Beyond Top Words: MonoTM for Topic Modeling with Interpretable Monosemantic Features

Una Joh and Bei Yu
School of Information Studies
Syracuse University
{sjoh01,byu}@syr.edu

Abstract

Topic models summarize large text corpora, but top-ranked words often provide only a limited representation of topic semantics. Sparse autoencoders (SAEs) offer a way to move beyond word-level descriptors by extracting interpretable features from dense representations, yet how feature interpretability relates to topic-inference quality remains unclear. We introduce **MonoTM**, an interpretable topic modeling framework that decouples these roles. Across three benchmark corpora, we show that document–topic mixture estimation and semantic interpretation favor different SAE configurations and feature subsets. MonoTM estimates mixtures from the full SAE bag-of-features representation and, with them fixed, learns topic descriptors over a separate vocabulary of corpus-grounded semantic features. This design preserves global topic structure while representing topics with semantic units more meaningful than individual words, making them more useful for downstream corpus analysis.

1 Introduction

Text is a primary source of evidence across the social sciences and the humanities, from political speeches and news coverage to historical archives and literary corpora. The scale of these collections motivates computational text-as-data methods; topic modeling, especially Latent Dirichlet Allocation (LDA) and its extensions, has therefore become a popular tool for exploratory corpus mapping and thematic summarization (Blei et al., 2003). Yet its role as scholarly evidence remains contested (Da, 2019; Grimmer and Stewart, 2013; Shadrova, 2021). A recent review of topic modeling validation in computational social science finds that validation and reporting practices remain heterogeneous and show little convergence across studies (Bernhard-Harrer et al., 2025).

One source of this issue is that most topic models define each topic as a probability distribution over words, and users typically interpret topics by their highest-probability words. However, assigning a concise, human-readable meaning or label to each inferred topic can be challenging in practice (Chang et al., 2009; Mei et al., 2007). Although recent neural topic models incorporate pretrained language models or document embeddings to capture semantic information beyond raw word counts, these models still typically extract top-ranked words or phrases post hoc to describe each topic to users (Bianchi et al., 2021a; Dieng et al., 2020; Grootendorst, 2022; Kardos et al., 2025).

This creates a gap between what topic models output—topics as distributions over words—and what many analysts want: semantic units that support interpretation at an intermediate level between reading individual documents and identifying coarse corpus-level themes. Beyond topic representation, words are rarely adequate semantic units for supporting meaningful downstream corpus analysis, such as analyzing relations between topics or documents based on word overlap.

Recent progress in mechanistic interpretability, particularly sparse autoencoders (SAEs) and autointerpretability methods, offers a way to extract human-readable semantic features from document embeddings produced by pretrained encoders. Motivated by our experimental findings, we propose **MonoTM**, an interpretable topic modeling framework that achieves competitive document–topic mixture estimation while producing validated, human-readable topic descriptors. [1]

[1]The official MonoTM repository is available at https://github.com/Una-J/MonoTM

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 645–682
November 6-10, 2026 ©2026 Association for Computational Linguistics

2 Background & Related Work

2.1 Mechanistic Interpretability and Sparse Autoencoders

Mechanistic interpretability seeks to explain neural networks by identifying internal components and how they compose into computations. A major obstacle is that neurons are often polysemantic, activating for multiple unrelated patterns (Olah et al., 2020). One influential account explains this via superposition, where models represent more features than they have dimensions by encoding features in an overcomplete set of directions, enabled by sparsity in the underlying factors (Elhage et al., 2022). This motivates moving from neurons to learned feature directions.

Two scalable strategies support this shift. First, Bills et al. (2023) employ language models to produce and score natural-language explanations of internal directions based on activation patterns. Second, sparse dictionary learning methods, especially sparse autoencoders (SAEs), learn sparse, overcomplete feature bases that often yield more monosemantic units than neurons and can support meaningful interventions (Bricken et al., 2023; Huben et al., 2024). Recent work also applies SAEs to dense text embeddings, recovering interpretable semantic factors useful for downstream control and analysis (O'Neill et al., 2024).

Building on these ideas, we investigate SAEs over document embeddings as a source of reusable semantic units for topic modeling, and adopt an Interpreter–Predictor style protocol to obtain validated feature labels for topic description.

2.2 Topic Modeling

Overview of topic models. Topic modeling began with probabilistic bag-of-words (BoW) generative models such as LDA (Blei et al., 2003), which represents each document as a mixture of topics and each topic as a distribution over words. Although LDA offers clear probabilistic semantics and interpretable topic descriptors, its BoW assumption ignores context and can produce frequency- and surface-form-sensitive topics. Neural topic models use inference networks for amortized variational inference instead of iterative per-document inference (ProdLDA; (Srivastava and Sutton, 2017)). Other neural parameterizations of document–topic mixtures include softmax and stick-breaking constructions (Miao et al., 2017). In parallel, embedding-aware models such as ETM represent topics in a word-embedding space to improve robustness to large vocabularies while maintaining interpretability (Dieng et al., 2020).

More recent approaches incorporate pretrained transformers to inject contextual semantics. A representative direction is contextualized topic modeling. CombinedTM uses contextual embeddings alongside BoW-style signals to improve topic quality (Bianchi et al., 2021a), and ZeroShotTM extends this idea with multilingual encoders to enable cross-lingual, zero-shot topic inference (Bianchi et al., 2021b). Beyond contextualization, recent work focuses on robustness and efficiency. ECRTM introduces embedding clustering regularization to discourage topic collapse and encourage more distinct topics (Wu et al., 2023). FASTopic further leverages pretrained transformers while modeling document–topic and topic–word relations in a unified framework designed to be fast, stable, and transferable (Wu et al., 2024b). Finally, S3 offers a decomposition-based alternative that treats topics as independent semantic axes in embedding space, discovered via ICA rather than clustering or neural decoders (Kardos et al., 2025).

SAE-based topic modeling. Recent work connects sparse autoencoders and topic modeling by leveraging mechanistic-interpretability insights and treating SAE latents as reusable semantic units in embedding/activation space. Girrbach and Akata (2025) view SAEs as continuous-space topic models by deriving the SAE objective as a MAP estimator under an LDA-like generative model. Zheng et al. (2025) propose Mechanistic Topic Models (MTMs), which featurize documents using a pretrained SAE over LLM activations and fit topic models over SAE feature counts. While their mLDA is closest to our LDA-with-bag-of-features approach, MTMs rely on pretrained SAEs and external feature descriptions.

Unlike these prior works, our goal is not simply to replace words with SAE latents as the vocabulary for topic modeling. We show that the features most useful for document–topic mixture estimation are not necessarily the same features whose labels can be reliably interpreted. This empirical mismatch motivates MonoTM, which decouples the statistical and interpretive roles of SAE features. Although this design is more costly than reusing pretrained SAE vocabularies, it targets settings such as social science and digital humanities, where corpus-specific interpretability, traceability,

and validation are often more important than fast topic discovery.

3 Research Questions

Our starting hypothesis is that SAE features can play two distinct roles in topic modeling. If a document embedding can be represented as a sparse, nonnegative linear combination of SAE features, and if those features are closer to monosemantic than the original embedding dimensions, then SAE features may serve both as human-readable semantic units and as useful units for estimating document–topic mixtures. These two roles are related but not identical: features that are easy to label may not be the same features that best support mixture estimation. We therefore first study these roles separately, and then ask how they can be combined in a single interpretable topic model:

1. Under what SAE configurations do corpus-trained features become reliably interpretable semantic units?

2. Under what SAE configurations do SAE features provide effective units for document–topic mixture estimation?

3. How can we use SAE features to build a topic model that produces interpretable semantic topic descriptors aligned with strong document–topic mixture estimates?

4 Experimental Setup

Datasets. To evaluate topic models with minimal subjectivity, we compare inferred topics against the gold category annotations provided by standard benchmarks (dataset statistics in Table 3). We use **(1) 20 Newsgroups** (user-generated Usenet newsgroup posts; scikit-learn version) (Lang, 1995), **(2) Web of Science** (WOS-46985) using abstracts as input (Kowsari et al., 2017), and **(3) Reuters** (Reuters-21578) with the ModApte split (Lewis, 1997) accessed via Hugging Face Datasets (Lhoest et al., 2021). Reuters is a multi-label and highly imbalanced collection of Reuters newswire stories. For Reuters, we keep labels appearing at least 30 times, yielding 47 topics.

Document embeddings and SAE training. We map each document d to a dense embedding $x_d \in \mathbb{R}^{4096}$ using NVIDIA's `llama-embed-nemotron-8b` model (Babakhin et al., 2025). We use the model's full context window of 32,768 tokens (NVIDIA, 2025); longer documents are truncated to their first 32,768 tokens. This affects only 18 documents in 20 Newsgroups.

For each corpus, we train top-K sparse autoencoders on dimension-wise standardized document embeddings. Each SAE learns a dictionary with mN latent features, where $m = 4096$ is the embedding dimension and N is the expansion factor. Thus, N controls the size of the learned feature dictionary, while K controls the maximum number of active features per document. Given a document embedding, the encoder produces nonnegative latent activations and retains only the largest K, yielding a sparse document–feature representation. The decoder reconstructs the standardized embedding from these active features.

The two SAE hyperparameters therefore have distinct roles: N controls dictionary capacity, while K controls per-document activity. We sweep 11 values of N from $1/64$ to 5 and eight values of K from 4 to 512, omitting configurations with $mN < K$. Full architectural, optimization, checkpoint-selection, and grid details are given in Appendix B.

Interpreter: feature label generation. Following and adapting the autointerpretability procedure of O'Neill et al. (2024), we pass the learned SAE features (decoder columns) to an autointerpretability module that proposes and later validates natural-language labels.

For each feature $f \in \{1, \ldots, mN\}$, we construct an Interpreter prompt using three sets of examples: **Max-activating examples**, the 10 documents with the highest activations $h_{d,f}$; **Typical-activating examples**, 10 documents sampled from the middle quantiles of the nonzero activation distribution for f; and **Zero-activating examples**, 10 documents where feature f is inactive ($h_{d,f} = 0$), sampled as negatives.

Given these examples, an *Interpreter LLM* outputs a short description ℓ_f (e.g., 4–10 words) intended to capture the single most prominent concept that is present in the activating texts but absent from the non-activating texts (see the Interpreter prompt in Figure 2).

Predictor: feature label validation. Labels produced by an Interpreter LLM can be plausible but unvalidated. For example, labels may describe a frequent corpus theme that is not specific to the feature. To test whether a label actually predicts feature behavior, we validate each ℓ_f using a separate Predictor LLM (Bills et al., 2023; O'Neill

et al., 2024).

For each feature f, we construct a balanced evaluation set $\mathcal{S}_f = \mathcal{S}_f^+ \cup \mathcal{S}_f^-$ by sampling 30 *positive* documents and 30 *negative* documents from the corpus. Positives satisfy $h_{d,f} > 0$ (feature f is active for document d), while negatives satisfy $h_{d,f} = 0$ (feature f is inactive). Let $y_{d,f} = \mathbf{1}[h_{d,f} > 0]$ denote the ground-truth activation indicator.

We run the Predictor as an independent per-document binary classification query. For each $d \in \mathcal{S}_f$, the Predictor LLM receives only the description ℓ_f and the raw document text, and outputs a binary prediction $\hat{y}_{d,f} \in \{0, 1\}$ indicating whether feature f would activate (see the Predictor prompt in Figure 2). We aggregate these document-level predictions into a feature-level interpretability score using the F1 score between $\{\hat{y}_{d,f}\}_{d \in \mathcal{S}_f}$ and $\{y_{d,f}\}_{d \in \mathcal{S}_f}$, and denote this score by $\mathrm{IS}(f)$.

Autointerpretability sweep and LLM setup. Running the Interpreter–Predictor protocol requires many LLM inference calls, so we apply it only to a subset of SAE configurations we expect to be most informative due to cost constraints: $N \in \{0.5, 1, 2, 3, 4, 5\}$ and $K \in \{4, 8, 16, 32\}$. Details of the LLMs used are in Appendix D.

5 Results

5.1 RQ1: When Do SAE Features Become Interpretable Semantic Units?

RQ1 asks whether corpus-trained SAE features can serve as reliable semantic units, and how this depends on SAE capacity and sparsity. We operationalize interpretability using the Interpreter–Predictor protocol described in Section 4: a feature is treated as interpretable when its natural-language label predicts held-out feature activations with an interpretability score above a validation threshold.

Empirical results. SAE features become interpretable semantic units most consistently under low-to-moderate per-document activity levels. Across all three datasets, configurations with smaller K have the highest fraction of activated features that pass validation (Tables 4, 5, and 6).

Increasing K changes this behavior. At permissive validation thresholds, larger K generally increases the absolute number of validated features, because more active features are exposed for labeling and validation. However, when averaged across the N values we evaluate, the fraction of activated features that are validated decreases monotonically with K on all datasets (Tables 4, 5, and 6).

The effect of dictionary capacity N is best understood conditional on the activity level K. On 20 Newsgroups and Reuters, increasing N at low K often reduces the number of activated features that surpass the activation threshold, while Web of Science is an exception: activated and validated counts increase with N even at low K (Tables 4, 5, and 6). This suggests that the value of increasing dictionary capacity depends on whether the corpus and activity level provide enough support for the larger feature space. At higher K, larger dictionaries are more often useful, as allowing documents to activate more features enables higher-capacity SAEs to expose a larger pool of validated semantic units.

Additionally, we report a secondary diagnostic analysis of feature granularity in Appendix F. This analysis suggests that increasing capacity and activity may shift the validated feature space toward finer semantic resolution.

Choosing (N, K) in practice. These results suggest that SAE configuration should be treated as a practical modeling choice. If the user wants a conservative, easy-to-audit feature vocabulary, the best starting point is a low- or moderate-activity SAE, such as $K = 4$ or $K = 8$, combined with a moderately large dictionary. This setting yields fewer features, but a larger share of them can be validated. If the user instead wants broader coverage or more fine-grained descriptors, it is reasonable to move toward larger K and larger N, but only with the expectation that post-hoc validation will discard a larger fraction of the activated feature space.

A useful practical procedure is therefore to begin with a high-precision anchor configuration and then expand only if the validated feature inventory is too coarse or too small. For corpora with many stable domain-specific distinctions, such as Web of Science, a moderate configuration such as $K = 8$ with a larger dictionary is a natural starting point. For noisier or more heterogeneous corpora such as 20 Newsgroups and Reuters, low-K configurations provide a cleaner initial semantic inventory.

5.2 RQ2: When Do SAE Features Support Document–Topic Mixture Estimation?

Parallel to RQ1's focus on interpretability, RQ2 asks whether corpus-trained SAE features can serve as useful statistical units for estimating

document–topic mixtures.

We test this hypothesis by replacing LDA's word vocabulary with SAE features. For each document, we treat active SAE features as pseudo-tokens and their activation magnitudes as token weights, producing a bag-of-features (BoF) representation. We deliberately choose LDA because it provides a transparent mixed-membership likelihood once document embeddings are converted into bags of SAE features. Neural topic models could also be applied to the same BoF inputs, but their neural parameterization would introduce additional representational capacity, making it harder to isolate whether the SAE feature vocabulary itself provides useful statistical units for topic inference.

RQ2 evaluates whether SAE features are useful for mixture estimation, regardless of their interpretability scores.

