. (2025). For each model, we extract the top 20 genes per topic and query GPT-4 to estimate how likely these genes are to co-participate in a known BP, effectively yielding a confidence score. This metric serves as a proxy for the biological coherence of gene sets (details in Appendix D).

[1] https://zenodo.org/records/10620607

Experimental Setup. We run both PRISM and MALLET on the same scRNA-seq dataset (details in Appendix D), using the same configuration across models: 10 runs for each setting of 10, 20, 30 topics. For PRISM, we estimated a β parameter in the same framework as done for textual corpora.

Results. As shown in Table 3, PRISM consistently outperforms MALLET in GPT-4-based confidence scores across all topic settings, indicating improved alignment with known Biological Processes (BPs). Prior work by Hu et al. (2025) has shown that GPT-4 confidence scores correlate with biological plausibility: low-confidence gene sets often failed to correspond to coherent BPs, while high-scoring sets typically aligned with well-established biological functions. Based on this, PRISM's consistently higher scores, suggest that the corpus-derived prior helps steer the model toward more accurate and biologically meaningful topic assignments. Although MALLET also performs reasonably well, PRISM's integration of corpus-intrinsic semantic structure offers a clear advantage in this biological setting.

Table 3: Comparison of GPT-4 confidence scores on the Breast Cancer dataset. Results averaged over 10 runs.

Model	10 BP	20 BP	30 BP
MALLET	.8721 (.0102)	.8831 (.0147)	.8701 (.0213)
PRISM (Ours)	**.9106 (.0138)**	**.9112 (.0114)**	**.8922 (.0207)**

7 Conclusion

We introduced PRISM, a corpus-driven initialization method for LDA that integrates semantic structure derived directly from the data, without relying on external embeddings. Across five diverse datasets—spanning news, social media, and biomedical texts—PRISM consistently improves coherence and interpretability over classical baselines and rivals embedding-based models. Its strong performance, despite operating entirely on corpus-internal signals, highlights the underexplored potential of structure-aware initialization in probabilistic models. This work opens new directions for enhancing topic models through data-intrinsic semantics—bridging the gap between classical transparency and modern representational strength.

Limitations

Our approach requires the number of topics K to be specified a priori, rather than inferred automatically. This design choice, common among classical topic models, necessitates treating K as a tunable hyperparameter. While this may limit full automation and scalability across diverse corpora, our empirical results suggest that PRISM remains robust across a range of K values.

Future Work

Future work could extend the data-driven initialization approach to the α parameter in LDA, which controls the document-topic distribution. Incorporating corpus-based statistical techniques for α may further improve topic sparsity and enhance document-level interpretability.

Acknowledgments

We thank Dr. Shahar Alon (https://www.alonlab.org/) for generously sharing high-quality scRNA-seq data generated using his lab's advanced gene detection and probing techniques.

References

Meta AI. 2024. Llama 3.3-70b instruct. Accessed: 2024-02-10.

Dimo Angelov. 2024. Top2vec at scale: Distributed and interpretable topic modeling using sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. 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 (ACL)*, pages 759–766.

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

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

John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 44(3):510–526.

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

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

Ronald R Coifman and Stéphane Lafon. 2006. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30.

Will Darling. 2019. Latent dirichlet allocation. https://coli-saar.github.io/cl19/materials/darling-lda.pdf. Lecture slides, Computational Linguistics, Saarland University.

Adji Bousso Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling with deep network priors. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 119:5330–5341.

Siddhant Garg, Dinesh Balasubramaniam, Aniruddha Guha, Dipankar Das, and Shubham Jain. 2023. Llm reading tea leaves: Automatically evaluating topic models with large language models. In *Findings of the Association for Computational Linguistics (ACL)*.

Maarten Grootendorst. 2022. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. *arXiv preprint*.

Minji Hu, Safa Alkhairy, Irene Lee, Ryan T Pillich, Daniel Fong, Kevin Smith, Rebecca Bachelder, Trey Ideker, and Daniel Pratt. 2025. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22(1):82–91.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Luyang Liu, Chenguang Zhu, Zhaopeng Tu, Michael Zeng, and Zhiting Hu. 2024. Ginopic: Graph neural inference for probabilistic topic modeling. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Technical software, University of Massachusetts Amherst University of Pennsylvania.

