

Vendi Clustering for Diversity-Preserving Topic Reduction

Sean Kim^{1,3} Adjì Bousso Dieng^{2,3}

¹Brown University ²Princeton University ³Vertaix
sean_s_kim@brown.edu adjì@princeton.edu

Abstract

Embedding-based topic models such as BERTopic routinely produce hundreds of fine-grained, overlapping topics that must be reduced before downstream use. Existing approaches couple reduction with outlier rejection, rely on local geometric linkage, or use costly and stochastic LLM prompting. We introduce Vendi Clustering (VC), a topic reduction method that reframes the problem as global diversity-preserving compression: at each step, VC selects the merge that induces the smallest loss in the Vendi Score, a spectral measure of diversity. A naive implementation of this global criterion is prohibitive at $O(n^6)$; we derive algebraic update rules at order $q = 2$ that evaluate each candidate merge in $O(1)$ via cached row-square sums and a lookahead Gram matrix, reducing total complexity to a tractable $O(n^3)$. On 20 Newsgroups and AG News, VC outperforms standard reduction baselines on coherence and diversity, surpasses LLM-prompted reduction on the majority of metrics against both an open-weight model and a frontier API model, with topic diversity gains of up to 35% while running over $40\times$ faster.

Keywords: Topic Modeling, Diversity, Clustering, Machine Learning, Vendi Score

1 Introduction

Modern embedding-based topic models such as BERTopic (Grootendorst, 2022) routinely produce hundreds of fine-grained, overlapping topics on real-world corpora, requiring a post-hoc reduction step before downstream use. The practical workflow therefore depends on topic reduction, and current methods force a difficult trade-off. Tuning density parameters (direct reduction) often results in a loss of granular information, as the algorithm struggles to distinguish between low-density topics and true background noise. Conversely, standard agglomerative clustering (indirect reduction) can

merge distinct topics that happen to be geometrically close while leaving genuine redundancies intact.

Recently, Large Language Model (LLM)-assisted topic reduction has emerged as the highest-quality option for topic reduction, with Janssens et al. (2025a) showing it improves both coherence and diversity over standard agglomerative baselines. However, these methods are fundamentally constrained by scalability; the prohibitive cost and latency of iterative prompting render them impractical for large-scale topic models, while their stochastic nature introduces inconsistency.

In this paper, we introduce Vendi Clustering (VC), which reframes topic reduction as a global diversity-preserving compression scheme. At each step, VC selects the merge that minimizes the loss in the Vendi Score, a spectral diversity measure invariant to duplication. This inverts the usual objective: by optimizing globally for diversity, coherence emerges as a consequence of preserving distinct themes. A naive evaluation of this algorithm costs $O(n^6)$; we derive algebraic update rules at order $q = 2$ via the Frobenius norm that compute each merge cost in $O(1)$, reducing total complexity to $O(n^3)$.

We validate Vendi Clustering on the task of topic reduction using the 20 Newsgroups and AG News text datasets. Our method yields higher coherence and diversity than standard baselines (e.g., Agglomerative Clustering) and matches or surpasses LLM-assisted reduction at a fraction of the cost. Our contributions are: (1) a topic reduction method optimizing global diversity via efficient Vendi Score updates, (2) an algebraic reformulation of the Vendi Score at $q = 2$ that enables global lookahead at $O(n^3)$ total complexity, and (3) an empirical demonstration that this approach matches, if not surpasses, the quality of LLM-prompted topic reduction without requiring a language model.

2 Related Work

Topic reduction is conventionally performed by agglomerative clustering over topic representations. Agglomerative methods iteratively merge clusters according to a local linkage criterion (Sneath and Sokal, 1973). The dominant criterion in topic reduction settings is centroid cosine similarity, a variant whose local nature inherits a long-known structural limitation of agglomerative methods: each merge is made greedily on a pairwise cost, with no objective evaluating the resulting partition globally (Dasgupta, 2016; Moseley and Wang, 2017; Cohen-addad et al., 2019). This limitation is amplified in embedding-based settings because pairwise distances concentrate in high dimensions (Beyer et al., 1999), undermining the discriminative power of any single nearest-pair decision. Vendi Clustering operates in the same indirect-reduction family but replaces the local criterion with a global spectral objective, addressing redundancy at the level of the partition rather than the level of pairwise distance.

2.1 Topic Modeling and the Reduction Taxonomy

Topic modeling frameworks have evolved from probabilistic generative models, such as Latent Dirichlet Allocation (LDA; Blei et al., 2003; Hofmann, 1999), to embedding-based approaches (Angelov, 2020; Dieng et al., 2020; Sia et al., 2020) that leverage pre-trained language model representations (Devlin et al., 2019; Wang et al., 2020). BERTopic (Grootendorst, 2022) exemplifies this modern paradigm, decomposing the topic modeling process into three modular stages: (1) embedding documents into a vector space, (2) dimensionality reduction on these document embeddings (e.g., UMAP; McInnes et al., 2020), and (3) clustering the reduced embeddings to discover latent topics (e.g., HDBSCAN; Campello et al., 2013). Recent work continues to refine the neural topic modeling paradigm: FASTopic (Wu et al., 2024a) proposes a transformer-based framework that is faster and more stable than BERTopic, while other methods use LLMs as topic generators directly (Pham et al., 2024; Mu et al., 2024; Wang et al., 2023).

Across these frameworks, the burden of topic granularity falls on the clustering or generation stage. HDBSCAN, the default choice in BERTopic, routinely produces hundreds of fine-grained overlapping topics on real-world corpora (Janssens

et al., 2025b). To produce a tractable and interpretable model, practitioners apply post-hoc reduction. Janssens et al. (2025b) provide the first systematic comparison of reduction strategies for BERTopic, categorizing existing methods into two families: direct reduction, which alters the initial clustering phase, and indirect reduction, which operates on the topics after they have been generated.

Direct vs. Indirect Reduction. *Direct Reduction* intervenes in the initial clustering phase, typically by raising HDBSCAN’s `min_cluster_size`, coupling reduction to outlier detection and biasing the final set toward compact, high-density regions (Janssens et al., 2025b). *Indirect Reduction* instead operates on already-generated topics, with the standard method being agglomerative clustering with cosine-similarity linkage on centroids. This decouples reduction from outlier handling but inherits local-linkage limitations: geometrically close but distinct themes may merge while redundant themes with drifted centroids remain separate. Topics are typically represented either as SBERT (Reimers and Gurevych, 2019) centroids or via class-based TF-IDF (c-TF-IDF; Grootendorst, 2022), with Janssens et al. (2025b) finding c-TF-IDF generally produces stronger reductions.

2.2 Modern LLM-Assisted Reduction

Large Language Models (LLMs) have recently been integrated into the topic modeling pipeline to enhance topic discovery and interpretability (Stammach et al., 2023; Pham et al., 2024; Wang et al., 2023; Mu et al., 2024). In the specific context of topic reduction, these frameworks attempt to replace mathematical distance with semantic reasoning. As proposed by Janssens et al. (2025a), this typically involves an iterative list-wise prompting strategy. In this workflow, the model is initialized with a set of m topics, each represented by its top keywords or an LLM-generated topic label. At each iteration, the entire list of current topics is fed into the LLM with a prompt instructing it to identify the indices of the single most semantically similar pair. These two topics are then merged, the list is updated, and the full prompt is re-submitted. This cycle repeats until a target k is reached.

While LLM-assisted reduction offers a semantic alternative to clustering heuristics, it faces distinct limitations. Most notably, it incurs substantial computational overhead, particularly as the full prompt must be re-processed for every iterative merge. Additionally, limited context windows re-

strict the number of topics that can be processed simultaneously (Liu et al., 2024), impeding scalability to larger topic counts. Furthermore, reliability is inconsistent; Janssens et al. (2025a) report that LLaMA-based models frequently fail to generate outputs when confronted with sensitive content.

3 Method

3.1 The Vendi Score

The Vendi Score (VS) was originally introduced as a similarity-based spectral metric of diversity (Friedman and Dieng, 2023). Given a collection of samples $\{X_i\}_{i=1,\dots,n}$ and a positive semi-definite similarity kernel $K_{i,j} = K(X_i, X_j)$ with $K_{ii} = 1$, VS is defined by:

$$\text{VS} = \exp \left(- \sum_{i=1}^n \lambda_i \log(\lambda_i) \right)$$

where the λ_i are the eigenvalues of the normalized similarity matrix K/n . The Vendi Score has been generalized to a broader family of diversity measures, parameterized by a tunable exponent q (Pasarkar and Dieng, 2024):

$$\text{VS}_q = \left(\sum_{i=1}^n \lambda_i^q \right)^{\frac{1}{(1-q)}}$$

3.2 The Vendi Clustering Framework

We introduce Vendi Clustering (VC), an agglomerative hierarchical clustering algorithm that treats cluster formation as a global diversity-preserving compression scheme. Unlike classical hierarchical methods, such as single, complete, or Ward’s linkage, which make merging decisions based on local pairwise distances between clusters, VC is guided by a global information-theoretic objective.

We define the Vendi Linkage between two candidate items (or clusters) a and b as the loss in total diversity induced by their merger. Specifically, let K be the similarity matrix of the current set of n clusters. The calculation of the post-merge matrix K' requires a policy for representing the union $a \cup b$. While the framework is compatible with various linkage functions, we focus on datasets where items can be represented as vectors in a metric space, allowing the merged cluster to be defined by its geometric centroid. If merging a and b results in a new similarity matrix K' of size $n - 1$, the linkage criterion is:

$$\Delta(a, b) = \text{VS}(K) - \text{VS}(K')$$

By selecting the pair $(a^*, b^*) = \arg \min \Delta(a, b)$ at each iteration, the algorithm greedily minimizes the per-step reduction in spectral diversity, or the effective number of distinct components in the dataset. This global criterion is designed to mitigate the premature collapse of diversity often seen in local linkage methods, where unique but proximal clusters are merged early due to local density.