BoF+LDA setup. For each dataset, we train a top-K SAE and construct a sparse document–feature matrix $X \in \mathbb{R}_{\geq 0}^{D \times mN}$ where $X_{d,j}$ is the activation magnitude of feature j in document d. We then fit LDA using scikit-learn's `LatentDirichletAllocation` with variational EM for 200 iterations, and set the number of topics to the number of gold categories (20 Newsgroups: T=20, Web of Science: T=7, Reuters: T=47).

Baselines. We compare against (i) classical BoW-LDA (Blei et al., 2003) and (ii) other state-of-the-art neural topic models that output document–topic mixtures: CombinedTM (Bianchi et al., 2021a), ZeroShotTM (Bianchi et al., 2021b), ECRTM (Wu et al., 2023), FASTopic (Wu et al., 2024b), and S^3 (Kardos et al., 2025). We do not include BERTopic because, following the taxonomy of Wu et al. (2024a), it is a clustering-based topic discovery method rather than a model for document–topic mixture estimation. Including it in the topic–label alignment benchmark would therefore conflate mixture-estimation quality with embedding-cluster separability.

For each baseline, we apply the same topic–label alignment procedure as for BoW+LDA. To ensure a fair comparison, all baselines that use document embeddings—CombinedTM, ZeroShotTM, FASTopic, and S^3—are run with the same `llama-embed-nemotron-8b` document embeddings used by MonoTM.

Evaluation. We do not use standard word-based topic coherence or topic diversity as evaluation metrics because doing so would defeat the purpose of MonoTM. Metrics such as PMI, NPMI, C_V, and topic diversity were designed for topic representations expressed as ranked lists of keywords: they measure whether top words co-occur in a reference or training corpus, or whether top-word lists are lexically redundant across topics (Newman et al., 2009; Mimno et al., 2011; Lau et al., 2014; Röder et al., 2015; Dieng et al., 2020). These metrics would require projecting our descriptors back into keywords that represent each topic, thereby reintroducing the very representation that our method is designed to replace.

We instead evaluate document–topic mixture quality by aligning inferred topics with gold benchmark labels and reporting Micro-F1 and Macro-F1. Full evaluation details, including the single-label and multi-label alignment procedures, are given in Appendix G. For all models and datasets, we run three trials with different random seeds and report mean scores in Table 1; full SAE-grid heatmaps are reported in Appendix H.

5.2.1 Effective mixture estimation requires balanced SAE scaling

BoF+LDA performance depends systematically on the SAE hyperparameters. The full Micro-F1 and Macro-F1 heatmaps are reported in Appendix H. Across datasets, mixture quality is weakest in two regimes: very wide dictionaries paired with very small K, and boundary cases where $mN = K$. These results suggest that SAE features are most useful for topic inference when the representation maintains an explicit sparsity bottleneck while still allowing enough active features per document.

The best-performing configurations instead lie in a balanced scaling band where dictionary capacity and per-document activity increase together. We therefore summarize BoF+LDA in Table 1 by averaging over a fixed robust subset of this region, using $N \in \{0.25, 0.5, 1, 2\}$ and $K \in \{128, 256, 512\}$. This gives a less brittle comparison than selecting the single best hyperparameter setting for each dataset.

5.2.2 Mixture quality and interpretability do not fully coincide

The balanced-scaling pattern identified above should not be read as an interpretability result. Although Figure 4 identifies a robust region for document–topic mixture estimation, the same region is not necessarily the one that maximizes the

Model	20 Newsgroups		Web of Science		Reuters	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
BoW-LDA	0.3673	0.3515	0.5147	0.4702	**0.3446**	0.2021
CombinedTM	0.4929	0.4513	0.5514	0.5201	0.2398	0.1728
ZeroShotTM	0.3704	0.3398	0.4904	0.4670	0.2051	0.1467
ECRTM	0.5774	0.5368	0.5176	0.4916	0.2410	0.1557
FASTopic	0.5492	0.4912	0.5032	0.4583	0.2535	0.1582
S^3	0.4578	0.4220	0.4233	0.3942	0.1513	0.1100
BoF+LDA (Ours)	**0.6770**	**0.6489**	**0.5753**	**0.5446**	0.3188	**0.2089**

Table 1: Micro F1 and Macro F1 scores on 20 Newsgroups, Web of Science, and Reuters datasets. The best scores are in bold.

number of interpretable SAE features. This can be seen by comparing the mixture-quality heatmaps in Figure 4 with the interpretability counts in Tables 4, 5, and 6. For example, in 20 Newsgroups, holding $N = 0.5$ fixed and increasing K from 8 to 32 reduces the fraction of dictionary features with IS ≥ 0.80 from $721/(4096 \cdot 0.5) = 35.2\%$ to $282/(4096 \cdot 0.5) = 13.8\%$, while Micro/Macro F1 increases from 0.56/0.52 to 0.66/0.63. Thus, neither the absolute number nor the dictionary-level fraction of validated features fully predicts document–topic mixture quality. These results suggest a tension between choosing SAE configurations for feature-level interpretability and choosing them for document–topic mixture quality.

We test this tension more directly by asking whether restricting BoF+LDA to highly interpretable SAE features improves document–topic mixture estimation. Specifically, we filter features according to their interpretability score, retain only those with $\mathrm{IS}(f) \geq \tau$, and then re-fit LDA on the resulting filtered BoF representation. Figures 6–8 show that this filtering generally hurts document–topic mixture quality, and that the degradation becomes more severe as τ increases. These results indicate that failing strict interpretability validation does not imply that an SAE feature is uninformative for mixture estimation.

5.2.3 Bag-of-features yields strong document–topic mixture estimates

Table 1 shows that BoF+LDA is competitive across all three benchmarks, achieving the strongest results on both single-label datasets and the best Macro-F1 on Reuters. These gains support the core intuition that SAE latents provide a sparse, semantically structured pseudo-token vocabulary, allowing LDA to pool co-occurrence evidence more effectively than when using raw words.

Results on Reuters are nuanced. Since it is multi-label and highly imbalanced, Micro-F1 is dominated by frequent classes while Macro-F1 is sensitive to performance on rare labels. Here, BoW-LDA attains the highest Micro-F1, but BoF+LDA achieves the best Macro-F1 overall. This pattern is consistent with BoF features providing more discriminative evidence for minority topics that have limited word overlap with majority classes.

5.3 RQ3: Interpretable Topic Modeling (MonoTM)

RQ3 asks how the two roles of SAE features identified above can be combined into a single interpretable topic model. The findings from RQ1 and RQ2 suggest that a single SAE representation should not be forced to serve both feature-level interpretability and document–topic mixture estimation. MonoTM implements this idea by separating document–topic mixture estimation from topic interpretation.

The remainder of this section presents MonoTM in three steps. Section 5.3.1 describes the MonoTM algorithm. Section 5.3.2 audits whether the validated descriptor vocabulary omits important topic semantics from lower-validation features. Section 5.3.3 gives a compact descriptor comparison against word-based and free-form LLM descriptor baselines.

5.3.1 The MonoTM Algorithm

MonoTM separates the statistical and interpretive roles of SAE features. Let D be the number of documents, T the number of topics, and V the number of validated interpretable features used for topic description. MonoTM estimates document–topic mixtures from a full SAE bag-of-features representation, then estimates a topic–feature distribution over validated interpretable features with the mixtures held fixed.

Stage 1: estimating document–topic mixtures from all SAE features. For each document, the mixture SAE produces a sparse nonnegative activation vector over all SAE features. We construct a full document–feature matrix X^{mix} from these activations and fit LDA with T topics, yielding document–topic mixtures

$$\Theta \in \mathbb{R}_{\geq 0}^{D \times T}, \qquad \theta_d \in \Delta^T.$$

This stage uses all active SAE features, rather than only validated interpretable features, because Section 5.2.2 shows that filtering to interpretable features degrades mixture quality.

Stage 2: constructing an interpretable document–feature matrix. For topic interpretation, we use a validated feature set $\mathcal{F}_\tau$. A feature is retained if it activates in at least 30 documents, has a nonempty label, and satisfies $IS(f) \geq \tau$, with $\tau = 0.8$ in our experiments. For each retained feature f, we construct an interpretable document–feature matrix $C \in \mathbb{R}_{\geq 0}^{D \times V}$ by weighting its activation by its interpretability score:

$$c_{d,f} = \max(\tilde{h}_{d,f}, 0) \cdot IS(f), \qquad f \in \mathcal{F}_\tau.$$

Stage 3: estimating topic–feature distributions with fixed mixtures. Given fixed document–topic mixtures Θ and interpretable feature matrix C, MonoTM estimates a topic–feature distribution

$$B \in \mathbb{R}_{\geq 0}^{T \times V}, \qquad \beta_t \in \Delta^V,$$

by maximizing the fixed-mixture likelihood

$$\mathcal{L}(B) = \sum_{d=1}^{D} \sum_{f=1}^{V} c_{d,f} \log \left(\sum_{t=1}^{T} \theta_{d,t} \beta_{t,f} \right).$$

We optimize this objective with a Dirichlet-smoothed EM procedure; full update equations and implementation details are given in Appendix K.

After estimating B, each topic t is represented by the highest-probability validated feature labels under β_t. Appendix J provides sample MonoTM topic–feature representations produced by this final descriptor layer.

5.3.2 Audit: do excluded features hide important topic semantics?

To test whether the validated descriptor vocabulary omits important topic semantics, we audit the excluded features most likely to matter for topic interpretation. For all three datasets, we estimate Θ from the full bag-of-features representation of an SAE with $N = 0.5$ and $K = 256$. For interpretation, we use the $N = 2, K = 16$ SAE. The final MonoTM descriptor vocabulary contains active, validated features with $\text{IS}(f) \geq .8$. For the audit, we estimate an auxiliary all-feature topic–feature distribution B^{all} over all active interpretation-SAE features, using raw feature activations instead of weighting by $\text{IS}(f)$. This avoids mechanically suppressing the lower-validation features that the audit is designed to inspect.

We then rank all active interpretation features by their mass under B^{all} for each topic, and focus on lower-validation features with $\text{IS}(f) < .8$ that would have ranked among the top 20 features for at least one topic. These features are the strongest possible challenge to the validated descriptor vocabulary: they are highly topic-associated, but absent from the final MonoTM representation. For each such feature, assigned to the topic where it obtains its best rank, we search for a validated same-topic neighbor among the top-50 validated features of that topic under B^{all}. Appendix L gives the full construction, row alignment, EM objective, filtering rules, and formal definitions.

Coverage is measured in document-activation space. For each excluded feature, we compute cosine similarity between its raw document-level interpretation-SAE activation vector and the corresponding vectors for validated same-topic descriptors. We compare the nearest validated same-topic descriptor to a random validated descriptor sampled from the same topic-specific candidate pool. Figure 1 shows that nearest validated descriptors are much closer than random validated descriptors, especially for 20 Newsgroups and Web of Science.

The Reuters results are less straightforward, so we further characterize the excluded high-association features using heuristic label-type flags based on keywords in the generated feature labels. (See Appendix L for the keyword rules.) These flags are used to summarize broad tendencies in the excluded feature set, not as a validation metric.

Table 2 shows a clear dataset difference. In 20 Newsgroups and Web of Science, excluded high-association features are mostly semantic or broad-discourse labels. In Reuters, by contrast, 70.7% are flagged as format/register/numeric cues, suggesting that many of these features capture financial-news register, reporting format, or numeric conventions rather than missing core topic semantics. Con-

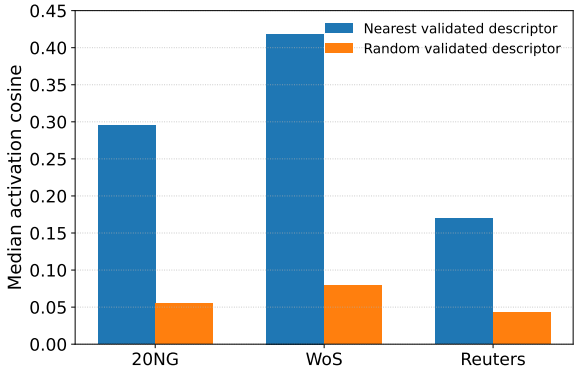


Figure 1: Coverage of lower-validation high-association features by validated same-topic descriptors. Bars show the median cosine similarity between feature activation profiles across documents.

Dataset	Semantic	Broad discourse	Format/register/ numeric
20NG	58.1%	36.4%	5.4%
WoS	88.5%	7.7%	3.8%
Reuters	6.3%	23.1%	70.7%

Table 2: Heuristic label-type split for lower-validation high-association features. Percentages are computed within each dataset.

crete examples of both patterns are provided in Appendix L.

Overall, while this audit does not prove that no excluded feature ever captures a meaningful omitted semantic dimension, it alleviates the concern that MonoTM arbitrarily omits a large class of independent topic-defining features.

5.3.3 Descriptor Comparison

Finally, we compare descriptor mechanisms under the same fixed document–topic mixtures Θ. This comparison is intended to isolate the descriptor layer by asking how different mechanisms describe the same inferred topics. We compare MonoTM against three descriptor baselines: θ-weighted TF-IDF terms, BERTopic-style c-TF-IDF terms (Grootendorst, 2022), and free-form LLM-generated descriptors. We include the LLM baseline because recent work increasingly treats LLMs as topic extractors (Lam et al., 2024; Mu et al., 2024; Pham et al., 2024). Full descriptor-generation details, prompts, and topic-name compression settings are provided in Appendix M. Full topic-name comparison tables are provided in Appendix O, and representative descriptor lists before topic-name compression are provided in Appendix P.

The comparison suggests two qualitative patterns. First, word-based descriptors often recover salient lexical anchors but remain surface-level. For example, in 20 Newsgroups, for the topic aligned with `talk.politics.misc`, the word-based baselines produce the topic name "Gun rights," whereas MonoTM yields "American culture wars," reflecting a broader mixture of gun-rights, sexuality, morality, and civil-liberties discussions (Table 12). Second, free-form LLM descriptors can be fluent but overly local to the sampled topic-associated documents. In Reuters, for example, the LLM baseline names the topic aligned with `crude` as "Ecuador oil crisis," whereas MonoTM yields the broader "OPEC oil market" (Table 14). These examples suggest that MonoTM descriptors offer a useful intermediate level of semantic abstraction: they bridge high-level topic labels and raw documents more effectively than word lists, while remaining more directly tied to the representations used for topic inference than ad hoc LLM descriptors.

Beyond topic naming, MonoTM also supports downstream analyses of relations among topics. Appendix Q illustrates this use case by comparing topic relations in the validated-feature space.

6 Conclusion

We introduced **MonoTM**, an interpretable topic modeling framework built on sparse autoencoders trained over document embeddings. Across three research questions, our results show that SAE features are useful for topic modeling in multiple but distinct ways. First, corpus-trained SAEs can produce relatively monosemantic units, but the availability of reliably interpretable features depends strongly on the choice of SAE configuration. Second, SAE activations can also serve as effective statistical units for estimating document–topic mixtures. However, the SAE configurations suited to these two roles do not necessarily coincide. Addressing RQ3, MonoTM resolves this mismatch by estimating document–topic mixtures from the full SAE representation and then learning topic–feature distributions over validated features with the mixtures fixed. Our analyses further show that excluded high-association features are typically covered by validated same-topic descriptors or reflect auxiliary non-core cues.