Van-Khanh Nguyen, Viet-Anh Ho, Minh-Le Nguyen, and Akira Shimazu. 2024. FASTopic: A fast and accurate graph-based topic modeling framework with corpus-only statistics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Douglas A Reynolds. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663.

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.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

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

Silvia Terragni, Xuan Tien Doan, Brian Davis, Bedoor AlShebli, Suppawong Tuarob, Pascal Frossard, Javier Borge-Holthoefer, Raheel AlGhamdi, and Peter Gloor. 2021. 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.

Leonardo Terrone, Federico Bianchi, Elisabetta Fersini, and Matteo Palmonari. 2021. 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.

J. Wang and X.-L. Zhang. 2021. Deep nmf topic modeling. *arXiv preprint arXiv:2102.12998*.

Kohei Watanabe. 2020. Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, 14(2):99–113.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Takuya Yoshida, Ryo Hisano, and Takuya Ohnishi. 2023. Gaussian hierarchical latent dirichlet allocation: Bringing polysemy back. *PLOS ONE*, 18(7):e0288274.

Wayne Xin Zhao, Songbo Tan, Rui Zhang, Xueqi Li, Yongdong Zhang, and Ji-Rong Liu. 2017. Weakly-supervised text classification with topic modeling. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, pages 2235–2238. ACM.

A Datasets

As described in Section 5, we evaluate our method on five benchmark datasets. Four of these are standard corpora included in the `octis` library, already preprocessed according to its internal pipeline. In addition, we include the *TrumpTweets* dataset, following its use in prior work by BERTopic (Grootendorst, 2022), to assess performance on short, noisy social media text. For statistics details view Table 4.

Dataset	# Docs	Vocab Size	Avg. Len.	# Labels	Topic Counts (K)
20NG	16,309	1,612	48.0	20	20, 25, 50
BBC	2,225	2,949	120.1	5	5, 10, 15
M10	8,355	1,696	5.9	10	10, 15, 20
DBLP	54,595	1,513	5.4	4	4, 10, 15
TT	18,239	1,988	9.0	–	10, 15, 20

Table 4: Statistics summary of the datasets used in our experiments.

The *TrumpTweets* dataset was obtained from the same source cited by BERTopic[2]. To ensure consistency across datasets, we applied the `octis` preprocessing module with basic filtering settings. The preprocessing configuration was as follows:

```
Preprocessing(
    vocabulary=None,
    lowercase=True,
    remove_numbers=True,
    min_words_docs=3,
    min_chars=3,
    min_df=0.01,
    max_df=0.9,
    max_features=2000,
    remove_punctuation=True,
    lemmatize=True,
    stopword_list="english"
)
```

This preprocessing configuration was selected to balance document retention with vocabulary quality, a trade-off particularly relevant when modeling short texts such as tweets.

B Metrics

We assess topic model quality using both statistical coherence metrics and a language-model-assisted interpretability evaluation. Below we describe each in detail.

[2] https://www.thetrumparchive.com/faq

B.1 Standard Topic Modeling Metrics

We report two standard automated metrics: c_v coherence and normalized pointwise mutual information (NPMI). These corpus-intrinsic metrics assess semantic consistency within topics and lexical distinctiveness across topics.

c_v Coherence. The c_v metric (Röder et al., 2015) combines pairwise NPMI scores with cosine similarity over context vectors derived from a sliding window over the corpus. Formally:

$$C_v = \frac{1}{|W|(|W|-1)} \sum_{i<j} \text{NPMI}(w_i, w_j) \cdot \cos(\mathbf{v}_i, \mathbf{v}_j),$$

where W is the set of top-N words in a topic. Each vector $\mathbf{v}_i \in \mathbb{R}^{|C|}$ encodes co-occurrence statistics of word w_i across the document set C. We compute c_v using `Gensim`'s `CoherenceModel` with default parameters.