3.3 Vendi Clustering for Topic Reduction

We now apply the general VC framework to the specific task of topic reduction. In this setting, the input comprises a set of m initial topic embeddings $\{e_i\}_{i=1}^m \subset \mathbb{R}^d$ and their corresponding document counts $\{n_i\}_{i=1}^m$, where n_i denotes the number of documents assigned to topic i .

We use the cosine kernel $K_{ij} = \mathbf{e}_i^\top \mathbf{e}_j$ on ℓ_2 -normalized topic embeddings. This matches the standard similarity used in BERTopic reduction and our agglomerative baseline. For normalized embeddings, K is a positive semidefinite Gram matrix with unit diagonal, making it a valid Vendi kernel and providing the trace condition used by our $q = 2$ acceleration.

When two topics a and b are selected for a merge, their new representation is calculated as the unit-normalized weighted centroid:

$$\mathbf{e}_{a \cup b} = \frac{n_a \mathbf{e}_a + n_b \mathbf{e}_b}{\|n_a \mathbf{e}_a + n_b \mathbf{e}_b\|_2}.$$

Weighting by n_i rather than averaging uniformly prevents low-prevalence topics, which may include noise or niche outliers, from dominating the merged representation.

Algorithm 1 shows Vendi Clustering in its most general form, applicable to any diversity order q . We present it here as a conceptual overview; the procedure we run in practice shares this structure but is substantially more involved, replacing the full Vendi Score recomputation at each candidate merge with the $O(1)$ updates derived in section 3.4. The optimized algorithm, with the cached row-square sums and lookahead Gram matrix that those updates require, is detailed in Appendix A.3.

3.4 Algebraic Acceleration and Complexity

Naively implementing Vendi Clustering is computationally prohibitive. At each iteration, the algorithm must evaluate every possible pair of clusters ($O(n^2)$ pairs) to identify the merge that maximizes global diversity. Since the Vendi Score is a spectral

Algorithm 1: Vendi Clustering for Topic Reduction

Input : Topic embeddings $\{\mathbf{e}_i\}_{i=1}^m$, topic sizes $\{n_i\}_{i=1}^m$, target k_{target}

Output : Reduced topic set

Compute similarity matrix K with $K_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j)$;
 $V \leftarrow \text{VendiScore}(K)$;

while $m > k_{\text{target}}$ and $m > 1$ **do**
 $(a^*, b^*, \Delta^*) \leftarrow \text{None}, \text{None}, \infty$;
 foreach topic pair (a, b) **do**
 Merge: $\mathbf{e}_{a \cup b} = \frac{n_a \mathbf{e}_a + n_b \mathbf{e}_b}{\|n_a \mathbf{e}_a + n_b \mathbf{e}_b\|_2}$;
 Update K' by replacing a, b with merged topic;
 $V' \leftarrow \text{VendiScore}(K')$;
 $\Delta = V - V'$;
 if $\Delta < \Delta^*$ **then**
 store $(a^*, b^*, \Delta^*, K', V')$;
 Apply best merge (a^*, b^*) ;
 update $K \leftarrow K', V \leftarrow V'$;
 $m \leftarrow m - 1$, and topic set;

return reduced topics;

measure, a standard calculation for a single candidate pair requires a full eigendecomposition of the similarity matrix ($O(n^3)$). Consequently, the total complexity to reduce n items to a single cluster scales at a prohibitive $O(n \times n^2 \times n^3) = O(n^6)$.

Fast Updates via Frobenius Norm. We overcome this barrier by leveraging the properties of the Vendi Score at order $q = 2$. For a similarity matrix K where $\text{tr}(K) = n$, the score simplifies to $\text{VS}_2(K) = n^2 / \|K\|_F^2$, where $\|K\|_F^2$ is the squared Frobenius norm. This equality holds because, for the eigenvalues λ_i of K/n ,

$$\sum_i \lambda_i^2 = \text{tr}((K/n)^2) = \|K\|_F^2 / n^2,$$

allowing us to bypass a full eigendecomposition of the similarity matrix.

Crucially, we derive an update rule that computes the change in $\|K\|_F^2$ for any candidate merge in constant $O(1)$ time. Because the merged centroid is a linear combination of parent vectors, the new similarity terms can be derived algebraically from quantities that depend only on existing rows. By caching row-square sums and maintaining a lookahead Gram matrix $G = KK^\top$, we can evaluate the diversity loss for all $O(n^2)$ candidate

pairs in $O(n^2)$ time per iteration. This optimization reduces the total algorithmic complexity from $O(n^6)$ to a tractable $O(n^3 + dn^2)$, which effectively scales as $O(n^3)$ during the iterative phase. We provide the full algebraic derivation in Appendix A.1, the full time complexity derivation in Appendix A.2, and the optimized caching pseudocode in Appendix A.3.

Complexity and Scalability. In practice, d is either a small constant (SBERT, ~ 384) or effectively sparse (c-TF-IDF), and n is the topic count (10 to 10^3), not the document count (10^4 to 10^6). Since the initial density-based clustering handles the document-level scale, VC operates only on the much smaller topic space, and the cubic term typically completes in seconds.

3.5 The Choice of $q = 2$

The choice of $q = 2$ is principled, not merely convenient. As shown in section 3.4, $\text{VS}_2(K)$ reduces to $n^2 / \|K\|_F^2$: a function of the squared Frobenius norm that connects it to a classical family of diversity measures. Treating the normalized eigenvalues of K/n as a distribution over spectral components (they sum up to one by construction), $1/\text{VS}_2$ recovers the Simpson concentration index, whose complement, $1 - 1/\text{VS}_2$, corresponds to the Probability of Interspecific Encounter (PIE; Simpson, 1949; Hurlbert, 1971) from ecology: the chance that two samples drawn at random fall into different components. Under this reading, VS_2 measures the effective number of distinct topics in the set: a set of n perfectly orthogonal topics gives $\text{VS}_2 = n$, while overlapping or redundant topics collapse it toward a smaller effective count. VC’s merge criterion therefore has a direct operational reading: select the merge that reduces this effective topic count by the smallest amount.

Lower orders ($q \leq 1$, including the original VS at $q = 1$) weight small eigenvalues more heavily, making the score sensitive to low-signal residuals (stop-word groups, near-empty clusters) that density-based methods often produce. The quadratic weighting at $q = 2$ dampens these and focuses on dominant thematic structure, while also enabling the $O(1)$ Frobenius update (section 3.4).

Empirically verifying the choice across orders is itself constrained by computation: $q \neq 2$ requires $O(n^6)$ in our setup, tractable only at small initial topic counts. We run a sensitivity ablation at the smallest practical scale: BERTopic’s output on 20 Newsgroups ($n = 139$, c-TF-IDF) reduced

Table 1: **Main Results.** Vendi Clustering (VC) vs. Agglomerative Clustering (AC) and LLM-based reduction (Gemma-3-12b-it, GPT-4o-mini) on 20 Newsgroups and AG News. Means $\pm$ standard deviations across three seeds; best per k in **bold**. Gemma’s seed 42 halted prematurely on 20 Newsgroups; statistics for $k \in \{25, 50\}$ exclude it.