7 Limitations

Remaining uncertainty about excluded features. Section 5.3.2 audits features from the interpretation SAE that fall below the validation threshold and finds that genuinely hard-to-validate features account for only a small share of high-association topic-feature mass. However, this audit does not prove that all excluded features are unimportant or fully understood. Some low-validation features may encode corpus-specific regularities, pragmatic cues, formatting patterns, or culturally specific concepts that are difficult to summarize with short semantic labels. Moreover, our validation scores depend on the Interpreter and Predictor LLMs, so systematic blind spots in those models could still affect which features are included in the final descriptor set. For high-impact applications, MonoTM should therefore be paired with additional coverage audits and, when appropriate, human expert review.

Sensitivity to the Interpreter and Predictor LLMs. Our interpretability scores depend on the capabilities and failure modes of the specific LLMs used as Interpreter and Predictor, as well as prompt details and decoding settings. Different LLM backends may vary in (i) their ability to abstract from examples into stable hypotheses, (ii) calibration on the binary prediction task, and (iii) robustness to domain-specific language. As a result, both the number of features that validate above a threshold and the apparent granularity of validated features could change with the choice of model(s). While we partially mitigate this by separating generation (Interpreter) from validation (Predictor), our validation remains an LLM-mediated measurement rather than a ground-truth guarantee of monosemanticity.

Cost and practicality of LLM-heavy autointerpretability. MonoTM's feature labeling relies on many LLM calls, which is computationally expensive and can be slow in practice. This cost is a real barrier for iterative modeling workflows. That said, MonoTM is aimed at settings where researchers are willing to spend more compute to obtain a single, carefully audited interpretive lens on an important corpus (e.g., social science and digital humanities analyses where interpretability and traceability are first-order goals). Still, improving efficiency is crucial. Promising directions include cheaper feature screening models for "likely interpretable" features and distilling the Predictor into a small classifier. Additionally, continuing progress in SAE feature interpretation and tooling (e.g., Lieberum et al., 2024) suggests that MonoTM could become cheaper to apply while preserving its core advantages.

Systematic biases can distort the interpretable topic representation. MonoTM estimates document–topic mixtures using the full SAE vocabulary but represents topics using only the subset of features that pass LLM-based validation. If the Interpreter or Predictor has systematic blind spots—for example, consistently under-interpreting certain registers, dialects, domains, or culturally specific concepts—then those features may be disproportionately excluded from the interpretable feature set. In that case, the topic descriptors produced by MonoTM could omit important aspects of a topic even if those aspects strongly influence the inferred mixtures. This limitation is especially salient if the LLMs' biases correlate with sensitive attributes or with particular styles of expression in the corpus. Mitigations include auditing interpretability coverage across document subpopulations, validating with multiple LLM backends, and incorporating human expert review for high-impact analyses.

References

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *Preprint*, arXiv:2511.07025.

Jana Bernhard-Harrer, Randa Ashour, Jakob-Moritz Eberl, Petro Tolochko, and Hajo Boomgaarden. 2025. Beyond standardization: a comprehensive review of topic modeling validation methods for computational social science research. *Political Science Research and Methods*, pages 1–19.

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

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, and 5 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Nan Z. Da. 2019. The Computational Case against Computational Literary Studies. *Critical Inquiry*, 45(3):601–639. Publisher: The University of Chicago Press.

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. Place: Cambridge, MA Publisher: MIT Press.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Leander Girrbach and Zeynep Akata. 2025. Sparse Autoencoders are Topic Models. *arXiv preprint*. ArXiv:2511.16309 [cs].

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint*. ArXiv:2203.05794 [cs].

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models.

Márton Kardos, Jan Kostkan, Kenneth Enevoldsen, Arnault-Quentin Vermillet, Kristoffer Nielbo, and Roberta Rocca. 2025. S^3 – Semantic Signal Separation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 633–666, Vienna, Austria. Association for Computational Linguistics.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–28, New York, NY, USA. Association for Computing Machinery.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

David D. Lewis. 1997. Reuters-21578 text categorization test collection. *Distribution 1.0*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and*

data mining, pages 490–499, San Jose California USA. ACM.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering Discrete Latent Topics with Neural Variational Inference. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2410–2419. PMLR. ISSN: 2640-3498.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10160–10171, Torino, Italia. ELRA and ICCL.

David Newman, Sarvnaz Karimi, and Lawrence Cavedon. 2009. External evaluation of topic models. In *Proceedings of the 14th Australasian Document Computing Symposium*, pages 1–8. University of Sydney.

NVIDIA. 2025. nvidia/llama-embed-nemotron-8b (hugging face model card/readme). `https://huggingface.co/nvidia/llama-embed-nemotron-8b`. Accessed 2025-12-31.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Charles O'Neill, Christine Ye, Kartheik G. Iyer, and John F. Wu. 2024. Towards Interpretable Scientific Foundation Models: Sparse Autoencoders for Disentangling Dense Embeddings of Scientific Concepts.

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.

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.

Anna Shadrova. 2021. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, 2021. Publisher: Episciences.org.

Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. Effective Neural Topic Modeling with Embedding Clustering Regularization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 37335–37357. PMLR. ISSN: 2640-3498.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. FASTopic: pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pages 84447–84481, Red Hook, NY, USA. Curran Associates Inc.

Carolina Zheng, Nicolas Beltran-Velez, Sweta Karlekar, Claudia Shi, Achille Nazaret, Asif Mallik, Amir Feder, and David M. Blei. 2025. Model Directions, Not Words: Mechanistic Topic Models Using Sparse Autoencoders. *arXiv preprint*. ArXiv:2507.23220 [cs].

A Dataset Details

See Table 3 for summary statistics of the datasets.

Dataset	Points	Cats	Multi	No Label	Mean Chars	SD Chars	Median	Min	Max
20 Newsgroups	18846	20	X	0	1902.53	3984.97	1175	115	160616
Web of Science	46985	7	X	0	1376.46	492.77	1355	95	8019
Reuters	12561	47	O	2304	843.7	874.97	546	56	13396

Table 3: Summary statistics of the datasets used. The abbreviations in the header are defined as follows: Points (Total number of data points), Cats (Number of categories), and Multi (Multi-label classification, where 'O' denotes multi-label and 'X' denotes single-label). No Label indicates the number of samples without a label. The character count statistics (Mean, SD, Median, Min, Max) describe the distribution of document lengths.

B SAE Architecture and Training Details

For each document $d \in \mathcal{D} = \{1, \ldots, M\}$, let $x_d \in \mathbb{R}^m$ denote its document embedding and let $\tilde{x}_d \in \mathbb{R}^m$ denote the dimension-wise standardized embedding, using the corpus-level mean and standard deviation. In our experiments, $m = 4096$.

We use a single-layer top-K sparse autoencoder with hidden dimensionality mN, where N is the expansion factor. Depending on N, the dictionary may be undercomplete, complete, or overcomplete relative to the embedding dimension. The decoder is a matrix $D \in \mathbb{R}^{m \times (mN)}$, and the encoder uses tied weights $D^\top$:

$$z = D^\top \tilde{x} + b, \tag{1}$$

$$a = \mathrm{ReLU}(z). \tag{2}$$

Sparsity is enforced with a hard top-K operator,

$$h = \mathrm{TopK}(a, K), \tag{3}$$

which yields a nonnegative latent activation vector $h \in \mathbb{R}^{mN}_{\geq 0}$ with at most K active coordinates.

We additionally introduce a learnable per-feature gain vector $g \in \mathbb{R}^{mN}_{>0}$ to modulate feature magnitudes independently of decoder direction norms. We parameterize g in log-space to enforce positivity and define

$$\tilde{h} = h \odot g. \tag{4}$$

The reconstruction of the standardized embedding is

$$\hat{x} = D\tilde{h}. \tag{5}$$

After each update, decoder columns are normalized to unit norm to avoid arbitrary rescaling between decoder weights and latent activations.

We train each SAE with Adam using learning rate 10^{-4} and batch size 1024 for up to 300k steps. We select the checkpoint with the lowest normalized reconstruction error on a fixed monitoring split containing 10% of documents. The training objective is

$$\mathcal{L} = \frac{\|\tilde{x} - \hat{x}\|_2^2}{\|\tilde{x} - \bar{x}\|_2^2 + \varepsilon}, \tag{6}$$

where $\bar{x}$ is the batch mean.

We train a grid of top-K SAEs over (N, K):

$$N \in \{0.015625, 0.03125, 0.0625, 0.125, 0.25, 0.5, 1, 2, 3, 4, 5\},$$
$$K \in \{4, 8, 16, 32, 64, 128, 256, 512\}.$$

This gives up to $11 \times 8 = 88$ configurations. We omit configurations with $mN < K$, since they cannot activate K distinct features. When $mN = K$, the top-K operator retains all units, so the model reduces to a dense tied-weight autoencoder; we keep this boundary case as a diagnostic comparison.

C Prompts for the Interpreter–Predictor Module

See Figure 2 for the Interpreter prompt and the Predictor prompt.

Interpreter Prompt

You are an NLP researcher conducting an investigation into one common basis vector shared by several document embeddings. Your goal is to determine what semantic phenomenon this vector encodes—specifically, which linguistic cue, tone, word choice, narrative framing, concept, or topic is associated with high activation along this dimension.

INPUT DESCRIPTION: You will be given three inputs:
1. Max-Activating Examples — up to 15 texts that produce high activation on this feature, with activation scores.
2. Typical-Activating Examples — up to 15 texts that activate this feature but are not necessarily among the highest activations.
3. Zero-Activating Examples — up to 15 texts that produce zero activation on this feature.

OUTPUT DESCRIPTION: Using the inputs, complete the following:
1. From the MAX-ACTIVATING EXAMPLES, list potential topics, concepts, themes, frames, and linguistic features they share. Consider multiple granularities. Give greater weight to items more prominent in higher-activation examples.
2. Cross-check those candidates against the TYPICAL-ACTIVATING EXAMPLES. Keep features that generalize across both max and typical positives; de-emphasize items that only appear in extreme cases.
3. Using the ZERO-ACTIVATING EXAMPLES, rule out any surviving topics/concepts/features that also appear in the zero-activating set. Be systematic.
4. Based on steps 1–3, provide a concise, rational analysis of the cues, tone, wording, frames, concepts, or topics—at appropriate granularity—that are likely to activate this neuron. Prefer the simplest explanation that fits all evidence.
5. Pick the single most prominent component that aligns with this vector, and express it in 4–10 words as: `FINAL:<description>`.

Here are the max-activating examples: `<max activating examples>`
Here are the typical-activating examples: `<typical activating examples>`
Here are the zero-activating examples: `<zero activating examples>`

Work through the steps thoroughly and analytically to interpret the vector. Output should only be in this format: `FINAL:<description>`. Do NOT return anything after these 4–10 words.

Predictor Prompt

You are an NLP researcher investigating one of the basis vectors of a document-embedding space. Your goal is to predict whether a given text exhibits the feature encoded by this vector.

OUTPUT DESCRIPTION: Based on the feature description, predict whether the vector will activate for this text. If you predict activation, output 1. If not, output 0.

Here is the description of the feature encoded by the vector: `<description>`
Here is the text to predict: `<text>`

Provide the output only in one of the following forms: `PREDICTION: 1` or `PREDICTION: 0`. Do not include anything else.

Figure 2: The prompts used for the Interpreter–Predictor Module. The Interpreter generates a semantic description of the vector, which is then fed into the Predictor to validate predictive accuracy.

D LLM Models and Inference Settings

For Interpreter and Predictor LLM runs, we use open-weight, instruction-tuned Gemma 3 models: Gemma-3-27B-IT (FP8) as the Interpreter and Gemma-3-12B-IT (bfloat16) as the Predictor.[2] Both models are accessed via the DeepInfra API.[3] We use the smaller model for the Predictor because the task is a comparatively simple binary decision. For both Interpreter and Predictor calls, we set the maximum context length to 131,072 tokens.

[2] https://huggingface.co/google/gemma-3-27b-it and https://huggingface.co/google/gemma-3-12b-it.

[3] https://deepinfra.com.

E Counts of Interpretable Features Across Interpretability Score Thresholds

See Table 4, Table 5 and Table 6.

N	K	IS≥ 0.95	IS≥ 0.90	IS≥ 0.85	IS≥ 0.80
0.5	4	73	239	374	453
1	4	89	211	339	423
2	4	103	217	312	372
3	4	87	193	280	328
4	4	83	184	261	302
5	4	95	199	272	307
0.5	8	53	237	483	721
1	8	67	269	502	689
2	8	62	237	413	547
3	8	46	184	320	434
4	8	61	159	274	357
5	8	49	142	250	322
0.5	16	7	92	292	577
1	16	22	161	481	923
2	16	22	218	564	1019
3	16	24	175	443	721
4	16	28	163	326	476
5	16	49	194	365	512
0.5	32	1	22	84	282
1	32	5	45	207	604
2	32	9	83	375	993
3	32	33	278	851	1759
4	32	73	462	1081	1795
5	32	91	540	1232	1923

Table 4: Number of features above interpretability score thresholds in the 20 Newsgroups dataset. IS denotes the interpretability score.

N	K	IS≥ 0.95	IS≥ 0.90	IS≥ 0.85	IS≥ 0.80
0.5	4	224	470	587	686
1	4	276	568	741	858
2	4	383	759	958	1063
3	4	438	832	1036	1169
4	4	467	861	1086	1188
5	4	509	936	1121	1254
0.5	8	247	648	934	1122
1	8	379	924	1294	1555
2	8	594	1231	1674	1983
3	8	655	1359	1816	2130
4	8	709	1443	1910	2195
5	8	657	1439	1924	2204
0.5	16	105	466	863	1222
1	16	343	1037	1733	2331
2	16	574	1734	2573	3274
3	16	704	1900	2806	3536
4	16	687	1899	2835	3567
5	16	664	1826	2855	3551
0.5	32	21	178	422	773
1	32	92	543	1199	1924
2	32	293	1301	2532	3857
3	32	403	1683	3169	4656
4	32	443	1808	3435	5120
5	32	376	1724	3360	5082

Table 5: Number of features above interpretability score thresholds in the Web of Science dataset. IS denotes the interpretability score.

N	K	IS≥ 0.95	IS≥ 0.90	IS≥ 0.85	IS≥ 0.80
0.5	4	59	152	213	277
1	4	50	120	187	226
2	4	58	122	180	210
3	4	42	108	169	204
4	4	46	116	163	196
5	4	47	108	158	190
0.5	8	39	123	263	406
1	8	43	152	270	372
2	8	38	128	202	279
3	8	42	105	184	261
4	8	32	93	166	236
5	8	31	78	143	192
0.5	16	21	102	238	433
1	16	24	125	266	467
2	16	21	106	206	347
3	16	31	102	193	284
4	16	42	146	251	358
5	16	55	168	282	400
0.5	32	3	30	121	285
1	32	8	66	220	515
2	32	33	177	446	756
3	32	57	269	570	881
4	32	58	272	584	983
5	32	55	289	654	1090

Table 6: Number of features above interpretability score thresholds in the Reuters dataset. IS denotes the interpretability score.