NPMI. We also report standalone NPMI (Bouma, 2009), defined as:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)},$$

where $P(w_i, w_j)$ is the empirical co-occurrence probability of words w_i and w_j. This metric normalizes PMI to the range $[-1, 1]$, enabling fairer comparison across corpora.

B.2 Word Intrusion Detection (WID)

To complement statistical metrics with a proxy for human interpretability, we use the Word Intrusion Detection (WID) task. General framework of the metric can be viewed in Figure 5.

The Word Intrusion Detection (WID) task (Chang et al., 2009) is a widely adopted human-centered evaluation method for assessing topic interpretability. In this task, annotators are shown the top-N words of a topic, one of which is an intruder—i.e., a word drawn from another topic that appears with low probability in the target topic but is prominent elsewhere. The annotator is asked to identify the word that does not semantically belong. Higher topic coherence typically results in easier and more consistent intruder identification, making WID an indirect yet effective proxy for human interpretability.

Recent studies have proposed leveraging large language models (LLMs) to automate WID (Garg et al., 2023), enabling scalable, consistent, and human-aligned evaluation. We adopt this paradigm by employing a LLM as an automatic evaluator within our WID framework (see Figure 5). This model is prompted to identify the intruder from each modified topic word list, effectively simulating human judgment without the need for manual annotation.

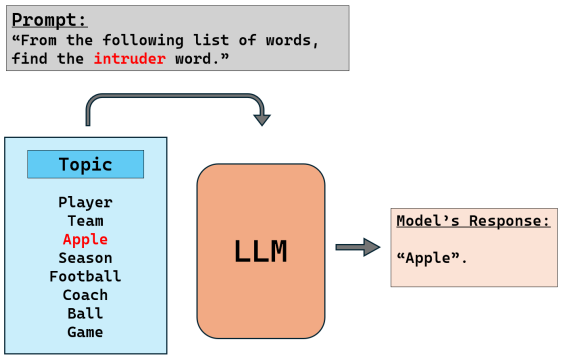

Figure 5: Illustration of the Word Intrusion Detection (WID) framework. A large language model is prompted to identify the word that does not belong in a list of top topic words (a.k.a. the intruder). The prompt shown here is illustrative; actual prompts used in our experiments follow a more structured format.

Pipline. Our pipeline leverages HuggingFace's `transformers` library. We initialize the tokenizer via `AutoTokenizer.from_pretrained`, explicitly setting the end-of-sequence (`eos`) token as the padding token to ensure consistent handling of short text inputs. The LLM is integrated with the `pipeline` API using `device_map="auto"` for efficient hardware mapping and `torch_dtype=torch.bfloat16` to reduce memory overhead.

During inference, each topic's word list is modified by injecting one intruder word. The model is then prompted to identify the semantic outlier. Its success in this task reflects the semantic cohesion of the topic, thus serving as an indirect interpretability metric that complements statistical scores.

The task involves identifying an intruder word inserted into an otherwise coherent topic word list. We evaluate performance using the Meta-LLaMA-3.3-70B-Instruct model (AI, 2024), which demonstrates strong alignment with human judgment.

Prompt Engineering. Inspired by Chain-of-Thought prompting (Wei et al., 2022) and role-play prompting (Kong et al., 2023), we crafted prompts to guide the LLM. The prompt included two exam-