k	Method	20 Newsgroups			AG News		
		C_v	NPMI	TD	C_v	NPMI	TD
25	VC (c-TF-IDF)	0.5815$\pm$0.0104	0.0498 $\pm$ 0.0031	0.8800$\pm$0.0069	0.8161$\pm$0.0026	0.2802$\pm$0.0049	0.9640$\pm$0.0080
	VC (SBERT)	0.5513 $\pm$ 0.0063	0.0529 $\pm$ 0.0007	0.6893 $\pm$ 0.0083	0.8057 $\pm$ 0.0100	0.2665 $\pm$ 0.0026	0.9453 $\pm$ 0.0061
	AC (c-TF-IDF)	0.5414 $\pm$ 0.0450	0.0249 $\pm$ 0.0232	0.8181 $\pm$ 0.0087	0.6078 $\pm$ 0.0234	0.1395 $\pm$ 0.0123	0.7931 $\pm$ 0.0146
	AC (SBERT)	0.5054 $\pm$ 0.0077	0.0169 $\pm$ 0.0060	0.4042 $\pm$ 0.0150	0.5466 $\pm$ 0.0204	0.0895 $\pm$ 0.0090	0.7222 $\pm$ 0.0168
	LLM (Gemma)	0.5763 $\pm$ 0.0034	0.0563 $\pm$ 0.0107	0.6520 $\pm$ 0.0962	0.7160 $\pm$ 0.0435	0.1990 $\pm$ 0.0313	0.9420 $\pm$ 0.0085
	LLM (GPT)	0.5785 $\pm$ 0.0190	0.0589$\pm$0.0057	0.8213 $\pm$ 0.0234	0.7547 $\pm$ 0.0217	0.2044 $\pm$ 0.0167	0.9267 $\pm$ 0.0180
50	VC (c-TF-IDF)	0.6189$\pm$0.0173	0.0779$\pm$0.0122	0.8667$\pm$0.0151	0.8247$\pm$0.0075	0.2708$\pm$0.0055	0.9327 $\pm$ 0.0031
	VC (SBERT)	0.5900 $\pm$ 0.0066	0.0635 $\pm$ 0.0057	0.7867 $\pm$ 0.0058	0.8150 $\pm$ 0.0139	0.2675 $\pm$ 0.0097	0.9420$\pm$0.0072
	AC (c-TF-IDF)	0.5753 $\pm$ 0.0180	0.0532 $\pm$ 0.0132	0.8238 $\pm$ 0.0066	0.7146 $\pm$ 0.0103	0.2010 $\pm$ 0.0089	0.8925 $\pm$ 0.0062
	AC (SBERT)	0.5426 $\pm$ 0.0180	0.0418 $\pm$ 0.0093	0.6177 $\pm$ 0.0082	0.6880 $\pm$ 0.0086	0.1681 $\pm$ 0.0045	0.8612 $\pm$ 0.0134
	LLM (Gemma)	0.5868 $\pm$ 0.0041	0.0564 $\pm$ 0.0160	0.6990 $\pm$ 0.0665	0.7214 $\pm$ 0.0574	0.1996 $\pm$ 0.0398	0.9310 $\pm$ 0.0156
	LLM (GPT)	0.6010 $\pm$ 0.0126	0.0622 $\pm$ 0.0106	0.8187 $\pm$ 0.0260	0.7922 $\pm$ 0.0134	0.2328 $\pm$ 0.0165	0.9093 $\pm$ 0.0023
100	VC (c-TF-IDF)	0.6326$\pm$0.0065	0.0734$\pm$0.0038	0.8443$\pm$0.0012	0.8297$\pm$0.0041	0.2725$\pm$0.0008	0.8940$\pm$0.0130
	VC (SBERT)	0.6174 $\pm$ 0.0066	0.0651 $\pm$ 0.0064	0.8207 $\pm$ 0.0093	0.8231 $\pm$ 0.0028	0.2621 $\pm$ 0.0027	0.8657 $\pm$ 0.0085
	AC (c-TF-IDF)	0.6160 $\pm$ 0.0034	0.0685 $\pm$ 0.0021	0.8202 $\pm$ 0.0090	0.7768 $\pm$ 0.0096	0.2374 $\pm$ 0.0057	0.8909 $\pm$ 0.0056
	AC (SBERT)	0.6200 $\pm$ 0.0091	0.0716 $\pm$ 0.0078	0.7694 $\pm$ 0.0132	0.7605 $\pm$ 0.0031	0.2156 $\pm$ 0.0025	0.8438 $\pm$ 0.0048
	LLM (Gemma)	0.6188 $\pm$ 0.0032	0.0618 $\pm$ 0.0084	0.7447 $\pm$ 0.0197	0.7716 $\pm$ 0.0030	0.2244 $\pm$ 0.0038	0.8515 $\pm$ 0.0064
	LLM (GPT)	0.6149 $\pm$ 0.0095	0.0644 $\pm$ 0.0046	0.8150 $\pm$ 0.0095	0.7908 $\pm$ 0.0041	0.2331 $\pm$ 0.0056	0.8483 $\pm$ 0.0202

to $k = 50$ for $q \in \{0.1, 0.5, 1.0, 1.5, 2.0, 5.0, \infty\}$. As shown in Appendix C, no order meaningfully outperforms $q = 2$: the total metric range across all q values averages just 3.1% of the mean, with $q = 2$ achieving the best or tied-best score on every metric. Combined with the $80\times$ speedup, this supports the choice on both principled and empirical grounds.

4 Experimental Setup

4.1 Datasets and Preprocessing

We evaluate Vendi Clustering on two standard text classification benchmarks that span contrasting evaluation regimes: 20 Newsgroups (18,846 long-form discussion documents, mean 181.6 words) and AG News (127,600 short-form news documents, mean 39.9 words).

To retain maximum contextual information for the transformer-based embedding module and ensure fair comparison across methods, the datasets are left completely unprocessed. We avoid traditional stop-word removal or lemmatization, as evidence suggests that Sentence-BERT (SBERT) embeddings perform most effectively on raw, natural language text (Janssens et al., 2025b).

4.2 Evaluation Metrics

To assess the quality of the reduced topic models, we utilize three standard lexical-statistical proxies that approximate human interpretability and topic diversity. Topic Coherence is evalu-

ated using Normalized Pointwise Mutual Information (NPMI), which quantifies the semantic relatedness of a topic’s top keywords based on their joint occurrence within the corpus. Ranging from -1 (complete divergence) to 1 (perfect co-occurrence), NPMI is widely used as an automatic proxy for local topic interpretability (Bouma, 2009; Röder et al., 2015), although it does not perfectly align with human judgments (Hoyle et al., 2021). We further include C_v coherence (Röder et al., 2015), which combines a sliding-window co-occurrence measure with cosine similarity between word context vectors to provide a more nuanced view of thematic consistency.

Topic Diversity (TD) is measured following the definition by Dieng et al. (2020) in the context of the Embedded Topic Model (ETM). It is calculated as the ratio of unique words present in the top-10 keywords across all topics to the total number of keywords. This metric provides a global assessment of the model’s ability to discover distinct, non-redundant themes while penalizing the overproduction of overlapping topics.

4.3 Experiment Protocols

We evaluate Vendi Clustering across three protocols, each targeting a different baseline family. All three start from a BERTopic base model with standard parameters.

Protocol 1: Vendi Clustering vs. Agglomerative Clustering. We employ a 2×2 factorial

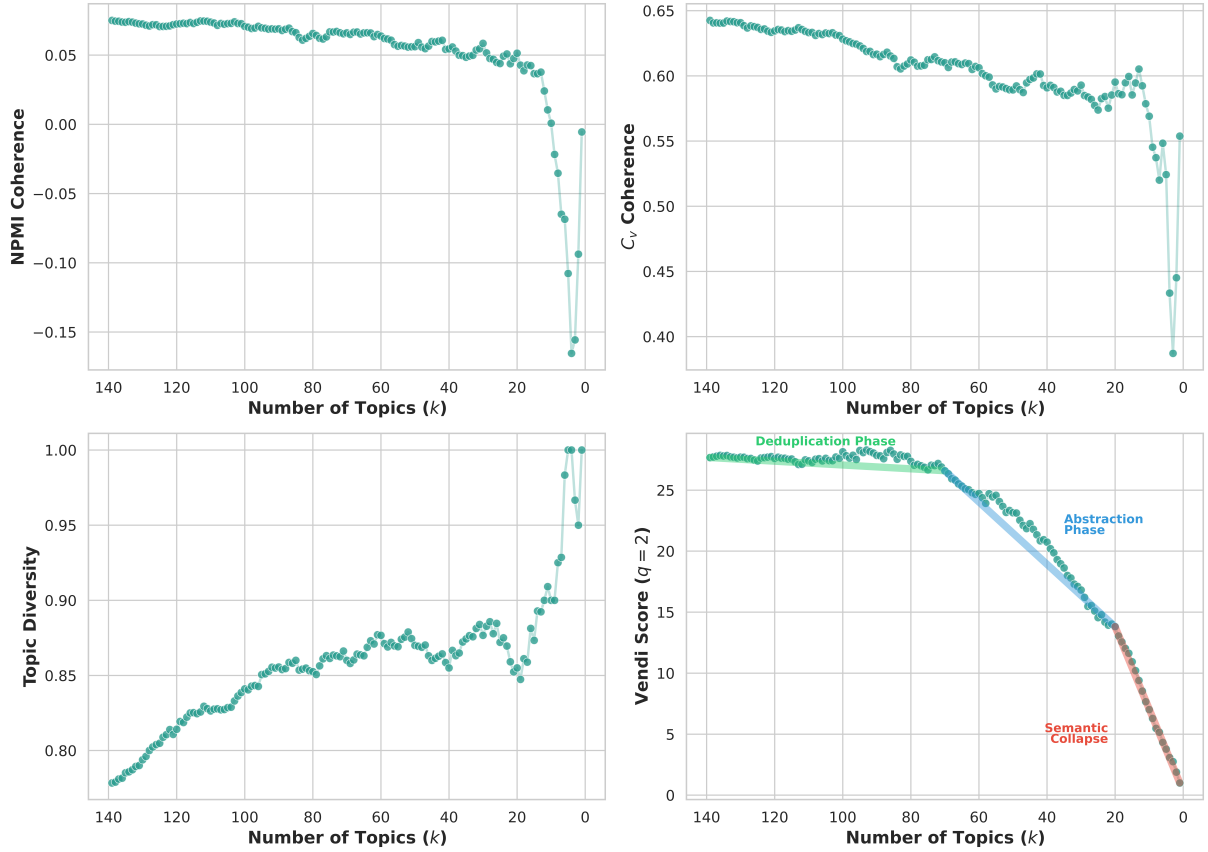


Figure 1: **Vendi Clustering Trajectory on 20 Newsgroups.** VS_2 , NPMI, C_v , and Topic Diversity over $k = 139 \rightarrow 1$.

design crossing reduction method (VC vs. AC) with topic representation (SBERT vs. c-TF-IDF), reducing the base model to $k \in \{25, 50, 100\}$. To account for UMAP stochasticity, all experiments were repeated across three random seeds (42, 43, 44), with reported metrics representing the mean and standard deviation across trials. This design allows us to determine if Vendi Clustering offers consistent improvements across different embedding types and levels of granularity, or if its benefits are specific to certain representations.