F Feature Granularity Across SAE Configurations

F.1 Motivation

In the main text, RQ1 focuses on whether corpus-trained SAE features can be reliably interpreted as semantic units. As a secondary diagnostic, we also ask whether SAE configurations differ in the granularity of the validated semantic features they recover. If larger or more active SAEs refine features learned by smaller SAEs, then document-level activations from the larger SAE should more easily reconstruct the validated-feature activations of the smaller SAE than vice versa.

This analysis is not required for MonoTM, but it helps characterize how SAE configuration affects the structure of the validated feature space.

F.2 Probe setup

Consider two SAEs trained on the same corpus but with different configurations (N, K), denoted A and B. Let $\mathcal{F}^{(A)}$ and $\mathcal{F}^{(B)}$ be the subsets of latent features that pass our label validation threshold. We use IS ≥ 0.80 throughout. Let $p = |\mathcal{F}^{(A)}|$ and $q = |\mathcal{F}^{(B)}|$, and let M be the number of documents.

F.3 Document–feature activation matrices

For each document d, each SAE produces a sparse top-K activation list $\{(i, a_{di})\}_{i \in \mathrm{TopK}(d)}$ with $a_{di} \geq 0$. We construct a dense document–feature activation vector by accumulating these latent activations into the coordinates corresponding to validated features. For SAE A, we define $x_d^{(A)} \in \mathbb{R}^p_{\geq 0}$ by

$$x_{dj}^{(A)} = \sum_{i \in \mathrm{TopK}^{(A)}(d)} a_{di}^{(A)} \, \mathbf{1}\left[i = f_j^{(A)}\right], \quad (7)$$

where $(f_1^{(A)}, \ldots, f_p^{(A)})$ is a fixed ordering of $\mathcal{F}^{(A)}$. Equivalently, we keep only validated feature coordinates and set all others to zero. Stacking over documents yields $X^{(A)} \in \mathbb{R}^{M \times p}_{\geq 0}$. We analogously construct $X^{(B)} \in \mathbb{R}^{M \times q}_{\geq 0}$.

F.4 Linear mapping probe

We fit an affine linear map in each direction:

$$\widehat{X}^{(B)} = X^{(A)} W_{A \to B}^\top + \mathbf{1}\, b_{A \to B}^\top, \quad (8)$$

$$\widehat{X}^{(A)} = X^{(B)} W_{B \to A}^\top + \mathbf{1}\, b_{B \to A}^\top, \quad (9)$$

where $W_{A \to B} \in \mathbb{R}^{q \times p}$ and $b_{A \to B} \in \mathbb{R}^q$, and symmetrically for $B \to A$. Parameters are estimated on a training split using ridge regression.[4]

F.5 Error metric and directional asymmetry

Because document–feature activations are nonnegative and sparse, we evaluate reconstruction with a weighted MSE that assigns higher weights to nonzero target entries:

$$\mathrm{WMSE}(\widehat{Y}, Y) = \frac{\sum_{d,j} w_{dj} (\widehat{y}_{dj} - y_{dj})^2}{\sum_{d,j} w_{dj}}, \qquad w_{dj} = 1 + (\lambda - 1)\, \mathbf{1}[y_{dj} > 0], \quad (10)$$

with $\lambda = 10$ in our experiments.

To make scores comparable across different target spaces, we normalize by the WMSE of a zero-predictor baseline computed on the same validation set:

$$\rho(A \to B) = \frac{\mathrm{WMSE}(\widehat{X}^{(B)}, X^{(B)})}{\mathrm{WMSE}(0, X^{(B)})}. \quad (11)$$

Lower ρ indicates more accurate reconstruction relative to predicting all-zero activations.

Our main statistic is the directional asymmetry

$$\Delta = \rho(\text{big} \to \text{small}) - \rho(\text{small} \to \text{big}), \quad (12)$$

where "big" and "small" refer to the ordered SAE configurations used in this probe. Negative values indicate that the larger SAE more easily reconstructs the validated-feature activation patterns of the smaller SAE than the reverse direction.

F.6 Resampling protocol

For each dataset, we repeat the full procedure over 10 random train/validation splits of documents. For each split, we fit both directions, $A \to B$ and $B \to A$, and compute ρ on the held-out validation documents. We evaluate all ordered pairs among the four SAE configurations shown in Figure 3.

F.7 Results

Figure 3 shows directional asymmetries in cross-SAE predictability. On Web of Science and Reuters, mappings from higher-capacity configurations to lower-capacity ones consistently achieve lower normalized error than the reverse direction

[4] Since our goal is to quantify reconstructability rather than minimize training error, we use ridge regression to control coefficient magnitudes and reduce sensitivity to resampling, yielding a lower-variance reconstructability estimate.

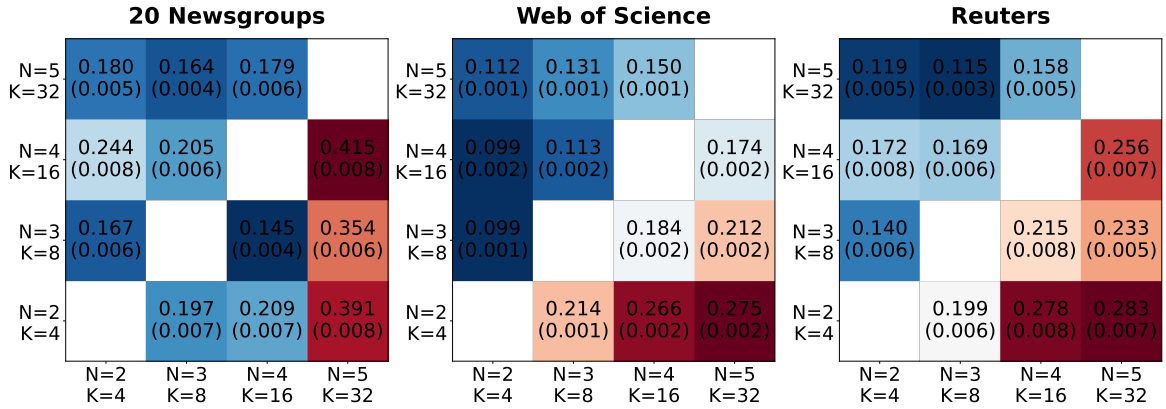


Figure 3: Heatmaps of linear cross-SAE reconstruction. Each cell shows the mean validation MSE ratio ρ for predicting the target SAE activations from the source SAE activations, with direction shown as row $\rightarrow$ column. Parentheses give the standard deviation over 10 train/validation splits. Lower is better.

Dataset	Median Δ	95% boot. CI	$\Pr(\Delta < 0)$
20 Ng	-0.111	$[-0.223,\ 0.052]$	0.667
WoS	-0.099	$[-0.164,\ -0.048]$	1.000
Reuters	-0.102	$[-0.139,\ -0.057]$	1.000

Table 7: Summary of cross-SAE predictability asymmetry across datasets. Reported are the median Δ and a 95% hierarchical bootstrap confidence interval over configuration pairs and random splits, along with $\Pr(\Delta < 0)$.

across essentially all configuration pairs. Aggregating across all pairs and seeds, the median Δ is negative on both datasets (Table 7), indicating that larger SAEs more easily reconstruct smaller SAEs than vice versa under this linear probe.

For 20 Newsgroups, the aggregate median Δ is also negative but less conclusive under resampling. The asymmetry is strongest when comparing the most separated configurations, $(N{=}2, K{=}4)$ and $(N{=}5, K{=}32)$, consistent with a larger gap in representational resolution.

F.8 Interpretation and limitations

Taken together, these results are consistent with a granularity shift as SAE capacity and activity increase. Features learned in larger SAEs appear to contain sufficient information to linearly reconstruct the validated-feature activations of smaller SAEs, while the reverse reconstruction is systematically harder.

This probe does not establish one-to-one feature correspondences, and it may underestimate nonlinear relationships between feature spaces. It should therefore be interpreted as a diagnostic of representational resolution rather than as direct evidence that individual features split cleanly across SAE configurations.

G Document–Topic Mixture Evaluation Details

We evaluate document–topic mixture quality through alignment with gold label topics. This evaluation is intended to measure whether the inferred topic mixtures recover the benchmark category structure, rather than whether topic descriptors form coherent word lists.

For 20 Newsgroups and Web of Science, which are single-label datasets, we assign each document to its highest-probability inferred topic,

$$\hat{z}_d = \arg\max_k \theta_{d,k}.$$

We then compute the confusion matrix between inferred topics and gold labels and use the Hungarian algorithm to find the one-to-one topic–label matching that maximizes total agreement. After applying this mapping, we treat the aligned topic assignments as multi-class predictions and report Micro-F1 and Macro-F1.

Reuters is multi-label and highly imbalanced, so we evaluate it as a multi-label topic–label alignment problem. For each candidate threshold

$$t \in \{0.1, 0.2, 0.3, 0.4, 0.5\},$$

we predict topic k for document d when $\theta_{d,k} > t$. We then compute pairwise F1 scores between each gold label topic and each inferred topic, use the Hungarian algorithm to obtain a one-to-one label–topic matching, and compute Micro-F1 and Macro-F1 over the matched pairs. We apply the same

threshold grid to all models and report each metric at its best threshold over this grid. Documents with no Reuters gold label are retained in the evaluation, so a model can avoid false positives on these documents only by assigning all topic probabilities below the selected threshold.

H Full SAE Hyperparameter Heatmaps for Mixture Estimation

Figure 4 reports Micro-F1 and Macro-F1 over the full SAE hyperparameter grid for BoF+LDA on each dataset. Figure 5 provides a complementary aggregate view by averaging performance across datasets after scaling each dataset's F1 scores to the $[0, 1]$ range. Together, these figures support the hyperparameter-selection discussion in Section 5.2.1 by showing how document–topic mixture quality varies with dictionary capacity N and per-document activity K.

Across datasets, two patterns are visible. First, very wide SAEs paired with very small K tend to perform poorly. In this regime, the dictionary has high capacity, but each document is allowed to activate only a small number of features. The resulting BoF representation is therefore too constrained: it exposes many possible pseudo-token types, but provides too little per-document evidence for stable topic inference. Second, boundary cases where $mN = K$ are also weak. Since the top-K operator retains all available units in this case, the model no longer imposes a meaningful sparsity bottleneck, which appears to reduce the usefulness of SAE latents as topic-modeling signals.

By contrast, stronger performance appears in a balanced scaling band where N and K increase together. These configurations preserve an effective sparsity constraint while allowing each document to express a richer set of SAE features. This pattern is visible in the per-dataset heatmaps in Figure 4 and remains apparent in the dataset-aggregated heatmaps in Figure 5. The purple region in Figure 4 marks the robust subset of this balanced regime used to summarize BoF+LDA in Table 1. We use this region rather than a single best cell to avoid overemphasizing dataset- or seed-specific hyperparameter choices.

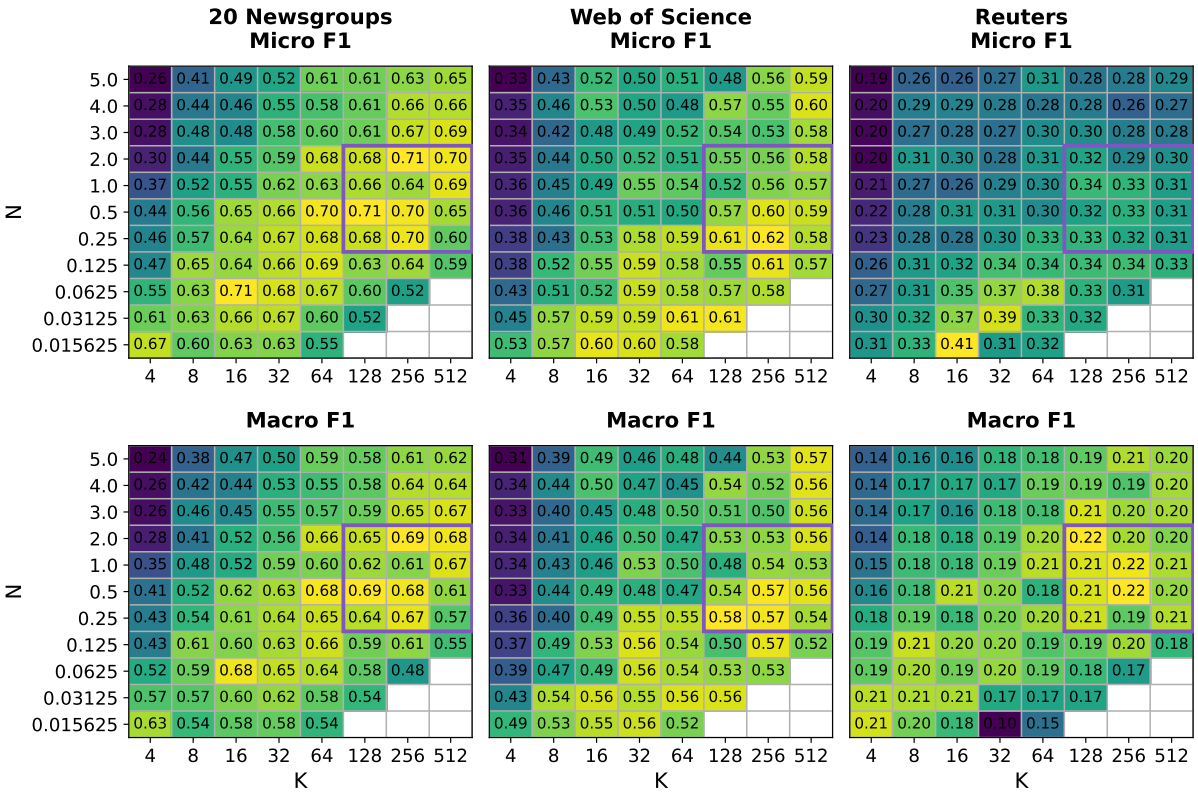


Figure 4: Heatmaps of Micro-F1 and Macro-F1 scores over the full SAE hyperparameter grid for BoF+LDA. The purple box indicates the (N, K) region whose values are averaged to produce the scores reported in Table 1.

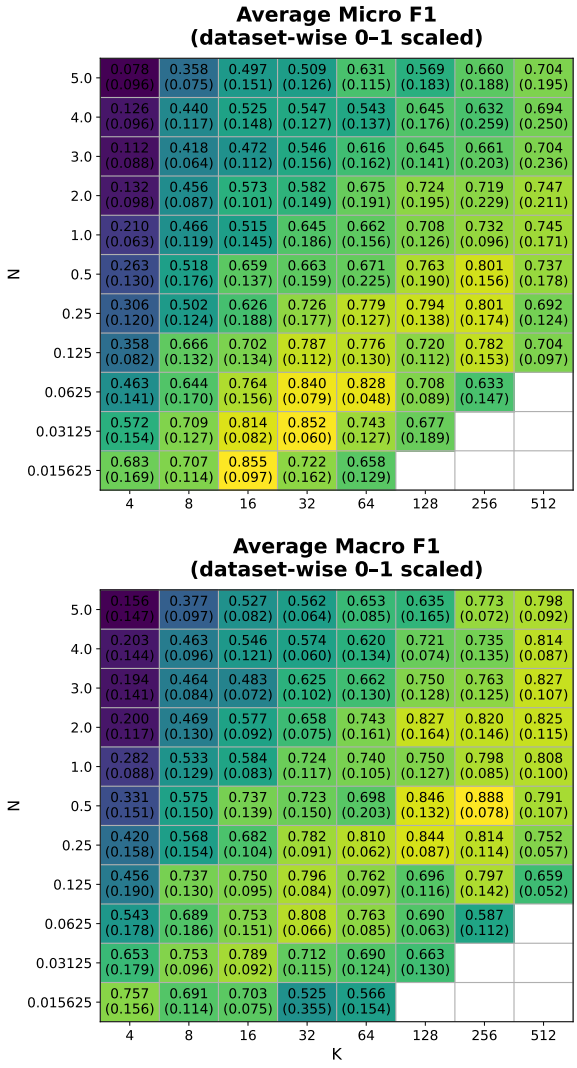


Figure 5: Heatmaps of Micro-F1 (top) and Macro-F1 (bottom) scores, shown as mean (standard deviation) over three datasets, computed from 9 data points for each (N, K). F1 scores were scaled to the $[0, 1]$ range on a per-dataset basis to prevent any single dataset from dominating performance.