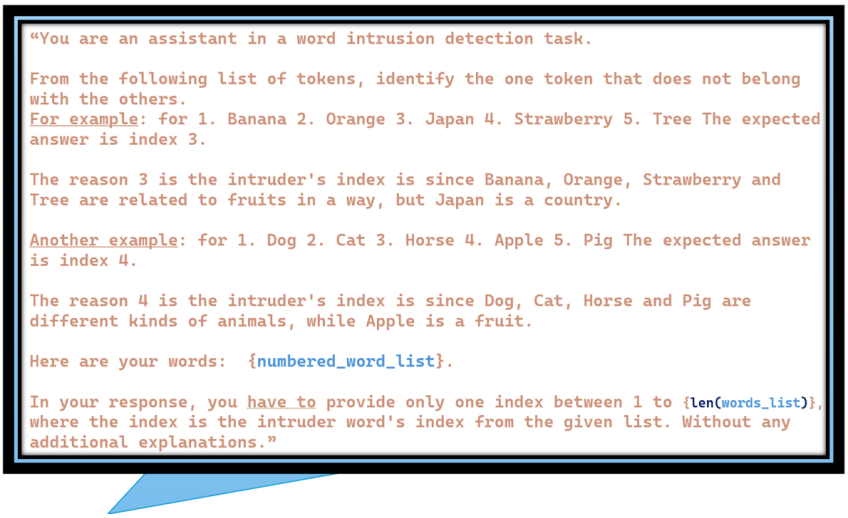

Figure 6: The prompt template used for the Word Intrusion Detection task provides clear instructions, illustrative examples, and a structured response format to assist the model in identifying the intruder word. The input variable `numbered_word_list` is dynamically integrated into the prompt during the evaluation process, enabling the model to process different word sets effectively.

ples to illustrate both the identification of intruder words and the expected response format as can be found in Figure 6.

Evaluation Metric. Accuracy was calculated as the proportion of correct intruder identifications:

$$\text{Accuracy} = \frac{\text{Count}\big(\text{LLM Response} = \text{Real Intruder}\big)}{K},$$

where K is the number of topics. Accuracy was reported separately for top-10, top-15, and top-20 word lists.

Full Pipeline. We present the pipeline for the Word Intrusion Detection task, following the configuration of all necessary settings.

Algorithm 1 Word Intrusion Detection Pipeline

Input: List of topic models $\mathcal{M}$ and Prompt Template
Output: Saved LLM Outputs, Model Accuracy Scores

for each model M in $\mathcal{M}$ **do**
  for each topic T in $M.topics$ **do**
    for each words W in $\{T10, T15, T20\}$ **do**
      $W_{index} \leftarrow$ AddIndices(W)
      $prompt \leftarrow$ FormatPrompt(W_{num})
      $result \leftarrow$ QueryLLM($prompt$)
      SaveOutput(result)
    end for
  end for
end for

EvaluateAccuracy($\mathcal{M}$)

This pipeline evaluates topic coherence by identifying intruder words—terms that do not semantically fit within topic-based word lists—using a LLM. For each topic model, the pipeline processes topics and their top-10, top-15, and top-20 word lists, adding indices to the words (`AddIndices`), formatting them into a structured prompt (`FormatPrompt`), and querying the LLM (`QueryLLM`) to detect the intruder. The model's outputs are saved (`SaveOutput`) and compared against the true intruders to calculate accuracy (`EvaluateAccuracy`), providing a quantitative measure of the topic model's coherence.

C Setup

To ensure a fair and reproducible comparison, we evaluated a diverse set of topic modeling baselines using their publicly available implementations and recommended configurations.

OCTIS Models. We ran the following models through the `octis` framework: `LDA`, `NMF`, `ETM`, `ProdLDA`, and `NeuralLDA`. All models were executed with default hyperparameters as defined in the `octis` documentation, using the library's standardized preprocessing pipeline.

For `ETM`, we evaluated two configurations:

Without pre-trained embeddings : the default `octis` configuration was used.

With pre-trained embeddings : we used GloVe embeddings with the following parameters:

```
ETM(num_topics=TOPICS_NUM,
   embeddings_path="filtered_glove.100d.vec",
    embedding_size=100)
```

This setup allows us to compare corpus-only training versus external semantic initialization. We got similar results with no detect improvement, thus we provided only the version without external knowledge.

BERTopic. We ran BERTopic using its official implementation[3] with default parameters, except for the number of topics. We explicitly set the number of topics to match the experimental setup and report the better result between the auto-detected and fixed-topic configurations.

Top2Vec. We ran Top2Vec in contextual embedding mode with the following configuration:

```
Top2Vec(
    documents,
    split_documents=True,
    contextual_top2vec=True,
    embedding_model="all-MiniLM-L6-v2",
    speed="deep-learn",
    workers=2
)
```

This follows the recommended usage from the official repository[4]. We also experimented with supplying a custom tokenizer, but it did not improve performance; thus, default tokenization was used in all reported results.