Protocol 2: Vendi Clustering vs. LLM-Assisted Merging. We benchmark against the iterative LLM-assisted reduction strategy of Janssens et al. (2025a), evaluating two LLMs used in their study on the same targets as Protocol 1: Gemma-3-12b-it (locally deployable) and GPT-4o-mini (frontier API). Gemma-3-12b-it represents the realistic local-deployment scenario: the kind of setup a researcher can stand up end-to-end on a single consumer-grade GPU without API access. Memory constraints on a single Ampere GPU under vLLM require 4-bit quantization to host the model end-to-end. The standard top-10 keyword represen-

tation also exceeds Gemma’s 8192-token context window at initial counts of $k \approx 150\text{--}200$, so we truncate each topic to its top-5 keywords, the minimum needed to keep the full prompt within budget. The model occasionally produced malformed outputs, halting reduction prematurely for some seeds (noted in section 5.1). GPT-4o-mini is run via the API with the standard top-10 keywords per topic. The full prompt template is in Appendix G.

Protocol 3: Vendi Clustering vs. Direct HDBSCAN Tuning. To evaluate whether VC can decouple reduction from outlier rejection, we use a shadowing design: we generate a fine-grained base model with low `min_cluster_size`, then create matched comparison states by training separate BERTopic models with stricter `min_cluster_size` values under both eom and leaf selection. For each target, we apply VC to the fine-grained base model until topic counts match, then compare the resulting outlier ratios (the proportion of documents classified as noise by HDBSCAN) of the resulting models to determine if Vendi Clustering is able to achieve the same level of compression as Direct Reduction while preserving

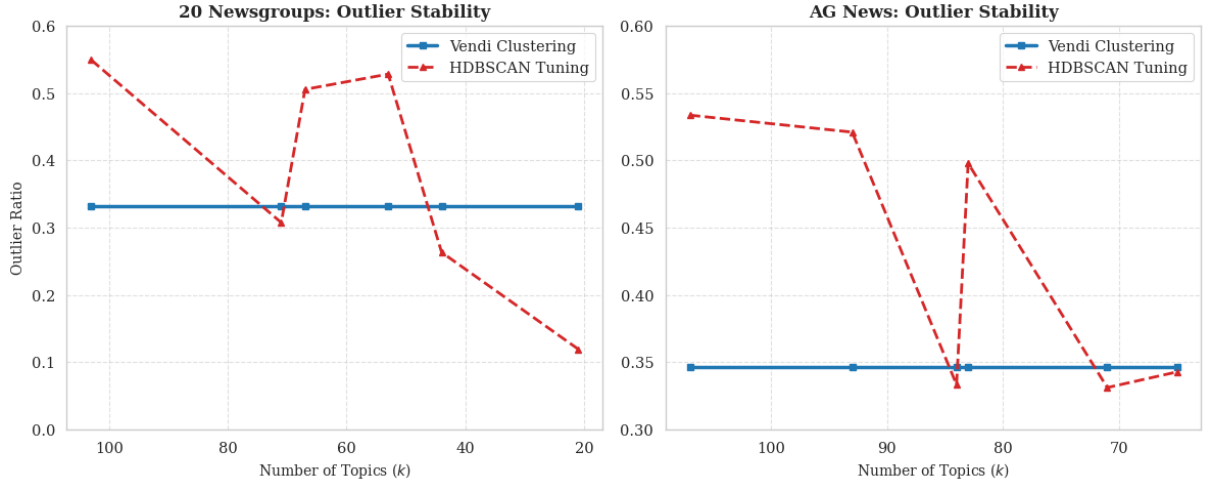


Figure 2: **Outlier Stability.** Outlier ratios for VC (solid blue) vs. Direct HDBSCAN tuning (dashed red) across topic counts (k). Left: 20 Newsgroups; Right: AG News.

document assignments that density-based tuning would otherwise discard.

Supplementary comparison with standalone topic models.

Our primary experiments hold the BERTopic base model fixed to isolate the effect of the topic reduction method. For broader context, Appendix F compares BERTopic reduced with VC against standalone LDA, CombinedTM, ETM, and FASTopic models. This supplementary experiment compares complete topic-modeling pipelines rather than reduction methods in isolation and may be useful to topic modeling practitioners choosing between native- k modeling and post-hoc topic reduction.

5 Results and Discussion

5.1 Indirect Reduction Methods Benchmark

We begin by evaluating Vendi Clustering against two distinct paradigms of indirect topic reduction. While all three methods share the same iterative mechanism, merging one topic pair per step until a target k is reached, they fundamentally differ in their selection criteria: Agglomerative Clustering minimizes local centroid distance, LLM-Merging utilizes semantic reasoning to identify redundancy, and Vendi Clustering optimizes for global spectral diversity.

Performance Analysis. Vendi Clustering with c-TF-IDF representations dominates across nearly all coherence and diversity metrics at every target k (Table 1), surpassing both the Agglomerative baselines and the SBERT-based variants on C_v and TD at every target. The sole exception is NPMI

on 20 Newsgroups at $k = 25$, where GPT-4o-mini leads narrowly; VC retains the lead on C_v and TD. Gemma shows a similar NPMI advantage at $k = 25$ but at a severe cost to coverage, with TD collapsing to 0.6520 against Vendi’s 0.8800.

Trajectory Analysis: The Three Phases of Reduction.

Figure 1 suggests that VC’s reduction trajectory unfolds in three phases, with changes in VS_2 mirrored by the other quality metrics. To test whether the visually apparent boundaries correspond to changes in slope, we fit a continuous piecewise-linear model to the VS_2 trajectory. The model identifies breakpoints at $k = 69$ and $k = 21$, closely matching the observed boundaries around $k = 70$ and $k = 20$ and thereby quantitatively supports this three-phase reading (Appendix D).

In the **Deduplication** phase ($k = 139 \rightarrow 70$), VS_2 is nearly flat, coherence (C_v , NPMI) holds steady, and TD increases, indicating duplicate removal. The **Abstraction** phase ($k = 70 \rightarrow 21$) shows VS_2 declining steadily while coherence and TD remain stable, as the algorithm folds sub-themes into broader categories without losing meaning. Below $k = 21$, the trajectory enters **Semantic Collapse**: C_v and NPMI fall sharply, while TD approaches 1.0 as a consequence of the small number of remaining topics rather than improved quality. Notably, $k = 20$ is exactly the number of ground-truth categories in 20 Newsgroups: suggestive rather than conclusive, but consistent with VC having resolved the dataset’s intrinsic dimensionality.

Robustness of Global Objective. Our results

Table 2: **Protocol 3 Results:** Comparison of Vendi Clustering (VC) vs. Direct HDBSCAN Tuning. Left: 20 Newsgroups; Right: AG News. VC consistently achieves higher coherence and diversity while matching target topic counts (k).

<i>Dataset: 20 Newsgroups</i>						<i>Dataset: AG News</i>					
Config	k	Method	C_v	NPMI	TD	Config	k	Method	C_v	NPMI	TD
mcs=30 eom	71	Vendi HDBSCAN	0.6077 0.5811	0.0646 0.0598	0.8718 0.6775	mcs=200 eom	84	Vendi HDBSCAN	0.8392 0.8011	0.2873 0.2346	0.9417 0.8476
mcs=30 leaf	103	Vendi HDBSCAN	0.6284 0.6451	0.0702 0.0816	0.8330 0.7282	mcs=200 leaf	107	Vendi HDBSCAN	0.8395 0.8283	0.2821 0.2544	0.9234 0.8308
mcs=40 eom	44	Vendi HDBSCAN	0.6104 0.5402	0.0642 0.0402	0.8705 0.5409	mcs=225 eom	71	Vendi HDBSCAN	0.8433 0.8132	0.2894 0.2377	0.9577 0.8606
mcs=40 leaf	67	Vendi HDBSCAN	0.6084 0.5940	0.0652 0.0653	0.8746 0.6299	mcs=225 leaf	93	Vendi HDBSCAN	0.8385 0.8346	0.2856 0.2563	0.9366 0.8409
mcs=50 eom	21	Vendi HDBSCAN	0.6033 0.5107	0.0544 -0.0013	0.8762 0.3190	mcs=250 eom	65	Vendi HDBSCAN	0.8463 0.8007	0.2931 0.2272	0.9646 0.8554
mcs=50 leaf	53	Vendi HDBSCAN	0.5928 0.5661	0.0584 0.0525	0.8698 0.5698	mcs=250 leaf	83	Vendi HDBSCAN	0.8410 0.8274	0.2883 0.2466	0.9434 0.8313

further indicate that the benefits of the Vendi Score are robust to the choice of underlying representation. While c-TF-IDF proved optimal, Vendi Clustering applied to SBERT embeddings frequently emerged as the second-best performing method, consistently outpacing the traditional Agglomerative approaches. For instance, on the AG News dataset (Table 1), Vendi (SBERT) surpassed both Agglomerative Clustering methods and Gemma on both coherence measures (C_v , NPMI) and topic diversity (TD).

Global Diversity Gains. Optimizing for global spectral diversity rather than local pairwise similarity allows VC to maintain high topic diversity even under aggressive compression. At $k = 25$ on 20 Newsgroups, VC (c-TF-IDF) improves TD by 35% over Gemma and 7% over GPT-4o-mini. This suggests a structural limitation of pairwise LLM-merging: while LLMs identify synonymy between two topics effectively, they appear less able to preserve the distinctness of the entire topic distribution, even at full precision with top-10 keyword context.

Computational Cost. Beyond quality, VC is significantly cheaper than LLM-prompted reduction. A full reduction completes in single-digit seconds on both datasets, whereas the LLM baselines require minutes per run because the entire topic list must be re-processed at every merge. Across both datasets, both representations, and all target k , VC is over $40\times$ faster than the LLM baselines, and up to $408\times$ faster in the SBERT setting. VC and standard agglomerative clustering are comparably

fast, so this speed comes at no runtime cost over the standard topic reduction baseline. Full wall-clock timings are reported in Appendix E.