I Document–Topic Estimation Using Only Interpretable Features

See Figure 6, Figure 7, and Figure 8.

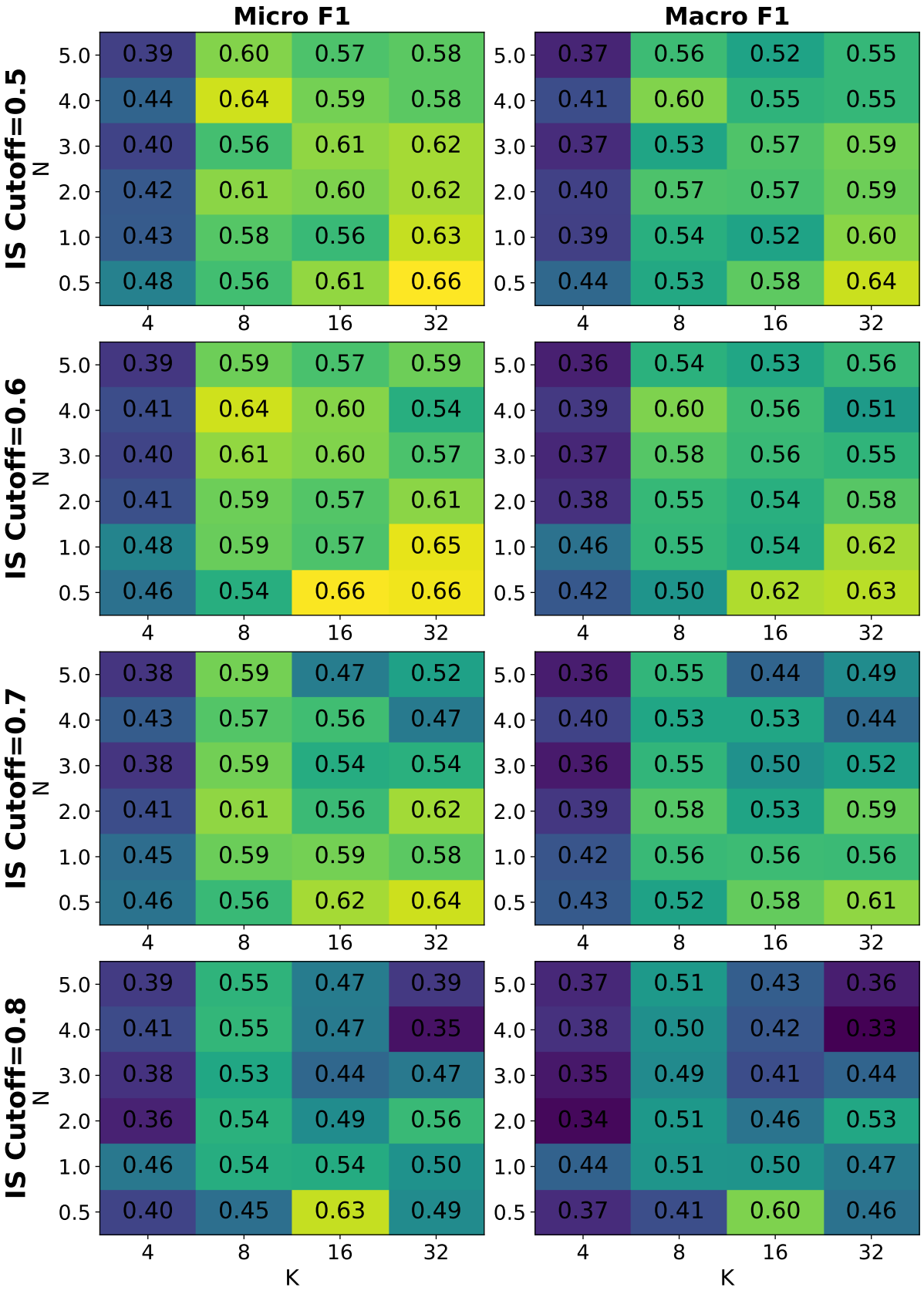


Figure 6: Heatmaps of micro-F1 and macro-F1 scores for document–topic mixtures on the 20 Newsgroups dataset, inferred using only features with interpretability scores (IS) above various cutoffs.

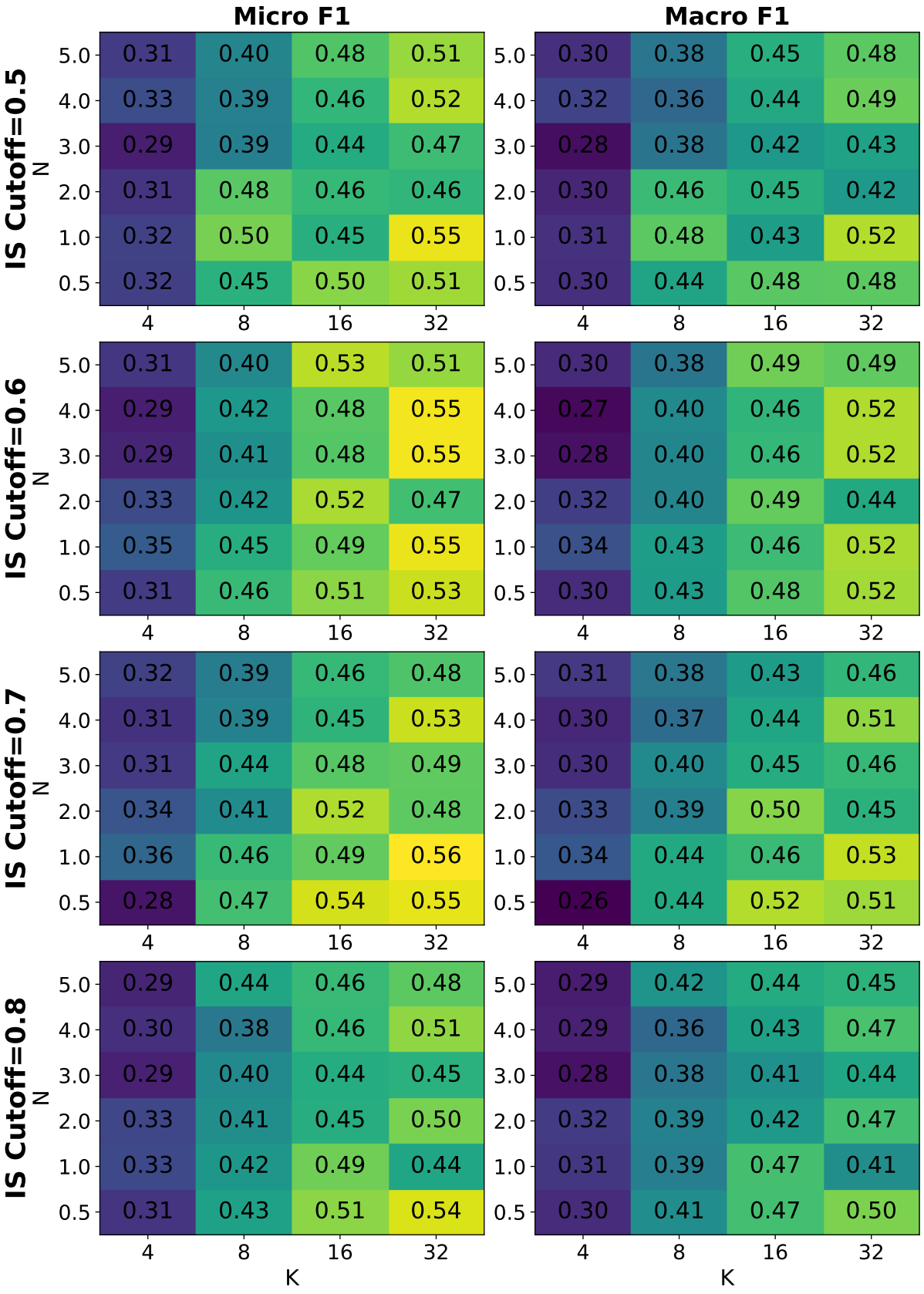


Figure 7: Heatmaps of micro-F1 and macro-F1 scores for document–topic mixtures on the Web of Science dataset, inferred using only features with interpretability scores (IS) above various cutoffs.

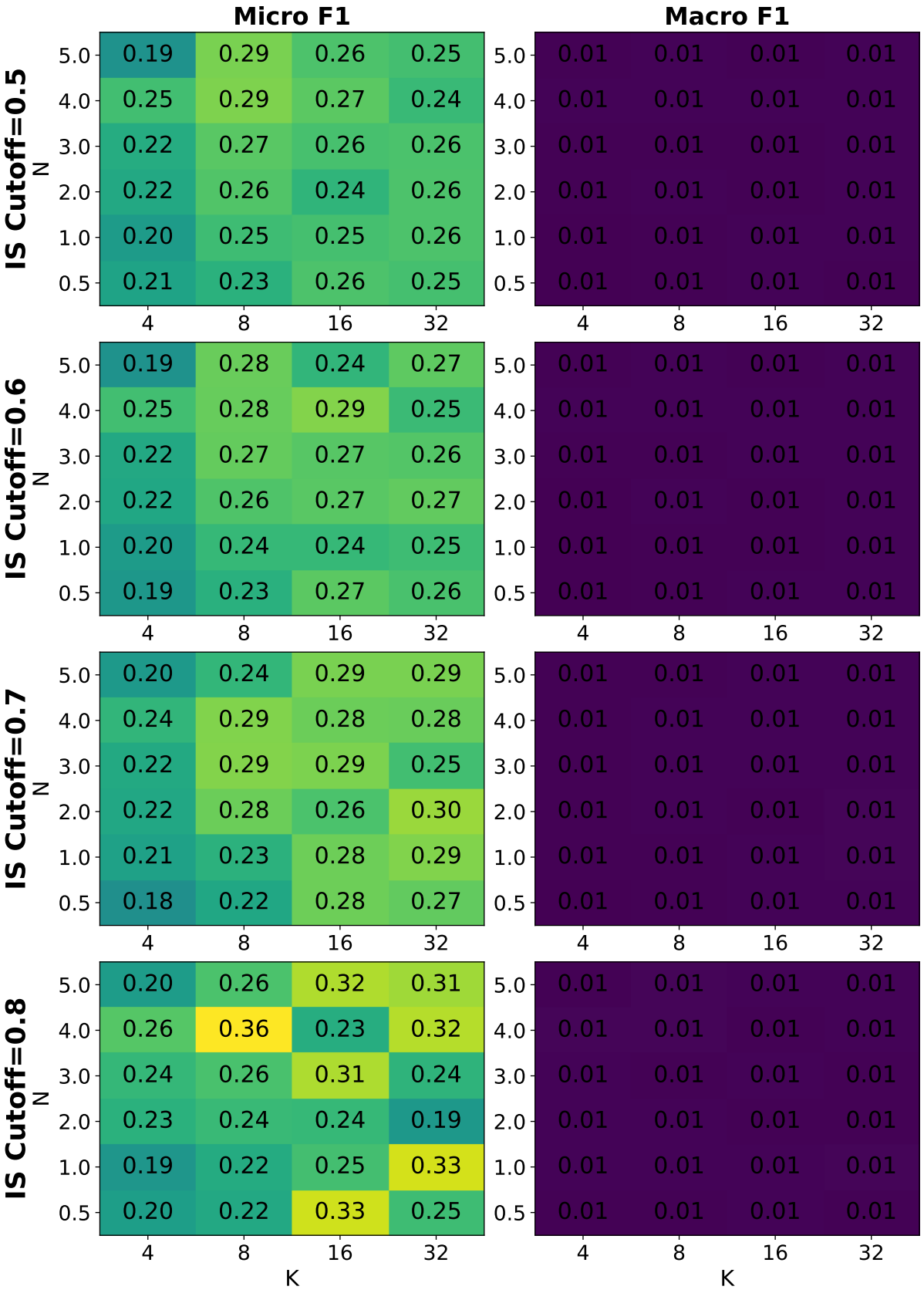


Figure 8: Heatmaps of micro-F1 and macro-F1 scores for document–topic mixtures on the Reuters dataset, inferred using only features with interpretability scores (IS) above various cutoffs. To evaluate against the multi-label gold labels, we treat topics whose inferred document–topic probabilities exceed 0.2 as predicted classes.

J Sample Topic–Feature Representations

Table 8 presents a subset of MonoTM's topic representation output, reporting the top-5 labeled features per topic for three topics in each dataset.

We use an interpretable-feature SAE with $(N{=}1, K{=}8)$ and filter features by $\mathrm{IS}(f) \geq 0.8$. This is not the richest configuration we tested: larger settings, such as $(N{=}5, K{=}32)$, yield substantially more interpretable features. However, the top feature labels for each topic show that even with interpretable features obtained from a relatively small SAE, the feature labels associated with high β values align well with the gold topic labels and provide a finer-grained understanding of the corpus than topic labels alone.