FASTopic. We used the FASTopic implementation from `topmost`[5], using the `Preprocess` utility as recommended. Each model was initialized as:

```
preprocessing =
    Preprocess(stopwords="English")
model =
    FASTopic(num_topics=topic_num,
             preprocess=preprocessing)
```

All hyperparameters followed the defaults in the official github[6], and no additional tuning was performed.

D Biological Experiments

Setup for Biological Data Experiments

To enable the use of raw gene expression matrices—analogous to document-term matrices in text—we adapted the MALLET input format to accept direct count data. Specifically, we constructed a serialized input object compatible with MALLET's internal representation, containing both the corpus alphabet (gene identifiers) and the expression counts per sample (document). This object was passed to MALLET via the `inputFile` argument, leveraging native support in the original MALLET codebase.

All models were trained within the same framework, with our estimated $\hat{\beta}$ supplied as an external input. This ensured a consistent inference pipeline across experiments, isolating the effect of our initialization from the generative process or hyperparameter settings.

Biological Evaluation via LLM Confidence

To assess the biological relevance of gene sets derived from topic models, we follow the LLM-based evaluation protocol introduced by (Hu et al., 2025).[7] The core idea is to query a large language model (GPT-4) with each gene set and evaluate whether it can (1) identify a coherent biological process (BP) associated with the gene set, and (2) express high confidence in that association.

Prompt Design. Each gene set is presented in a natural language prompt, instructing the model to infer a shared biological process based on the listed genes. Prompts are carefully crafted to be neutral and avoid leading the model toward specific functions. The model is then asked to (i) name the most likely BP, and (ii) rate its confidence on a scale from 0 to 1.

Scoring. The model's textual output is manually inspected to verify whether the inferred BP matches a plausible biological function supported by external evidence (e.g., GO annotations). Confidence scores are recorded for each gene set and aggregated to assess overall coherence across topics.

Interpretation. As shown in prior work, gene sets yielding low confidence often correspond to functionally inconsistent or noisy groups, whereas high-confidence predictions align with known biological pathways. Thus, the LLM confidence score serves as a proxy for functional coherence and interpretability of the gene sets.

E More Qualitative Findings

To complement the qualitative findings presented in the main paper, we include additional quantitative

[3] https://github.com/MaartenGr/BERTopic
[4] https://github.com/ddangelov/Top2Vec
[5] https://github.com/yfsong0709/TopMost
[6] https://github.com/bobxwu/FASTopic
[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11725441

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
company	technology	law	win	election
market	phone	case	good	government
rise	mobile	court	play	party
sale	game	government	game	labour
firm	service	rule	film	plan
price	music	claim	award	tory
share	user	legal	player	public
growth	computer	charge	back	country
economy	net	police	show	work
month	firm	ban	world	minister

(a) MALLET Top 10 words

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
government	film	company	win	technology
election	good	market	game	computer
party	award	firm	play	phone
labour	music	rise	player	mobile
plan	win	sale	good	service
tory	show	price	back	user
law	include	economy	match	game
issue	star	share	team	firm
public	top	growth	club	net
minister	actor	month	final	music

(b) PRISM Top 10 words

Figure 7: Top 10 words per topic over the **BBC dataset** inferred by MALLET (a) and PRISM (b), with $K = 5$. Each column represents a distinct topic. Colors denote manually interpreted topic categories: politics, entertainment, business, sports, and technology. Lighter shades indicate weaker relevance to the topic, while white denotes no clear association.

analysis for BBC dataset here.

BBC Dataset. Figure 7a (MALLET) and Figure 7b (PRISM) display the top words for each topic on the BBC dataset with 5 topics. While MALLET produces reasonable topics, it redundantly captures politics in two separate themes and fails to isolate the entertainment domain. In contrast, PRISM yields distinct and semantically meaningful topics, effectively covering all major themes in the corpus. These results suggest that while MALLET offers a solid inference framework, our initialization method pushes the model further toward more coherent and semantically distinct topics, indicating that the observed improvements stem from our approach rather than the base model alone.