5.2 Granularity vs. Information Loss

Direct HDBSCAN tuning bundles topic reduction with outlier rejection: a single parameter controls both granularity and noise tolerance, forcing practitioners to retune both jointly. We evaluate whether Vendi Clustering can decouple these conflicting objectives. We compare our method against two Direct HDBSCAN Reduction strategies: Leaf selection (which prioritizes fine-grained structure) and Excess of Mass (EOM) (which prioritizes data retention).

Decoupling Outlier Detection from Reduction. As illustrated in Figure 2, Direct Reduction exhibits high volatility in data retention. For example, on the 20 Newsgroups dataset, attempting to reduce topics via HDBSCAN’s Leaf selection caused the outlier ratio to spike from roughly 30% to over 54%. Conversely, while EOM selection successfully minimized outliers (dropping as low as 11.9%), it did so at the cost of semantic coherence, with NPMI scores collapsing to near-zero (-0.0013) in the low-granularity settings. This suggests EOM achieves data retention by absorbing low-density regions into broader clusters, at the cost of coherence within those clusters.

In contrast, Vendi Clustering maintains a constant outlier ratio across all target topic counts. By fixing the initial density threshold to a fine-grained base model, VC allows the practitioner to lock in a noise tolerance up front and reduce topics without

retuning it. Note that this stability is a property of *predictability*, not *minimization*: an appropriately tuned HDBSCAN configuration can potentially achieve lower outlier rates at specific k values, but the rate is a sensitive and non-monotonic function of the parameter setting. VC offers practitioners a fixed outlier budget regardless of target topic count, which is qualitatively different from competing on outlier rate directly.

Quality at Fixed Granularity. Even when constrained to match the specific topic counts produced by HDBSCAN tuning, VC consistently yields higher quality metrics. In the AG News benchmark (table 2), when VC was applied to reach the same k as the HDBSCAN baselines, it outperformed them on C_v coherence and Topic Diversity (TD) in every single trial. Notably, when matching the $k = 83$ target set by HDBSCAN (AG News, `min_cluster_size=250`), VC achieved a C_v of 0.8410 compared to the baseline’s 0.8274, while simultaneously maintaining a significantly lower outlier ratio (34.6% vs 49.8%) than the comparable leaf configuration.

These results indicate that Vendi Clustering allows users to retain the granularity of fine-grained density settings while achieving the compactness of low- k models.

6 Conclusion

In this work, we introduced Vendi Clustering, a hierarchical framework that fundamentally reframes the clustering objective from local distance minimization to global diversity preservation. By integrating the Vendi Score, a spectral measure of the effective number of components, into the merge criterion, our method eliminates the “myopic” behaviors inherent to standard agglomerative linkage and the “high-density bias” of direct parameter tuning. We further show that the theoretical benefits of this global criterion need not come at a prohibitive computational cost. Through our derivation of algebraic updates for the Frobenius norm, we reduced the complexity of the diversity lookahead from $O(n^6)$ to $O(n^3)$, making spectral diversity optimizations tractable for real-world applications.

Empirically, Vendi Clustering produced topic sets more coherent and diverse than standard baselines, while decoupling topic granularity from outlier rejection: practitioners can topic-model once and reduce to any target k without the volatility of density-based retuning. Most notably, it outper-

formed LLM-assisted reduction on the majority of coherence and diversity metrics while remaining deterministic and orders of magnitude faster.

Limitations

Vendi Clustering is designed for post-hoc topic reduction, where the input size n is the number of latent topics (typically 10 to 10^3) rather than the document count. At this scale, the $O(n^3)$ time complexity and $O(n^2)$ memory cost complete in seconds. The method is therefore not intended for direct document-level clustering at a corpus scale (n in the $10^5 - 10^7$ range), where memory alone could become prohibitive. For such settings, density-based clustering remains the appropriate first stage, with VC applied subsequently to the resulting topic set.

Our $O(1)$ per-pair update rule is specific to $q = 2$: the Frobenius-norm decomposition that enables it does not extend to general q . While our ablation (section 3.5) shows that $q = 2$ produces topic quality comparable to other diversity orders across the configurations tested, applications that require a particular q value would need the slower eigendecomposition-based approach. Approximation methods such as Nyström (Williams and Seeger, 2000) offer a possible direction for scaling beyond $q = 2$, which we leave to future work.

Our evaluation is restricted to two English-language benchmarks. While 20 Newsgroups and AG News represent contrasting regimes: long-form discussion (~ 180 words/document) and short-form news (~ 40 words/document), they do not exhaust the conditions under which topic reduction is practically deployed. Noisier corpora such as social media, domain-specific corpora such as scientific or legal text, and multilingual settings may surface failure modes that English long-form and news text do not, and we leave systematic evaluation in these regimes to future work.

We evaluate topic quality through automatic lexical-statistical metrics (C_v , NPML, Topic Diversity). While this protocol mirrors standard practice in directly comparable topic modeling work, automatic metrics are imperfect proxies for human interpretability (Hoyle et al., 2021). Human evaluation and LLM-judge evaluation would complement the present analysis.

Acknowledgments

AI assistants were used during this project primarily to support code development, including structuring portions of the experimental codebase and assisting with implementation details. Use in writing was limited to occasional editing for clarity and grammar on author-drafted text. All scientific contributions, including the design of the Vendi Clustering method, the algebraic acceleration at $q = 2$, the experimental design, analysis, and interpretation of results, are the authors' own work.

References

- Dimo Angelov. 2020. [Top2vec: Distributed representations of topics](#). *Preprint*, arXiv:2008.09470.
- Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When is “nearest neighbor” meaningful? In *Database Theory — ICDT’99*, pages 217–235, Berlin, Heidelberg. Springer Berlin Heidelberg.
- 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 and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.
- D. Blei, A. Ng, and M. 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 2009*, pages 31–40.
- Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Vincent Cohen-addad, Varun Kanade, Frederik Mallmann-trenn, and Claire Mathieu. 2019. [Hierarchical clustering: Objective functions and algorithms](#). *J. ACM*, 66(4).
- Sanjoy Dasgupta. 2016. [A cost function for similarity-based hierarchical clustering](#). In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC ’16, pages 118–127, New York, NY, USA. Association for Computing Machinery.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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.
- Dan Friedman and Adji Bousso Dieng. 2023. [The vendi score: A diversity evaluation metric for machine learning](#). *Transactions on Machine Learning Research*.
- Maarten Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#). *Preprint*, arXiv:2203.05794.
- Thomas Hofmann. 1999. [Probabilistic latent semantic indexing](#). In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’99*, pages 50–57, New York, NY, USA. Association for Computing Machinery.
- Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. [Is automated topic model evaluation broken? the incoherence of coherence](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.
- Stuart H. Hurlbert. 1971. [The nonconcept of species diversity: A critique and alternative parameters](#). *Ecology*, 52(4):577–586.
- Wannes Janssens, Matthias Bogaert, and Dirk Van den Poel. 2025a. [Llm-assisted topic reduction for bertopic on social media data](#). *Preprint*, arXiv:2509.19365.
- Wannes Janssens, Matthias Bogaert, and Dirk van den Poel. 2025b. [A comparative analysis of topic reduction techniques for bertopic](#). *IEEE Access*, 13:204087–204103.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Leland McInnes, John Healy, and James Melville. 2020. [Umap: Uniform manifold approximation and projection for dimension reduction](#). *Preprint*, arXiv:1802.03426.
- Benjamin Moseley and Joshua Wang. 2017. [Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search](#). In *Advances in Neural Information Processing Systems*, volume 30, pages 3094–3103. Curran Associates, Inc.

- 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.
- Amey P. Pasarkar and Adji Bousso Dieng. 2024. [Cousins of the Vendi score: A family of similarity-based diversity metrics for science and machine learning](#). In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3808–3816. PMLR.
- Chau Minh 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.
- 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.
- 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, WSDM ’15*, pages 399–408, New York, NY, USA. Association for Computing Machinery.
- Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. [Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.
- E. H. Simpson. 1949. [Measurement of diversity](#). *Nature*, 163:688.
- Peter H. A. Sneath and Robert R. Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W. H. Freeman, San Francisco.
- Dominik Stammach, 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.
- Anton Thielmann, Arik Reuter, Quentin Seifert, Elisabeth Bergherr, and Benjamin Säfken. 2024. [Topics in the haystack: Enhancing topic quality through corpus expansion](#). *Computational Linguistics*, 50(2):619–655.
- Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. 2023. [Prompting large language models for topic modeling](#). In *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.
- Christopher Williams and Matthias Seeger. 2000. [Using the nystrom method to speed up kernel machines](#). In *Advances in Neural Information Processing Systems*, volume 13, pages 682–688. MIT Press.
- Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024a. [Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 84447–84481. Curran Associates, Inc.
- Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2024b. [Towards the TopMost: A topic modeling system toolkit](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Bangkok, Thailand. Association for Computational Linguistics.

A Derivations and Complexity Analysis

In this appendix, we provide the mathematical foundations for the efficient implementation of Vendi Clustering with $q = 2$. We first derive the algebraic update rules for the Frobenius norm, then analyze the computational complexity, and finally present the pseudocode of the optimized algorithm.