Dataset	Topic	Feature label	β
20 Newsgroups	rec.sport.hockey	NHL hockey game and player information	.402
20 Newsgroups	rec.sport.hockey	Sports scores reported on Usenet newsgroups	.048
20 Newsgroups	rec.sport.hockey	Frustrated fan complaints about sports broadcasting decisions	.038
20 Newsgroups	rec.sport.hockey	Passionate sports discussion in online forums	.031
20 Newsgroups	rec.sport.hockey	Sports stats requests and early internet culture	.030
20 Newsgroups	soc.religion.christian	Intense Christian theological debate and interpretation	.348
20 Newsgroups	soc.religion.christian	Religious debate, interpretation, and personal belief defense	.061
20 Newsgroups	soc.religion.christian	Traditional Catholic theology and liturgical concerns	.041
20 Newsgroups	soc.religion.christian	Skepticism towards religious belief and dogma	.040
20 Newsgroups	soc.religion.christian	Christian theological discussion and apologetic reasoning	.030
20 Newsgroups	rec.motorcycles	Motorcycle riding advice and safety discussion	.481
20 Newsgroups	rec.motorcycles	Motorcycle countersteering technique and related debate	.035
20 Newsgroups	rec.motorcycles	Discussions of drunk driving and motorcycles	.033
20 Newsgroups	rec.motorcycles	Motorcycle culture, camaraderie, road-based social signaling	.032
20 Newsgroups	rec.motorcycles	Motorcycle enthusiast discussion: ownership, specs, and banter	.031
Web of Science	Medical	Cardiovascular disease research and clinical information	.016
Web of Science	Medical	Gastrointestinal function, microbiome, and symptomology	.016
Web of Science	Medical	Bladder dysfunction, urological assessment, and clinical research	.015
Web of Science	Medical	Osteoporosis research, bone density, and skeletal health	.014
Web of Science	Medical	Neurobiological basis of psychiatric disease research	.013
Web of Science	Psychology	Internal psychological states and research analysis	.029
Web of Science	Psychology	Parenting challenges, family stress, child well-being	.025
Web of Science	Psychology	Social judgment and character attribution processes	.020
Web of Science	Psychology	Psychiatric comorbidity, distress, and vulnerability factors	.019
Web of Science	Psychology	Autism spectrum disorder and related neurodevelopmental features	.017
Web of Science	CS	Advanced computer and network security threat analysis	.030
Web of Science	CS	Computer vision algorithm development and evaluation	.023
Web of Science	CS	Parallel computing, performance optimization, technical implementation details	.022
Web of Science	CS	Cryptographic security and data privacy techniques	.019
Web of Science	CS	Machine learning methodology and technical rigor	.016
Reuters	earn	Comparative financial performance reporting, earnings vs prior period	.516
Reuters	earn	Comparative financial performance, "X vs Y" reporting	.175
Reuters	earn	Quarterly financial report key metrics comparison	.073
Reuters	earn	Financial turnaround, profit versus loss reporting	.063
Reuters	earn	Corporate quarterly financial results reporting and metrics	.054
Reuters	acq	Corporate acquisition and ownership transfer announcements	.311
Reuters	acq	Corporate mergers and acquisitions announcements, financial details	.089
Reuters	acq	Mergers, acquisitions, and corporate restructuring events	.088
Reuters	acq	Corporate merger and acquisition announcements (LOI)	.081
Reuters	acq	Energy sector financial deal reporting details	.051
Reuters	money-fx	Central bank money market interventions and rates	.366
Reuters	money-fx	Bank of England money market liquidity revisions	.227
Reuters	money-fx	Central bank money market intervention reporting	.097
Reuters	money-fx	Bank of England liquidity intervention reporting	.075
Reuters	money-fx	British monetary policy and currency market signals	.033

Table 8: Part of the final MonoTM output showing topic–feature distributions. The table lists the top five validated feature labels for three topics from each dataset. Gold labels from the original datasets are assigned using the alignment procedure in Section 5.2 and are shown only for orientation. β denotes the probability of each feature within the topic–feature distribution.

K Full MonoTM Algorithm and Inference Details

This appendix expands the MonoTM algorithm summarized in Section 5.3.1. Let $\mathcal{D} = \{1, \dots, D\}$ index documents and let T be the number of topics. MonoTM estimates document–topic mixtures from a full SAE bag-of-features representation, then estimates a topic–feature distribution over validated interpretable SAE features with the mixtures held fixed.

We allow the SAE used for mixture estimation and the SAE used for interpretation to be different. We denote their activations by $\tilde{h}^{\text{mix}}$ and $\tilde{h}^{\text{int}}$, respectively. The single-SAE case is recovered by setting these two representations to be the same. For each document d, let $\theta_d \in \Delta^T$ denote its document–topic mixture, and let $\Theta \in \mathbb{R}_{\geq 0}^{D \times T}$ collect these mixtures row-wise. Let $\mathcal{F}_\tau = \{1, \dots, V\}$ denote the set of validated interpretable SAE features used for topic description. Each feature $f \in \mathcal{F}_\tau$ has a natural-language label ℓ_f and interpretability score $\text{IS}(f)$. Our goal is to estimate a topic–feature matrix $B \in \mathbb{R}_{\geq 0}^{T \times V}$ whose row $\beta_t \in \Delta^V$ ranks validated interpretable features for topic t.

Stage 1: estimating document–topic mixtures. For mixture estimation, MonoTM uses the full mixture-SAE representation rather than only validated features. For a top-K SAE with expansion factor N, each document embedding yields a sparse nonnegative activation vector $\tilde{h}_d^{\text{mix}} \in \mathbb{R}_{\geq 0}^{mN}$ with at most K nonzero entries. We construct a document–feature matrix $X^{\text{mix}} \in \mathbb{R}_{\geq 0}^{D \times mN}$ by setting

$$X_{d,j}^{\text{mix}} = \tilde{h}_{d,j}^{\text{mix}}, \tag{13}$$

and fit LDA with T topics on X^{mix} to obtain Θ. This stage uses all active SAE features because Section 5.2.2 shows that strict interpretability filtering generally hurts mixture estimation.

Stage 2: constructing the interpretable feature matrix. For topic description, MonoTM restricts to interpretation-SAE features that (i) activate in at least 30 documents, (ii) have a nonempty natural-language label, and (iii) satisfy $\text{IS}(f) \geq \tau$, with $\tau = 0.8$ in our experiments. The interpretable document–feature matrix $C \in \mathbb{R}_{\geq 0}^{D \times V}$ is defined by

$$c_{d,f} = \max(\tilde{h}_{d,f}^{\text{int}}, 0) \cdot \text{IS}(f), \qquad f \in \mathcal{F}_\tau. \tag{14}$$

This weighting makes the final topic descriptors emphasize features with stronger validation evidence.

Stage 3: estimating topic–feature distributions. Given Θ and C, MonoTM estimates B by maximizing the fixed-Θ mixture log-likelihood

$$\mathcal{L}(B) = \sum_{d=1}^{D} \sum_{f=1}^{V} c_{d,f} \log \Big(\sum_{t=1}^{T} \theta_{d,t} \beta_{t,f} \Big), \tag{15}$$

subject to $\beta_t \in \Delta^V$ for all t. After estimating B, topic t is represented by its ranked feature distribution β_t; the reported descriptor list consists of the top-ranked labels ℓ_f under $\beta_{t,f}$, with associated probabilities.

K.1 Dirichlet-smoothed EM

We optimize Eq. (15) with an EM algorithm that treats latent topic assignments as missing data. Let

$$q_{d,f,t} = p(z = t \mid d, f) = \frac{\theta_{d,t} \beta_{t,f}}{\sum_{t'} \theta_{d,t'} \beta_{t',f}}. \tag{16}$$

The E-step computes expected topic–feature counts:

$$N_{t,f} = \sum_{d=1}^{D} c_{d,f}\, q_{d,f,t}. \tag{17}$$

The M-step updates B with symmetric Dirichlet smoothing $\eta > 0$:

$$\beta_{t,f} = \frac{N_{t,f} + \eta}{\sum_{f'} (N_{t,f'} + \eta)}. \tag{18}$$

We iterate until relative improvement in Eq. (15) falls below a tolerance. To avoid zero-probability features and stabilize topics when C is sparse, we use the data-dependent heuristic

$$\eta = \frac{1}{T} + \frac{\bar{L}}{V}, \tag{19}$$

where $\bar{L}$ is the average number of nonzero validated features per document.

K.2 Using Separate Mixture and Interpretation SAEs

MonoTM does not require a single SAE to simultaneously optimize mixture estimation and interpretability. Even when a single SAE is used end-to-end, we do not interpret topics using the topic–feature distribution implicitly produced during Θ inference. Instead, topic interpretation is based on a separately estimated topic–feature distribution over validated interpretable features. When the SAE configuration that yields the most useful validated feature set is suboptimal for estimating Θ, a user can train a second SAE optimized for mixture quality while using an interpretability-optimized SAE to construct C.

L Excluded-feature audit details

This appendix gives the full procedure for the excluded-feature audit in Section 5.3.2. The audit asks whether features excluded from the final MonoTM descriptor vocabulary nevertheless contain topic-defining semantic information that is absent from the validated descriptor set.

Configurations. For all datasets, the document–topic mixture matrix Θ is the same one used in the main RQ3 analysis. It is estimated from the mixture-quality SAE with expansion factor $N = 0.5$ and top-$K = 256$, using seed 123. The number of topics is fixed to the number of gold categories: $T = 20$ for 20 Newsgroups, $T = 7$ for Web of Science, and $T = 47$ for Reuters. The interpretation-side SAE used for the audit has $N = 2$ and $K = 16$.

All-active interpretation feature matrix. Let $\mathcal{F}_{\text{act}}$ denote the set of interpretation-SAE features that activate in at least 30 documents, and let ℓ_f be the generated label for feature f. The validated descriptor set used by MonoTM is

$$\mathcal{F}_{\text{val}} = \{f \in \mathcal{F}_{\text{act}} : \text{IS}(f) \geq .8 \text{ and } \ell_f \neq \emptyset\}.$$

For the audit, we instead construct an all-active feature matrix $A \in \mathbb{R}_{\geq 0}^{D \times |\mathcal{F}_{\text{act}}|}$ with entries

$$a_{d,f} = \max(\tilde{h}_{d,f}, 0),$$

where $\tilde{h}_{d,f}$ is the gain-adjusted top-K activation of the interpretation SAE. Unlike the final descriptor estimation step, the audit does not weight feature activations by $\text{IS}(f)$. This avoids building the conclusion into the measurement: interpretability-score weighting would mechanically reduce the mass of lower-validation features. Document rows are aligned between Θ, the interpretation-SAE activations, and raw documents using saved row IDs.

All-feature audit distribution. Using fixed Θ and the raw all-active matrix A, we estimate an auxiliary topic–feature distribution B^{all} by maximizing

$$\mathcal{L}(B^{\text{all}}) = \sum_{d=1}^{D} \sum_{f \in \mathcal{F}_{\text{act}}} a_{d,f} \log \left(\sum_{t=1}^{T} \theta_{d,t} \beta_{t,f}^{\text{all}} \right),$$

subject to $\beta_t^{\text{all}} \in \Delta^{|\mathcal{F}_{\text{act}}|}$ for every topic t. We use the same Dirichlet-smoothed EM routine as in the fixed-Θ MonoTM descriptor step. This distribution is used only for auditing; the final MonoTM descriptors are still estimated over the validated feature vocabulary.

High-risk excluded set. For each topic t, all active features are ranked by $\beta_{t,f}^{\text{all}}$. We define

$$\text{rank}_t^{\text{all}}(f) = 1 + \left| \{f' : \beta_{t,f'}^{\text{all}} > \beta_{t,f}^{\text{all}}\} \right|.$$

For each feature f with $\text{IS}(f) < .8$, let $t_f^\star$ be the topic where it has its best rank. The high-risk excluded set is

$$\mathcal{R}_{20} = \{(t_f^\star, f) : \text{IS}(f) < .8,\ \text{rank}_{t_f^\star}^{\text{all}}(f) \leq 20\}.$$

This set contains lower-validation features that are sufficiently associated with at least one topic to be plausible missing descriptors.

Coverage test. For each $(t, f) \in \mathcal{R}_{20}$, we compare f with validated same-topic descriptors. The candidate pool is

$$\mathcal{V}_{t,M} = \{g \in \mathcal{F}_{\text{val}} : \text{rank}_t^{\text{all}}(g) \leq M\}.$$

We report $M = 50$ in the main text and include $M = 20$ and $M = 100$ as sensitivity settings. The $M = 20$ setting asks whether the feature is covered by the most compact validated descriptor region; $M = 50$ asks whether it is covered by the broader high-ranked validated descriptor distribution; $M = 100$ checks sensitivity to a wider candidate set.

Let a_f be the document-level activation vector for feature f. We compute

$$\cos_{\text{act}}(f, g) = \frac{a_f^\top a_g}{\|a_f\|_2 \|a_g\|_2}.$$

The nearest validated same-topic descriptor is

$$g^\star(f, t) = \arg \max_{g \in \mathcal{V}_{t,M}} \cos_{\text{act}}(f, g).$$

For comparison, we sample random validated descriptors from the same candidate pool and compute the random-baseline activation cosine. Table 9 reports the aggregate results.

Heuristic label-type flags. We also assign a coarse heuristic label type to each excluded feature based only on keywords in its generated feature label. These flags are not manual annotations and are not used as a validation metric. They are included to characterize broad patterns, especially the difference between Reuters and the other datasets.

Labels are flagged as *format/register/numeric* if they include terms related to document form, boilerplate, reporting style, technical artifacts, numerical data, or financial-news announcements,

Dataset	Candidate pool	n	Med. nearest	Med. random	Med. Δ	$\%\cos_{\text{act}} \geq .2$
20NG	top-20	129	.239	.072	.154	60.5
20NG	top-50	129	.295	.055	.218	72.9
20NG	top-100	129	.312	.044	.262	78.3
WoS	top-20	26	.360	.104	.235	76.9
WoS	top-50	26	.419	.079	.295	88.5
WoS	top-100	26	.442	.066	.349	96.2
Reuters	top-20	334	.170	.059	.105	41.6
Reuters	top-50	334	.170	.043	.130	45.2
Reuters	top-100	334	.188	.029	.153	47.9

Table 9: Coverage audit details for lower-validation high-association features. The high-risk excluded set is fixed as features with $\text{IS}(f) < .8$ and best all-feature topic rank at most 20. Candidate pool size controls how many validated same-topic descriptors are allowed to provide coverage. Median random cosine is computed from random validated descriptors sampled from the same topic-specific candidate pool.

including keywords such as: *format, boilerplate, template, placeholder, metadata, header, posting, email, message, announcement, reporting, reports, quarterly, qtly, dividend, earnings, record date, subscription, unsubscribe, FAQ, manual, instructions, technical support, code, script, program, software package, release, version, archive, file, conversion, table, tabular, numerical, numeric, quantitative, percentage, percent, price, market data, Reuter, newswire*. Labels are flagged as *broad discourse* if they include terms related to debate, critique, controversy, policy, institutional or political discussion, economic conflict, or broad framing, including keywords such as: *discussion, debate, critique, controversy, tensions, policy, institutional, economic, financial, trade, political, rhetoric, argument, concerns, issues, context, general, broad*. Remaining labels are assigned the default *semantic* flag. The rules are applied in this order: format/register/numeric first, broad discourse second, and semantic as the default.

Examples from the excluded-feature audit. Table 10 gives deterministically selected examples from the label-type audit. To avoid cherry-picking, examples are selected by a fixed rule: within each dataset and heuristic label-type category, we sort features by best all-feature topic rank, break ties by larger $\beta^{\text{all}}_{t,f}$, and show the first available feature. The table is intended to make the heuristic categories transparent, not to serve as a separate evaluation.

Concrete high-association cases illustrate the two main patterns. In 20 Newsgroups, an excluded feature labeled "Israeli policy critique with Nazi analogies" is ranked first under B^{all} for the topic aligned to `talk.politics.mideast`. Its nearest validated same-topic descriptor is "Israeli-Palestinian conflict, criticism, and political dispute," with activation cosine .650. In Web of Science, the excluded feature "Rainwater harvesting for water resource management" is ranked first for the Civil topic, but its nearest validated same-topic descriptor is "Rainwater harvesting for sustainable water management," with activation cosine .732. These cases look less like independent missing topics and more like lower-validation variants of semantic regions already covered by validated descriptors. By contrast, Reuters features such as "Financial dividend and earnings report announcements" or "Dividend and earnings financial report language" are systematic and topic-associated, but they primarily describe reporting format or financial-news register rather than substantive topical content.

Dataset	Type	IS(f)	Excluded label	Nearest validated label	$\cos_{\text{act}}$
20NG	semantic	.410	Patient narratives of challenging medical experiences	Alternative medicine debate and medical authority skepticism	.337
20NG	broad	.783	Detailed, analytical discussion of professional hockey	Toronto Maple Leafs playoff game discussion & fandom	.301
20NG	format/reg.	.727	Sports news and scores reporting via Usenet	Ironic, irreverent, and self-aware internet discourse	.176
WoS	semantic	.333	Rainwater harvesting for water resource management	Rainwater harvesting for sustainable water management	.732
WoS	broad	.750	Distributed computing system architecture and design discuss	Intelligent interconnected devices and adaptive environments	.443
WoS	format/reg.	.750	Novel materials for photocatalytic energy conversion	Photocatalysis, solar energy, material science research	.692
Reuters	semantic	.667	Bank ownership stake acquisitions and control	Corporate deals, mergers, and joint ventures reporting	.341
Reuters	broad	.769	International trade disputes and policy tensions	International trade disputes and government intervention	.315
Reuters	format/reg.	.256	US-USSR grain trade reporting, quantitative data	U.S. farm policy and commodity support programs	.312

Table 10: Deterministically selected examples from the excluded-feature label-type audit. Examples are selected by sorting, within each dataset and heuristic type, by best all-feature topic rank and then by larger audit-distribution topic mass.

M Descriptor Baseline and Topic-Naming Details

For the controlled descriptor comparison in Section 5.3.3, all methods describe the same fixed document–topic mixtures Θ. We generate descriptors for each inferred topic using four mechanisms: θ-weighted TF-IDF terms, BERTopic-style c-TF-IDF terms, free-form LLM-generated descriptors, and MonoTM validated SAE-feature descriptors.

Word-based descriptor preprocessing. Both word-based baselines are implemented with `scikit-learn` vectorizers. We do not apply stemming or lemmatization. The vectorizers lowercase all text, remove English stopwords, and use unigrams only. We use the token pattern

```
(?u)\b[a-zA-Z][a-zA-Z0-9_\-]{2,}\b,
```

which keeps tokens that begin with an ASCII letter and have length at least three, while allowing later alphanumeric, underscore, and hyphen characters. This removes many short tokens, numbers-only tokens, and punctuation artifacts. We use a maximum vocabulary size of 50,000 terms.

θ-weighted TF-IDF. For the θ-weighted TF-IDF baseline, we fit a `TfidfVectorizer` on the full aligned corpus with `lowercase=True`, `stop_words=english`, `ngram_range=(1,1)`, `min_df=2`, `max_df=0.95`, `max_features=50000`, `norm=None`, and `sublinear_tf=True`. For each topic t, let D_t^{50} be the 50 documents with the largest $\theta_{d,t}$. We normalize topic weights within this top-document set,

$$\bar{\theta}_{d,t} = \frac{\theta_{d,t}}{\sum_{d' \in D_t^{50}} \theta_{d',t}},$$

and score each term w by

$$s_t^{\text{tfidf}}(w) = \sum_{d \in D_t^{50}} \bar{\theta}_{d,t} \, \text{tfidf}_{d,w} \, .$$

We report the 30 highest-scoring terms for each topic.

c-TF-IDF. For the c-TF-IDF baseline, we form one class document for each topic by concatenating the same 50 top-θ documents. We then fit a `CountVectorizer` with `lowercase=True`, `stop_words=english`, `ngram_range=(1,1)`, `min_df=1`, `max_features=50000`, and the same token pattern as above. Let $n_{t,w}$ be the count of term w in topic class document t, and let $L_t = \sum_w n_{t,w}$ be the class-document length. We compute

$$\text{tf}_{t,w} = \frac{n_{t,w}}{L_t}.$$

Let $\text{df}_{\text{class}}(w)$ be the number of topic class documents in which w appears, and let $\bar{L}$ be the average class-document length across topics. We use the class-based inverse-document-frequency term

$$\text{idf}_w = \log\left(1 + \frac{\bar{L}}{\max(\text{df}_{\text{class}}(w), 1)}\right),$$

and rank terms by

$$s_t^{\mathrm{ctfidf}}(w) = \mathrm{tf}_{t,w}\,\mathrm{idf}_w\,.$$

We report the 30 highest-scoring terms for each topic.

Free-form LLM descriptor baseline. The LLM summary baseline is intended to be a strong post-hoc descriptor baseline. For each topic, the prompt receives 50 topic-associated documents. The 25 highest-θ documents are always included, and the remaining 25 documents are sampled without replacement from the top 200 documents for that topic using a fixed random seed. Each document is first whitespace-normalized and then truncated to at most 1000 model tokens; if tokenization support is unavailable, the implementation falls back to whitespace-token truncation. We do not apply an additional character cap.

We use OpenAI `gpt-5-2025-08-07` through the Responses API. For LLM-generated descriptors, we use `reasoning_effort=low`, `verbosity=low`, JSON-object output mode, and a maximum output budget of 3000 tokens. We leave temperature unspecified. If the response cannot be parsed into exactly 10 descriptors, we retry up to three times. The prompt asks the model to produce exactly 10 concise, non-redundant descriptors grounded only in the provided excerpts.

Free-form LLM descriptor prompt

You are given excerpts from documents that are highly associated with the same latent topic.
Task: Produce exactly 10 concise descriptors for this topic.
Rules: Each descriptor must be 4 to 10 words. Descriptors should be specific, non-redundant, and grounded only in the excerpts. Do not output a single topic name; output a ranked list of descriptors. Return JSON only, with this exact schema: `{"descriptors": ["...", "..."]}`.
Topic id: `<topic>`
Documents: `<document excerpts>`

MonoTM descriptors. For MonoTM, we use the topic–feature distribution B estimated by the fixed-Θ descriptor model described in Appendix K. We retain validated SAE features with $\mathrm{IS}(f) \geq 0.8$ and report the top 10 feature labels under $\beta_{t,f}$ for each topic. Unlike the free-form LLM descriptor baseline, MonoTM does not generate each topic's descriptor list freely from top documents. Instead, the descriptor list is selected from a fixed validated feature vocabulary grounded in document-embedding activations.

Topic naming for compact qualitative comparison. For compact comparison in Tables 12, 13, and 14, we generate high-level topic names from each descriptor list. The naming model receives only the descriptor list and the descriptor method; it does not receive the gold label or the candidate-label list. For word-based methods, the naming prompt receives the 30 ranked terms. For the free-form LLM descriptor baseline and MonoTM, it receives the 10 semantic descriptors.

We use the same OpenAI model, `gpt-5-2025-08-07`, through the Responses API with `reasoning_effort=low`, `verbosity=low`, JSON-object output mode, and a maximum output budget of 300 tokens. We again leave temperature unspecified and retry up to three times if the JSON response cannot be parsed.

Topic naming prompt

I have a ranked list of `<top words/phrases or semantic descriptors>` derived from a single latent topic.
Descriptor method: `<method>`
Ranked descriptors: `<descriptor list>`
Task: Identify the high-level topic that these descriptors likely came from.
Rules: Output a concise topic name in 1 to 3 words. Do not include quotes, explanations, bullets, or extra text. Return JSON only with this exact schema: `{"name": "..."}`.

N High-Level Topic Labels Across SAE Configurations

To provide an intuitive qualitative check of MonoTM's output, we take each topic's top-10 interpretable feature labels (ranked by $\beta_{t,f}$) and ask `gpt-4.1-2025-04-14` (temperature 0) to produce a 1–3 word topic name using the prompt in Appendix M. Table 11 shows these names for three SAE configurations ($(N{=}1, K{=}8)$, $(N{=}3, K{=}16)$, $(N{=}5, K{=}32)$). For an informative comparison, we align inferred topics to gold categories using the process described in Section 5.2. In many cases, the generated topic labels closely paraphrase the matched gold labels.

Gold Label	N=1, K=8	N=3, K=16	N=5, K=32
20 Newsgroups			
rec.sport.hockey	Online Hockey Discussions	NHL Hockey Fans	Professional Hockey Discourse
soc.religion.christian	Christian Theology Debates	Religious Controversies	Religion and Morality
rec.motorcycles	Motorcycle Riding	Motorcycle Safety	Motorcycle Culture
rec.sport.baseball	Baseball Online Communities	Baseball Discussion	Baseball Analysis
sci.crypt	Cypherpunk Movement	Cryptography and Surveillance	Clipper Chip Controversy
Web of Science			
Medical	Human Disease Research	Human Health Research	Human Health Disorders
Psychology	Child and Adolescent Psychology	Psychological Research	Social Psychology
CS	Computer Science	Computer Science	Computer Science
biochemistry	Biomedical Science	Biomedical Research	Biomedical research
ECE	Control Engineering	Electrical Engineering	Electrical Engineering
Reuters			
earn	Quarterly Financial Reporting	Financial Reporting	Financial Performance Comparison
acq	Mergers and Acquisitions	Financial News Events	Mergers and Acquisitions
money-fx	Central Bank Operations	Bank of England Interventions	Central Bank Interventions
grain	Agricultural Trade Markets	Agricultural Trade Reporting	Agricultural Trade
crude	Global Oil Markets	Global Oil Markets	OPEC Oil Markets

Table 11: Topic labels generated from MonoTM outputs across three SAE configurations. The gold labels are shown only for orientation and are not provided during topic naming.

O Additional Descriptor Comparison Results

Tables 12–14 report additional topic-name tables for the controlled descriptor comparison in Section 5.3.3. As in the main text, all methods describe the same inferred topics under fixed document–topic mixtures Θ, and reference labels are shown only for orientation.

Reference	MonoTM	θ-TF-IDF	c-TF-IDF	LLM-generated
alt.atheism	Religious debate	Objective morality	Objective morality	Meta-ethics
comp.graphics	Computer Graphics	Image format conversion	Image file formats	Computer Graphics
comp.os.ms-windows.misc	PC hardware troubleshooting	Hard drive setup	PC hard drives	DOS disk configuration
comp.sys.ibm.pc.hardware	Technical troubleshooting	Lead-acid batteries	Lead-acid batteries	Lead-acid battery storage
comp.sys.mac.hardware	PC hardware troubleshooting	Macintosh hardware	Macintosh hardware	Macintosh video hardware
comp.windows.x	X11 troubleshooting	X11 programming	X11 Motif	X11 programming
misc.forsale	Tech classifieds	Comics for sale	Marvel Comics	Classifieds
rec.autos	Automotive discussion	Cars	Cars	Performance cars
rec.motorcycles	Motorcycle Culture	Motorcycles	Motorcycles	Motorcycle riding techniques
rec.sport.baseball	Baseball Analytics	Baseball	Baseball	Baseball sabermetrics
rec.sport.hockey	Ice hockey	NHL Playoffs	NHL hockey	NHL Playoffs
sci.crypt	Clipper Chip Debate	Clipper Chip	Clipper Chip	Clipper Chip
sci.electronics	Computer hardware	Electronic circuits	Electronic circuits	Clickless audio switching
sci.med	Medical skepticism	Candida yeast syndrome	Candida overgrowth	Candida overgrowth
sci.space	Space commercialization	Lunar exploration	Spaceflight	Lunar habitation prize
soc.religion.christian	Theological debates	Christianity	Christian theology	Biblical hermeneutics
talk.politics.guns	Waco siege	Waco siege	Waco siege	Waco Siege
talk.politics.mideast	Ethno-national conflicts	Armenian Genocide	Armenian Genocide	Armenian genocide controversy
talk.politics.misc	American culture wars	Gun rights	Gun rights	Gay rights debate
talk.religion.misc	Usenet discussions	Email	Usenet newsgroups	Mailing list administration

Table 12: High-level topic names generated from descriptor outputs on 20 Newsgroups. All methods describe the same inferred topics under fixed document–topic mixtures Θ. Reference labels are shown only for orientation and are not provided to the topic-naming model.

Reference	MonoTM	θ-TF-IDF	c-TF-IDF	LLM-generated
biochemistry	Biomedical research	Cancer cell signaling	Molecular oncology	Signal transduction
Civil	Civil and Environmental Engineering	Rainwater harvesting	Rainwater Harvesting	Rainwater harvesting
CS	Computer Science Research	Big Data Computing	Big Data	Cloud and Distributed Systems
ECE	Control Systems Engineering	Power converter control	Power Converter Control	Digital Inverter Control
MAE	Computational mechanics	Composite Materials	Composite materials	Mechanical behavior of materials
Medical	Clinical epidemiology	Male hypogonadism	Male hypogonadism	Male hypogonadism
Psychology	Psychology research	Parenting and child behavior	Child Development	Prosocial Behavior

Table 13: High-level topic names generated from descriptor outputs on Web of Science. All methods describe the same inferred topics under fixed document–topic mixtures Θ. Reference labels are shown only for orientation and are not provided to the topic-naming model.

Reference	MonoTM	θ-TF-IDF	c-TF-IDF	LLM-generated
earn	Earnings reports	Earnings Reports	Earnings reports	Corporate Earnings
acq	Mergers and Acquisitions	Bank mergers	Bank Mergers	Bank Mergers
money-fx	Money market interventions	Money market operations	Money market	Open market operations
grain	Agricultural Trade	Export Enhancement Program	Grain export subsidies	Export Enhancement Program
crude	OPEC oil market	Oil production	OPEC oil production	Ecuador oil crisis
trade	US-Japan Trade Disputes	US-Japan trade dispute	US-Japan trade dispute	U.S.-Japan trade dispute
interest	Interest Rate Changes	Prime rate	Interest rates	Prime rate changes
ship	Persian Gulf conflict	Persian Gulf Tanker War	Gulf Tanker War	Operation Nimble Archer
wheat	Corporate financial distress	Earnings reports	Earnings reports	Nonrecurring items
corn	Mergers and Acquisitions	Retail mergers and acquisitions	Business news	Consumer retail M&A
oilseed	Financial reporting	Corporate earnings	Corporate earnings	Corporate earnings
sugar	Agricultural trade	Sugar crop production	Crop production	Global sugar production
dlr	International monetary policy	Louvre Accord	Louvre Accord	G7 currency coordination
gnp	Macroeconomic Indicators	Economic Growth	Macroeconomic Indicators	OECD economic outlook
coffee	International Coffee Trade	International coffee trade	International Coffee Agreement	International Cocoa Agreement
veg-oil	Trade and M&A	Mergers and acquisitions	Mergers and acquisitions	HBJ takeover battle
money-supply	Federal Reserve operations	Open market operations	Open market operations	Open market operations
gold	Mining Industry Finance	Gold mining	Gold mining	Gold mining
nat-gas	Airline Mergers	Airline mergers	Airline mergers	TWA USAir takeover
livestock	US Agricultural Policy	Livestock futures	Livestock futures	Meatpacking labor disputes

Table 14: High-level topic names generated from descriptor outputs on the 20 most frequent Reuters labels. Because Reuters is multi-label and imbalanced, reference labels should be read as orientation labels rather than exhaustive topic descriptions.

P Representative Descriptor Lists

Table 15 provides representative descriptor lists before topic-name compression for the controlled descriptor comparison in Section 5.3.3. All methods describe the same fixed document–topic mixtures Θ. The examples illustrate the different roles of the descriptor mechanisms: word-based methods expose lexical anchors, free-form LLM-generated descriptors often summarize salient document clusters, and MonoTM descriptors are selected from a fixed vocabulary of validated SAE features ranked by a model-estimated topic–feature distribution.