A.1 Derivation of the Fast Update Rule

In this section, we derive a constant-time expression for evaluating each candidate merge under the Vendi Score of order $q = 2$. Recall that for a similarity matrix $K \in \mathbb{R}^{n \times n}$ with $\text{tr}(K) = n$, the Vendi Score is given by:

$$\text{VS}_2(K) = \frac{n^2}{\|K\|_F^2} = \frac{n^2}{\sum_{i,j} K_{ij}^2} \quad (1)$$

At each step of the algorithm, we consider merging two clusters a and b into a new cluster $p = a \cup b$. This transforms the matrix K (size $n \times n$) into K'

(size $(n-1) \times (n-1)$). Since $\text{VS}_2(K) = \frac{n^2}{\|K\|_F^2}$ and all candidate matrices K' at a given iteration share the same size $(n-1)$, minimizing the diversity loss $\Delta(a, b)$ is equivalent to minimizing the Frobenius norm $\|K'\|_F^2$. Thus, rather than recomputing the full score for every candidate merge, our algorithm searches for the pair (a, b) that yields the minimum $\|K'\|_F^2$.

Linearity and Normalization. Let the unnormalized centroid be $\mathbf{v}_p = n_a \mathbf{e}_a + n_b \mathbf{e}_b$. To maintain unit-norm representations, so that cosine similarity equals the inner product, we must normalize it: $\mathbf{e}_p = \mathbf{v}_p / \|\mathbf{v}_p\|$. The normalization constant is given by:

$$C_{ab}^2 = \|\mathbf{v}_p\|^2 = n_a^2 + n_b^2 + 2n_a n_b K_{ab} \quad (2)$$

Consequently, the similarity between the new cluster p and any existing cluster j is:

$$K_{pj} = \langle \mathbf{e}_p, \mathbf{e}_j \rangle = \frac{n_a K_{aj} + n_b K_{bj}}{\sqrt{C_{ab}^2}} \quad (3)$$

Decomposition of the Frobenius Norm. Let $R_i = \sum_{j=1}^n K_{ij}^2$ denote the squared norm of row i . We express the post-merge Frobenius norm as

$$\begin{aligned} \|K'\|_F^2 &= \|K\|_F^2 - \mathcal{D}_{ab} + \mathcal{A}_{ab}, \\ \mathcal{D}_{ab} &= 2R_a + 2R_b - 2K_{ab}^2 - 2, \\ \mathcal{A}_{ab} &= 2 \sum_{j \neq a, b} K_{pj}^2 + 1, \end{aligned} \quad (4)$$

where $\mathcal{D}_{ab}$ collects the entries removed with rows and columns a and b , and $\mathcal{A}_{ab}$ collects the entries introduced by the merged cluster p . The factors of 2 account for the symmetry of K . Within $\mathcal{D}_{ab}$, the term $-2K_{ab}^2$ corrects for double-counting the two off-diagonal entries K_{ab} and K_{ba} , while the constant terms account for replacing the two unit diagonal entries K_{aa} and K_{bb} with the single entry $K_{pp} = 1$.

Constant-Time Calculation via the Lookahead Matrix. The remaining bottleneck is the squared norm of the new similarity row, excluding the merged clusters a and b . Define

$$u_{ab,j} := n_a K_{aj} + n_b K_{bj}. \quad (5)$$

By the normalized centroid expansion above, $K_{pj} = u_{ab,j} / C_{ab}$. Therefore, defining $T_{ab} := \sum_{j \neq a, b} K_{pj}^2$, we obtain

$$T_{ab} = \frac{1}{C_{ab}^2} \sum_{j \neq a, b} u_{ab,j}^2. \quad (6)$$

To evaluate T_{ab} without iterating over j , we use the cached row-square sums $R_i = \sum_{j=1}^n K_{ij}^2$ and the lookahead Gram matrix $G = K K^\top$. Excluding indices a and b gives

$$\sum_{j \neq a, b} K_{aj}^2 = R_a - 1 - K_{ab}^2, \quad (7)$$

$$\sum_{j \neq a, b} K_{bj}^2 = R_b - 1 - K_{ab}^2, \quad (8)$$

$$\sum_{j \neq a, b} K_{aj} K_{bj} = G_{ab} - 2K_{ab}. \quad (9)$$

The final identity follows from the symmetry of K and the unit-diagonal conditions $K_{aa} = K_{bb} = 1$. Expanding $u_{ab,j}^2$ and substituting these identities into eq. (6) yields

$$\begin{aligned} C_{ab}^2 T_{ab} &= n_a^2 (R_a - 1 - K_{ab}^2) \\ &\quad + n_b^2 (R_b - 1 - K_{ab}^2) \\ &\quad + 2n_a n_b (G_{ab} - 2K_{ab}). \end{aligned} \quad (10)$$

Consequently, T_{ab} , and hence the added contribution $\mathcal{A}_{ab} = 2T_{ab} + 1$, can be evaluated in $O(1)$ using cached quantities.

A.2 Complexity Analysis

We analyze the computational cost of the optimized algorithm 2. The complexity is defined by two primary phases: Initialization and the Iterative Merging Loop.

Initialization: $O(dm^2 + m^3)$. The algorithm begins by computing the pairwise cosine similarity matrix $K \in \mathbb{R}^{m \times m}$ from the input topic embeddings $E \in \mathbb{R}^{m \times d}$. Computing K requires $\frac{m(m-1)}{2}$ dot products of dimension d , yielding a cost of $O(dm^2)$. Following this, the algorithm initializes the lookahead structures. The Gram matrix $G = K K^\top$ involves multiplying two $m \times m$ matrices. While naively $O(m^3)$, this occurs only once. Similarly, the row-square-sum vector R requires a single pass over K , costing $O(m^2)$. The total initialization cost is dominated by $O(dm^2)$ in high-dimensional settings or $O(m^3)$ in low-dimensional ones. It is worth noting that in most practical topic modeling scenarios, this initialization is highly efficient. For sparse lexical vectors (e.g., c-TF-IDF), the effective d scales only with the number of non-zero entries. Conversely, for dense semantic embeddings (e.g., Sentence-BERT), d is typically a small constant (e.g., 384 or 768) often comparable to the topic count m , ensuring that initialization effectively scales as $O(m^3)$.

Iterative Merging Phase: $O(m^3)$. The algorithm performs $(m - k_{\text{target}})$ merge iterations. Let t denote the current number of topics, decreasing from m to k_{target} . The iterative phase consists of two steps: Search and Update. In the **Search Step**, the algorithm evaluates $\frac{t(t-1)}{2} \approx O(t^2)$ candidate pairs. Thanks to the derivation in appendix A.1, evaluating the diversity change Δ_Ω for a single pair involves only constant-time algebraic operations on cached values $(R_a, R_b, G_{ab}, K_{ab}, n_a, n_b)$. The total search cost per iteration is therefore:

$$\text{Search Cost} = O(t^2) \times O(1) = O(t^2)$$

In the **Update Step**, once the optimal pair is identified, the algorithm computes the new centroid e_{new} in $O(d)$ and the new similarity row K_{new} in $O(t)$. Crucially, updating the Gram matrix G after a merge requires updating $O(t^2)$ cached entries, costing $O(t^2)$. The update cost per iteration is:

$$\text{Update Cost} = O(t^2 + d) \approx O(t^2)$$

Total Complexity. Summing the costs over all iterations yields the total runtime complexity:

$$\begin{aligned} T(m) &\approx \sum_{t=k}^m (O(t^2) + O(t^2)) \\ &\approx O\left(\sum_{t=1}^m t^2\right) = O(m^3). \end{aligned} \quad (11)$$

Thus, the total complexity of Vendi Clustering is $O(m^3 + dm^2)$. This cubic scaling is comparable to classical hierarchical procedures with non-trivial global or geometric updates. The early iterations, when $t \approx m$, dominate the total cost, while later iterations become progressively cheaper as the topic set shrinks.

Comparison with Naive Approach:

Without the algebraic acceleration derived in this work, evaluating the Vendi Score requires a full eigendecomposition $O(t^3)$ for every candidate pair. This results in a search cost of $O(t^2 \times t^3) = O(t^5)$ per iteration, leading to a prohibitive total complexity of $O(m^6)$. Our optimization therefore reduces the asymptotic complexity from $O(m^6)$ to $O(m^3)$.

A.3 Optimized Implementation

While Algorithm 1 in the main text illustrates the conceptual flow, the efficient implementation utilizes the cached structures described above. Here,

$\text{NormChange}(a, b)$ evaluates Δ_Ω in constant time using the current cached quantities R, G, K , and the topic sizes, as derived in eqs. (4) and (10).

Algorithm 2: Optimized Vendi Clustering (Fast Updates)

Input : Topic embeddings $\{e_i\}_{i=1}^m$, topic sizes $\{n_i\}_{i=1}^m$, target k_{target}

Output : Reduced topic set

Compute similarity matrix K with $K_{ij} = \cos(e_i, e_j)$;

Compute Gram matrix $G \leftarrow KK^\top$ and row-square sums R ;

while $m > k_{\text{target}}$ **do**

$(a^*, b^*, \Omega^*) \leftarrow \text{None}, \text{None}, +\infty$;

foreach topic pair (a, b) **do**

$\Delta_\Omega \leftarrow \text{NormChange}(a, b)$;

if $\Delta_\Omega < \Omega^*$ **then**

$\Omega^* \leftarrow \Delta_\Omega$;

$(a^*, b^*) \leftarrow (a, b)$;

$n_{\text{new}} \leftarrow n_{a^*} + n_{b^*}$;

Compute weighted centroid e_{new} for $a^* \cup b^*$;

Calculate normalized similarity row K_{new} from K_{a^*} and K_{b^*} ;

Replace a^* and b^* with the merged topic, updating K, G, R , and the topic sizes;

$m \leftarrow m - 1$;

return reduced topics;

B Implementation Details

This appendix details the embedding, dimensionality reduction, clustering, and topic representation components of the BERTopic base model from which all reduction methods start. All unspecified parameters are left at the corresponding library defaults.