Reference	Method	Top descriptors
sci.crypt	MonoTM	Distrust of government and defense of liberty; government skepticism, privacy concerns, argumentative discourse; technical discourse on cryptographic algorithms and implementation; secure voice tech, compression, early internet security; government surveillance and cryptographic privacy concerns
sci.crypt	θ-TF-IDF	clipper; encryption; keys; escrow; chip; government; technology; secure; nsa; key
sci.crypt	c-TF-IDF	encryption; clipper; chip; government; technology; key; keys; escrow; nsa; secure
sci.crypt	LLM-generated	Government key-escrowed Clipper Chip proposal; secret Skipjack algorithm and family key; law enforcement access via escrowed unit keys; LEAF transmits session key encrypted for authorities; critics warn of mandated, tappable encryption standard
talk.politics.misc	MonoTM	Distrust of government and defense of liberty; contentious debate surrounding homosexuality and morality; moral debates about sexuality and statistics; gun control, self-defense arguments, statistical debate; contentious social debate, group identity, and conflict
talk.politics.misc	θ-TF-IDF	cramer; optilink; clayton; homosexuals; people; gay; gun; militia; rights; sexual
talk.politics.misc	c-TF-IDF	people; cramer; edu; optilink; militia; homosexuals; gay; gun; sexual; rights
talk.politics.misc	LLM-generated	Debating Kinsey versus Guttmacher homosexuality prevalence statistics; claims homosexuals exaggerate numbers to sway politicians; arguments over gay rights anti-discrimination employment laws; libertarian stance: private hiring without government interference; accusations linking gays with child molestation, NAMBLA
ECE	MonoTM	Engineered control systems and PID control design; electrical power system analysis and control; high-performance electric motor design and analysis; analog circuit design and performance metrics; engineered system optimization and energy efficiency analysis
ECE	θ-TF-IDF	converter; control; digital; voltage; proposed; power; current; controller; circuit; system
ECE	c-TF-IDF	control; proposed; voltage; converter; power; current; controller; circuit; digital; system
ECE	LLM-generated	FPGA-based digital control for power converters; digital PLLs for single-phase grid synchronization; discrete-time modulation introducing transport delays; finite control set MPC implemented fully on FPGA; deadbeat predictive current control in inverters
Civil	MonoTM	Applied hydrology, water resource engineering, rainwater harvesting; engineered system optimization and energy efficiency analysis; sustainable building and environmental responsibility concepts; construction project management, analytical methodology, performance evaluation; water pollution assessment and scientific reporting
Civil	θ-TF-IDF	water; rainwater; rwh; harvesting; demand; tank; system; saving; efficiency; supply
Civil	c-TF-IDF	water; rainwater; rwh; harvesting; use; demand; tank; saving; efficiency; system
Civil	LLM-generated	Rainwater harvesting supplying toilets, laundry, irrigation; economic feasibility via payback and NPV analyses; optimal storage tank sizing using daily simulations; water saving efficiency and reliability metrics; integrating rainwater with greywater reuse systems
crude	MonoTM	Oil price stability and OPEC market control; OPEC oil production, pricing, and market regulation; OPEC oil price and production control; OPEC oil production, pricing, and regulation reports; oil market data and energy economics
crude	θ-TF-IDF	bpd; oil; ecuador; opec; barrels; crude; prices; pipeline; production; exports
crude	c-TF-IDF	said; oil; bpd; opec; ecuador; crude; barrels; pipeline; production; prices
crude	LLM-generated	Earthquake halts Ecuador crude exports, declares force majeure; Lago Agrio–Balao pipeline severely damaged, costly repairs; temporary pipeline link to Colombia's Tumaco port; Venezuela lends Ecuador crude for exports, domestic needs; Ecuador plans over-quota OPEC production to repay loans
ship	MonoTM	Government economic intervention and trade policy; US-Iran military conflict and retaliation narratives; international military conflict and geopolitical tension; Persian Gulf naval conflict and escalation threat; Persian Gulf oil security, U.S. military response
ship	θ-TF-IDF	iran; gulf; iranian; attack; kuwaiti; oil; reflagged; ship; tanker; platform
ship	c-TF-IDF	iran; said; gulf; iranian; oil; attack; kuwaiti; ship; platform; tanker
ship	LLM-generated	U.S. destroyers shell Iranian Rostam oil platform; retaliation for Silkworm strike on Sea Isle City; Navy raids second platform, destroys radar communications; Iran vows crushing retaliation for platform attacks; escorting reflagged Kuwaiti tankers through Gulf

Table 15: Longer representative descriptor lists before topic-naming. Word-based descriptors often expose terms or named entities; LLM-generated descriptors often describe salient document clusters; MonoTM descriptors are selected from validated SAE features and ranked by a topic–feature distribution.

Q Affordance Demonstration: Topic Relations in a Shared Semantic-Feature Space

This appendix demonstrates a downstream affordance of MonoTM's shared semantic-feature representation. The goal is not to introduce a new quantitative benchmark, but to illustrate how MonoTM can support analyses beyond compact topic description. Because MonoTM represents both documents and topics in a shared vocabulary of validated SAE features, users can inspect which features are shared across documents, which features distinguish neighboring topics, and which semantic attributes cut across topic boundaries. We demonstrate this affordance by analyzing how topics are related in the Web of Science corpus.

Construction details. We perform the analysis on Web of Science under the same fixed document–topic mixtures Θ used in the controlled descriptor comparison. For MonoTM, we use the estimated topic–feature distribution B and retain the top 200 validated features per topic. We keep the raw $\beta_{t,f}$ masses rather than renormalizing them. For the lexical analogue, we use the top 50 terms per topic from the θ-weighted TF-IDF topic–term representation, with term scores normalized to sum to one within each topic.

For two topics i and j with MonoTM topic–feature distributions β_i and β_j, we define their shared feature mass as

$$O(i,j) = \sum_f \min(\beta_{i,f}, \beta_{j,f}).$$

More generally, for either representation, we compute topic-pair overlap as

$$O(i,j) = \sum_x \min(v_{i,x}, v_{j,x}),$$

where x indexes validated SAE features for MonoTM and lexical terms for the TF-IDF analogue. To explain each edge, we rank shared items by $\min(v_{i,x}, v_{j,x})$ and report the top shared features or terms. The visualized networks display the eight strongest topic-pair overlaps, and the tables below report representative shared items from those edges.

To visualize the resulting topic relations, Figure 9 shows the strongest topic-pair overlaps as networks for MonoTM and the θ-weighted TF-IDF analogue. Figure 10 provides the corresponding full pairwise overlap matrices, showing whether the strongest network edges reflect isolated pairwise relations or broader overlap structure across topics.

Qualitative comparison. Figures 9 and 10 show that the two representations yield qualitatively different topic-relation structures and edge explanations. Table 16 reports representative shared items for several high-overlap topic pairs. MonoTM links topics through semantic features such as clinical disease investigation, severe mental illness and treatment, efficient algorithm implementation and hardware optimization, applied engineering education, and engineered system optimization and energy efficiency. By contrast, the θ-weighted TF-IDF analogue often explains topic links using lexical items such as *model*, *systems*, *method*, *study*, *results*, *paper*, and *based*. Thus, a word-based representation can indicate that two topics overlap lexically, but MonoTM can make the semantic basis of the relation more explicit.

Overall, this demonstration illustrates a broader affordance of MonoTM: the same validated feature vocabulary can support topic description, topic comparison, and inspection of semantic attributes that cut across benchmark category boundaries.

Topic pair	MonoTM shared validated features	θ-TF-IDF shared terms
biochemistry–Medical	Pulmonary disease research and clinical investigation focus.; Cancer research, molecular mechanisms, and clinical outcomes.; Gastrointestinal distress, IBS/IBD, patient well-being	study
Medical–Psychology	Severe mental illness, clinical diagnosis, and treatment.; Anxiety, depression, and psychological distress research.; Child development, family well-being, healthcare support.	study; results; levels
CS–ECE	Efficient algorithm implementation and hardware optimization.; Applied technology for infrastructure and efficiency.; Mathematical modeling and algorithmic development in data science	paper; based; model; performance; proposed
MAE–ECE	Magnetic material design and analysis in engineering.; Precise sensor measurement and instrumentation details.; Detailed thermodynamic system analysis and optimization.	model; experimental; method; using; results
Civil–ECE	Engineered system optimization and energy efficiency analysis.; Engineered thermal systems and heat transfer analysis.; Applied technology for infrastructure and efficiency.	model; systems; performance; using
MAE–Civil	Engineered thermal systems and heat transfer analysis.; Applied engineering education and curriculum design.; Technical design, optimization, and performance analysis.	model; using

Table 16: Representative shared items explaining Web of Science topic-pair edges. MonoTM edge explanations are validated SAE features ranked by shared β mass, while the lexical analogue explains edges through shared θ-weighted TF-IDF terms.

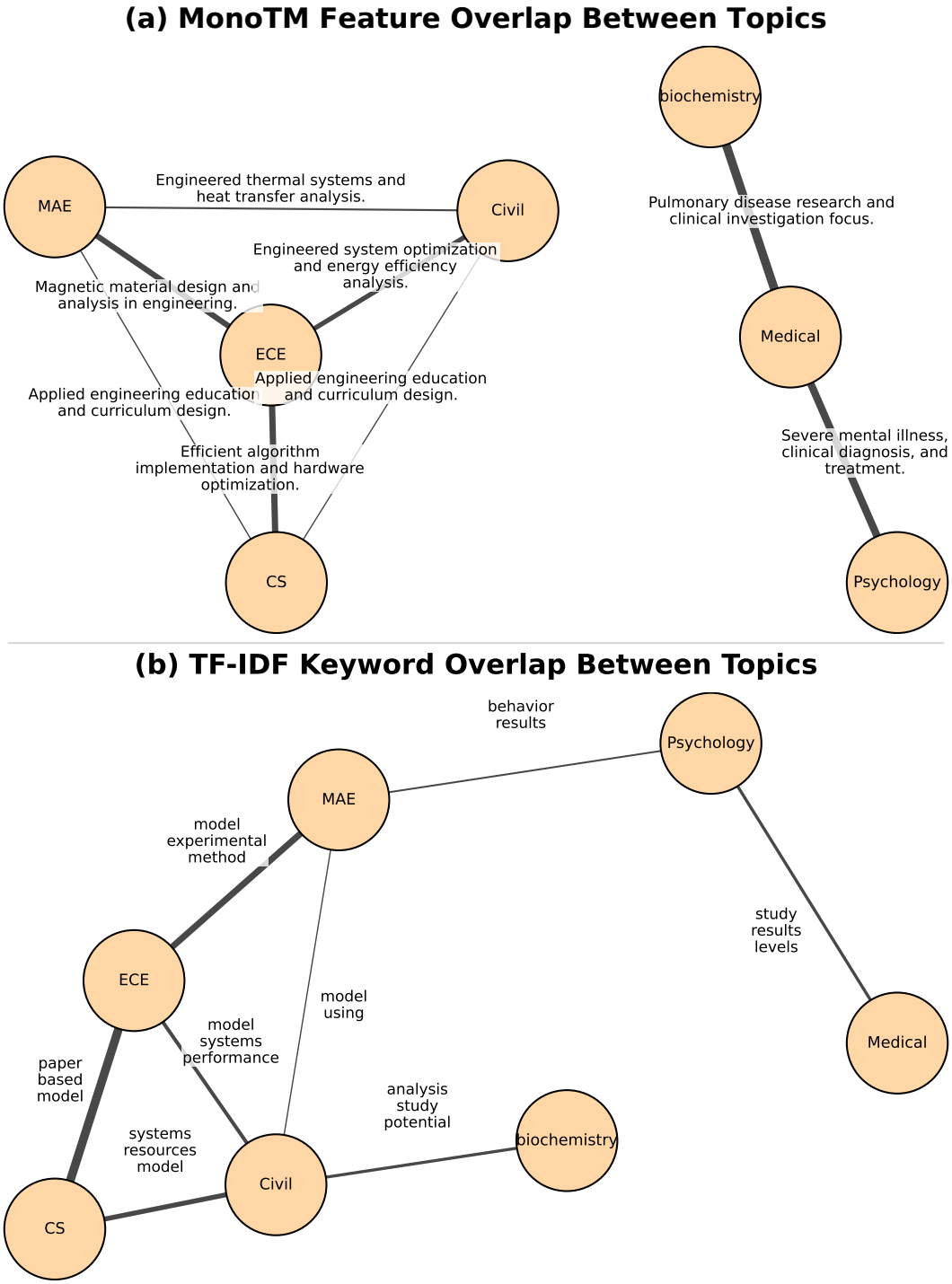


Figure 9: Web of Science topic relation networks under fixed document–topic mixtures. Edges show the strongest topic-pair overlaps, with edge width proportional to shared mass. (a) MonoTM computes overlap in the validated SAE-feature space, so edge labels identify the semantic features that explain each topic connection. (b) A θ-weighted TF-IDF analogue computes overlap in a lexical term space, so edge labels identify shared keywords. The comparison illustrates how MonoTM supports feature-level interpretation of topic relations beyond lexical overlap.

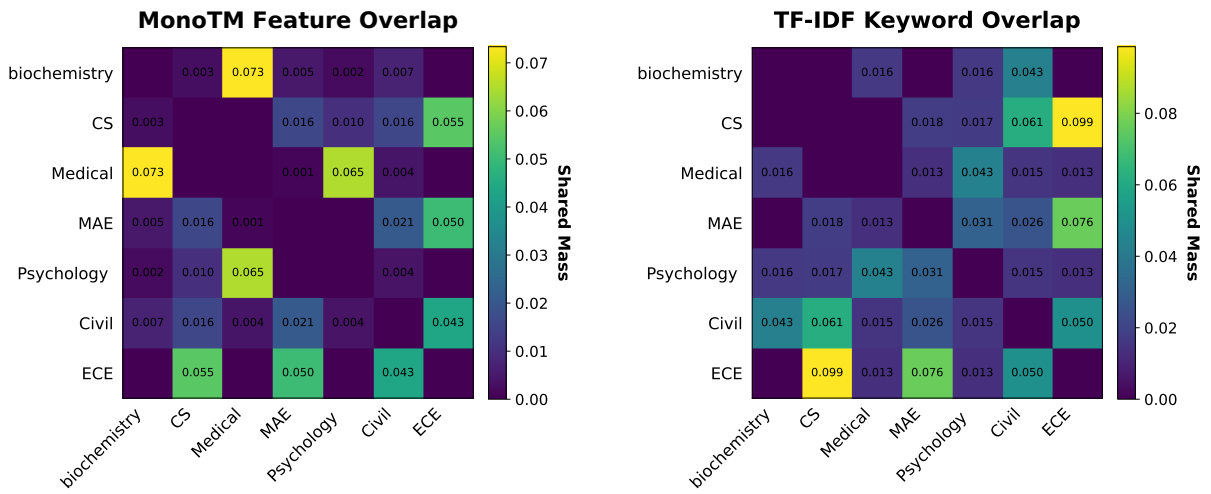


Figure 10: Topic-pair overlap matrices for Web of Science. Left: MonoTM shared-feature mass using the top 200 validated features per topic from B. Right: θ-weighted TF-IDF shared-term mass using the top 50 lexical terms per topic. Diagonal entries are set to zero.