Embedding model. We embed documents with sentence-transformers/all-MiniLM-L6-v2, producing 384-dimensional vectors. The same checkpoint is used for both datasets and across all three protocols.

Dimensionality reduction. Document embeddings are projected with UMAP (McInnes et al., 2020) prior to clustering, using `n_neighbors=15`, `n_components=5`, `min_dist=0.0`, and cosine distance. The `random_state` is set to the trial seed (42, 43, or 44) to make UMAP’s stochastic initialization reproducible within each trial; reported

Table 3: **Protocol 3 Results (SBERT)**: Comparison of Vendi Clustering (VC) vs. Direct HDBSCAN Tuning.

<i>Dataset: 20 Newsgroups</i>						<i>Dataset: AG News</i>					
Config	k	Method	C_v	NPMI	TD	Config	k	Method	C_v	NPMI	TD
mcs=30, eom	71	Vendi HDBSCAN	0.5954 0.5811	0.0605 0.0598	0.8254 0.6775	mcs=200, eom	84	Vendi HDBSCAN	0.8161 0.8011	0.2548 0.2346	0.9095 0.8476
mcs=30, leaf	103	Vendi HDBSCAN	0.6132 0.6451	0.0611 0.0816	0.8136 0.7282	mcs=200, leaf	107	Vendi HDBSCAN	0.8152 0.8283	0.2566 0.2544	0.8888 0.8308
mcs=40, eom	44	Vendi HDBSCAN	0.5759 0.5402	0.0591 0.0402	0.7591 0.5409	mcs=225, eom	71	Vendi HDBSCAN	0.8203 0.8132	0.2606 0.2377	0.9296 0.8606
mcs=40, leaf	67	Vendi HDBSCAN	0.5851 0.5940	0.0552 0.0653	0.8164 0.6299	mcs=225, leaf	93	Vendi HDBSCAN	0.8159 0.8346	0.2571 0.2563	0.9043 0.8409
mcs=50, eom	21	Vendi HDBSCAN	0.5585 0.5107	0.0518 -0.0013	0.6810 0.3190	mcs=250, eom	65	Vendi HDBSCAN	0.8183 0.8007	0.2605 0.2272	0.9323 0.8554
mcs=50, leaf	53	Vendi HDBSCAN	0.5875 0.5661	0.0624 0.0525	0.7962 0.5698	mcs=250, leaf	83	Vendi HDBSCAN	0.8154 0.8274	0.2546 0.2466	0.9145 0.8313

metrics are mean and standard deviation across the three seeds.

Table 4: Base-model min_cluster_size by protocol and dataset. HDBSCAN’s min_samples inherits this value by default.

Protocol	20 NG	AG News
Protocols 1, 2 (vs. AC, LLM)	15	100
Protocol 3 (vs. HDBSCAN)	15	50

Clustering. Topics are discovered by HDBSCAN (Campello et al., 2013) on the UMAP-reduced embeddings, using Euclidean distance and cluster_selection_method=eom. For Protocols 1 and 2 (section 5.1), this base model is reduced by VC, agglomerative clustering, or LLM-prompted merging. For Protocol 3 (section 5.2), the same base model is reduced by VC and compared against *separate* BERTopic models trained with stricter min_cluster_size values; the grids for these stricter comparison models are reported in Table 2.

C Additional Experimental Results

This appendix provides supplementary experiments and performance benchmarks that complement the main evaluation. Table 5 examines sensitivity of Vendi Clustering to the diversity order of q , while table 3 reports supplementary SBERT-based results for Protocol 3.

D Quantitative Analysis of the Trajectory

To assess whether the three regions observed in Figure 1 correspond to changes in the rate of diversity loss, we fit a continuous piecewise-linear model to

VS_2 as a function of k (Figure 3). We selected two integer breakpoints by exhaustive search, enforcing continuity between segments. The best fit identifies breakpoints at $k = 69$ and $k = 21$ and explains 99.5% of the variation in VS_2 ($R^2 = 0.9949$). The corresponding slopes are 0.0061, 0.2648, and 0.6856, capturing the near-flat, moderate-decline, and sharp-decline regimes described in the main text. The close agreement between these breakpoints and the boundaries observed across VS_2 , coherence, and topic diversity provides quantitative support for the three-phase interpretation.

E Timing Results

Table 6 reports wall-clock reduction time in seconds for Vendi Clustering (VC), Agglomerative Clustering (AC), and the LLM-prompted baselines for $k \in \{25, 50, 100\}$. VC and AC timings cover both the c-TF-IDF and SBERT representations; LLM timings cover Gemma-3-12b-it (local, 4-bit)

Table 5: q -sensitivity on 20 Newsgroups ($k = 50$). Vendi Clustering uses c-TF-IDF representations. Metric differences fall within seed variation, whereas $q = 2$ is approximately $80\times$ faster than the other orders.

q	C_v	NPMI	TD
0.1	0.612 ± 0.021	0.075 ± 0.011	0.863 ± 0.013
0.5	0.616 ± 0.018	0.077 ± 0.010	0.865 ± 0.017
1.0	0.619 ± 0.021	0.078 ± 0.011	0.860 ± 0.012
1.5	0.616 ± 0.023	0.077 ± 0.013	0.863 ± 0.008
2.0	0.619 ± 0.017	0.078 ± 0.012	0.867 ± 0.015
5.0	0.615 ± 0.020	0.076 ± 0.012	0.859 ± 0.015
∞	0.615 ± 0.020	0.076 ± 0.012	0.859 ± 0.015

q	0.1	0.5	1.0	1.5	2.0	5.0	∞
Time (s)	321.1	334.9	348.0	366.2	4.2	362.6	333.8

Table 6: **Reduction time in seconds.** Mean $\pm$ standard deviation across three seeds unless otherwise noted.

Method	20 Newsgroups			AG News		
	$k = 25$	$k = 50$	$k = 100$	$k = 25$	$k = 50$	$k = 100$
VC (c-TF-IDF)	2.84 ± 0.13	2.97 ± 0.08	2.82 ± 0.11	6.05 ± 0.50	5.83 ± 0.52	6.09 ± 0.32
VC (SBERT)	1.61 ± 0.04	1.86 ± 0.06	2.08 ± 0.03	2.98 ± 0.03	3.07 ± 0.05	3.28 ± 0.09
AC (c-TF-IDF)	2.07 ± 0.51	2.12 ± 0.53	2.50 ± 0.68	3.11 ± 0.24	3.37 ± 0.31	3.73 ± 0.26
AC (SBERT)	1.95 ± 0.13	2.06 ± 0.05	2.24 ± 0.10	2.87 ± 0.02	3.14 ± 0.19	3.31 ± 0.06
LLM (GPT-4o-mini)	330.6 ± 7.9	251.2 ± 5.1	117.9 ± 4.6	996.3 ± 340.4	786.1 ± 168.1	534.6 ± 18.3
LLM (Gemma-3-12b-it)*	656.4 ± 18.5	547.0 ± 20.5	293.5 ± 14.2	1005.8 ± 59.4	888.7 ± 55.8	626.3 ± 60.5

*For 20 Newsgroups, Gemma seed 42 halted at $k = 83$ and therefore did not reach $k \in \{25, 50\}$. The corresponding cells are averaged over seeds 43 and 44. At $k = 100$, all three seeds reached the target. This matches the seed handling in Table 1. All AG News cells and all other rows use three seeds.

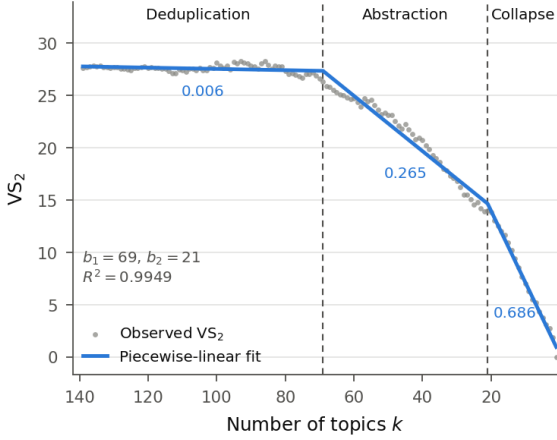


Figure 3: **Piecewise-linear fit to the VS_2 reduction trajectory.** Dashed lines mark the fitted breakpoints at $k = 69$ and $k = 21$, and blue values report the segment slopes dVS_2/dk (the k -axis runs from high to low). The fit captures the near-flat Deduplication regime, the steadier decline during Abstraction, and the sharp decline during Semantic Collapse ($R^2 = 0.9949$).

and GPT-4o-mini (API). Values are mean $\pm$ standard deviation across all three seeds (42, 43, 44) unless otherwise noted in the table footnotes. VC and AC are comparably fast, both completing in a few seconds, whereas the LLM baselines require hundreds to thousands of seconds; relative to the LLM baselines, VC’s speedup exceeds $40\times$ in every configuration and reaches $408\times$ in the fastest (SBERT) setting.

F Standalone Topic Models

Our primary experiments compare topic-reduction methods while holding the underlying BERTopic model fixed. A complementary question is whether post-hoc reduction is useful relative to fitting a topic model directly at the desired number of topics. We therefore compare Vendi-reduced BERTopic

with four standalone topic models trained natively at the target topic count: LDA (Blei et al., 2003), CombinedTM (Bianchi et al., 2021), ETM (Dieng et al., 2020), and FASTopic (Wu et al., 2024a). As shown in table 7, all five methods are evaluated at exactly $k \in \{25, 50, 100\}$ topics. Because these methods constitute different modeling pipelines, this experiment should be interpreted as a whole-pipeline comparison rather than a controlled evaluation of topic reduction alone.

F.1 Models and training

LDA, CombinedTM, and ETM are fitted with TopMost (Wu et al., 2024b). FASTopic is fitted with the reference fastopic package wrapped by TopMost. To reduce variation introduced by model-specific preprocessing choices, all four standalone models are trained on the same TopMost-processed corpus: lower-cased text with punctuation and English stop-words removed, excluding tokens containing digits or fewer than three characters, and truncated to the 10,000 most document-frequent terms.

LDA is fitted with Gensim’s LdaModel for 20 passes, with fixed symmetric priors $\alpha = \eta = 1/k$. CombinedTM is trained for 200 epochs with Adam (learning rate 0.002, batch size 200, 200 encoder units, and dropout 0.4). ETM is trained for 200 epochs with Adam (learning rate 0.002, batch size 200, 800 encoder units, and no dropout). Its word embeddings are initialized from pre-fitted 200-dimensional GloVe vectors (glove-wiki-gigaword-200) and held fixed during training. FASTopic is trained for 200 epochs in full-batch mode with Adam (learning rate 0.002); its transport parameters $DT_\alpha = 3.0$, $TW_\alpha = 2.0$, and temperature $\tau = 1.0$ are left at their defaults. FASTopic and CombinedTM receive the identical cached all-MiniLM-L6-v2 document vectors sup-

plied to BERTopic, whereas LDA operates only on bag-of-words counts, and ETM combines bag-of-words counts with pre-fitted GloVe word embeddings; neither uses document embeddings.

BERTopic retains the preprocessing and representation pipeline used throughout our primary experiments. The aim of this comparison is to evaluate the existing BERTopic + VC pipeline against models trained directly at the desired topic count, rather than to impose an identical training pipeline across fundamentally different model families. We instead standardize the downstream evaluation protocol across all methods.

F.2 Shared evaluation protocol

Although the standalone models share a common preprocessing pipeline, the modeling families still differ substantially in their representations, inference procedures, vocabularies, and mechanisms for constructing representative topic words. To make their outputs as directly evaluable as possible, we score every model using the same reference corpus, dictionary, and metric implementations. The reference corpus consists of the 18,846 20 Newsgroups documents with headers, footers, and quotes removed, tokenized into lowercased unigrams using BERTopic’s CountVectorizer analyzer.

Results are means and standard deviations over seeds 42, 43, and 44. For the standalone models, the seed changes model initialization while preprocessing remains fixed.

F.3 Interpreting cross-model metrics

Even under this shared evaluation protocol, the resulting scores are not equivalent to a controlled comparison of topic-reduction methods. In particular, differences in C_v and NPMI can reflect properties of how each modeling pipeline constructs and ranks its topic words in addition to differences in topic quality. This concern is less pronounced in our primary topic-reduction experiments because every method begins with the same BERTopic base model and inherits the same vocabulary and topic representations. Accordingly, the results in table 7 should be interpreted as complementary whole-pipeline evidence rather than as an isolated comparison of topic-reduction methods.

F.4 Embedding-based evaluation

Alongside C_v , NPMI, and Topic Diversity, we report embedding coherence (COH) (Thielmann

et al., 2024). COH is computed as the mean pairwise cosine similarity among the embeddings of each topic’s top ten words, averaged across topics, using paraphrase-MiniLM-L6-v2. Because COH evaluates semantic proximity in an external embedding space rather than corpus-level word co-occurrence, it provides a complementary view when comparing models that construct topic words differently.

G LLM Prompting Details

To ensure reproducibility, Figure 4 illustrates the exact prompt template used for the LLM-assisted reduction baseline (Protocol 2). The template is adapted directly from Janssens et al. (2025a), and both LLMs are prompted with this identical structure: the System Message, Demonstration Prompt, Demonstration Answer, and Instruction Prompt are the same across the two models. The *only* difference between the two configurations is the number of keywords shown per topic.

GPT-4o-mini is prompted with the full top-10 keywords per topic, exactly as shown in Figure 4. Gemma-3-12b-it is prompted with the same template, but each topic line is truncated to its top-5 keywords throughout. In the demonstration prompt this means topic 0 reads “astronaut, rocket, planet, galaxy, NASA” (the trailing five terms telescope, satellite, mission, orbit, cosmos are omitted), topic 2 reads “computer, laptop, technology, device, screen” (keyboard, processor, portable, battery, software are omitted), and so on for every topic in both the demonstration and the candidate topic list inserted at inference. The system message, demonstration answer, and instruction prompt are unchanged. This top-5 truncation is the minimum modification required to fit Gemma’s 8192-token context window when initial topic counts are in the $k \approx 150$ –200 range; we keep all other prompt content fixed so that any performance gap reflects model capability rather than prompt engineering.

Table 7: **Comparison with standalone topic models on 20 Newsgroups.** BERTopic with Vendi Clustering (VC) is compared with LDA, CombinedTM, ETM, and FASTopic trained directly at each target k . Values are means $\pm$ standard deviations across three seeds; best per k in **bold**.

k	Model	C_v	NPMI	COH	TD
25	BERTopic + VC	0.5815 $\pm$ 0.0104	0.0498 $\pm$ 0.0031	0.2552 $\pm$ 0.0031	0.8800 $\pm$ 0.0069
	LDA	0.6416 $\pm$ 0.0310	0.0591 $\pm$ 0.0230	0.2918 $\pm$ 0.0106	0.8680 $\pm$ 0.0243
	CombinedTM	0.5719 $\pm$ 0.0179	-0.0398 $\pm$ 0.0146	0.2423 $\pm$ 0.0169	0.9800 $\pm$ 0.0040
	ETM	0.5802 $\pm$ 0.0138	0.0219 $\pm$ 0.0122	0.3124 $\pm$ 0.0117	0.9240 $\pm$ 0.0106
	FASTopic	0.5576 $\pm$ 0.0396	-0.0911 $\pm$ 0.0080	0.2666 $\pm$ 0.0127	1.0000 $\pm$ 0.0000
50	BERTopic + VC	0.6189 $\pm$ 0.0173	0.0779 $\pm$ 0.0122	0.2454 $\pm$ 0.0055	0.8667 $\pm$ 0.0151
	LDA	0.6069 $\pm$ 0.0140	0.0242 $\pm$ 0.0071	0.2771 $\pm$ 0.0029	0.8740 $\pm$ 0.0040
	CombinedTM	0.5290 $\pm$ 0.0138	-0.0874 $\pm$ 0.0184	0.2199 $\pm$ 0.0008	0.9600 $\pm$ 0.0020
	ETM	0.5583 $\pm$ 0.0189	0.0038 $\pm$ 0.0066	0.3083 $\pm$ 0.0113	0.9020 $\pm$ 0.0069
	FASTopic	0.5759 $\pm$ 0.0127	-0.0882 $\pm$ 0.0080	0.2539 $\pm$ 0.0052	0.9893 $\pm$ 0.0042
100	BERTopic + VC	0.6326 $\pm$ 0.0065	0.0734 $\pm$ 0.0038	0.2484 $\pm$ 0.0032	0.8443 $\pm$ 0.0012
	LDA	0.5391 $\pm$ 0.0009	-0.0337 $\pm$ 0.0072	0.2631 $\pm$ 0.0042	0.8960 $\pm$ 0.0075
	CombinedTM	0.5161 $\pm$ 0.0122	-0.0878 $\pm$ 0.0081	0.2103 $\pm$ 0.0034	0.9220 $\pm$ 0.0130
	ETM	0.5331 $\pm$ 0.0081	-0.0282 $\pm$ 0.0042	0.3018 $\pm$ 0.0063	0.8550 $\pm$ 0.0108
	FASTopic	0.5595 $\pm$ 0.0078	-0.1106 $\pm$ 0.0026	0.2536 $\pm$ 0.0043	0.9523 $\pm$ 0.0191

System Message:

You are a helpful assistant tasked with identifying overlapping topics.

Demonstration Prompt:

Below I will provide a numbered list of topic words. Each line represents the topic words of a topic. Return the index of the two most similar topics in Python list format.

0: astronaut, rocket, planet, galaxy, NASA, telescope, satellite, mission, orbit, cosmos

1: apple, fruit, red, sweet, tree, orchard, juicy, vitamin, snack, fresh

2: computer, laptop, technology, device, screen, keyboard, processor, portable, battery, software

3: orange, fruit, citrus, vitamin, juicy, tropical, snack, fresh, tree, peel

4: car, vehicle, engine, road, drive, wheels, speed, travel, gasoline, transportation

5: book, library, read, pages, author, fiction, genre, chapter, knowledge, cover

Return only the Python list and nothing more!

Demonstration Answer: [1, 3]

Instruction Prompt:

Below I will provide a numbered list of topics. Each line represents the topic words of a topic. Return the index of the two most similar topics in Python list format.

[TOPIC_LIST_INSERTED_HERE]

Return only the Python list and nothing more!

Figure 4: **LLM Prompt Template (top-10 keywords, as used for GPT-4o-mini).** The pipeline provides a System Message, a Demonstration Example, and finally the Instruction Prompt containing the candidate topics. Gemma-3-12b-it uses the *identical* template, with each topic line truncated to its top-5 keywords (i.e., the trailing five terms of every line are removed) to fit its 8192-token context window. No other element of the prompt is altered between the two models.