Model Directions, Not Words: Mechanistic Topic Models Using Sparse Autoencoders

Carolina Zheng[*◇] **Nicolas Beltran-Velez**[*◇] **Sweta Karlekar**[*◇] **Claudia Shi**[◇]
Achille Nazaret[◇] **Asif Mallik**[‡] **Amir Feder**[◇†] **David M. Blei**[◇]
[◇]Columbia University, USA [†]Google Research, USA [‡]Independent
{cz2539, nb2838, ssk2275}@columbia.edu

Abstract

Traditional topic models are effective at uncovering latent themes in large text collections. However, due to their reliance on bag-of-words representations, they struggle to capture semantically abstract features. While some neural variants use richer representations, they are similarly constrained by expressing topics as word lists, which limits their ability to articulate complex topics. We introduce Mechanistic Topic Models (MTMs), a class of topic models that operate on interpretable features learned by sparse autoencoders (SAEs). By defining topics over this semantically rich space, MTMs can reveal deeper conceptual themes with expressive feature descriptions. Moreover, uniquely among topic models, MTMs enable controllable text generation using topic steering vectors. To properly evaluate MTM topics against word list approaches, we propose *topic judge*, an LLM-based pairwise comparison evaluation framework. Across eight datasets, MTMs match or exceed traditional and neural baselines on coherence metrics, are consistently preferred by topic judge, and enable effective LLM steering.[1]

1 Introduction

Topic models are a family of unsupervised algorithms that automatically discover thematic structure in document collections (Blei, 2012). Given a corpus of texts, they produce a predefined number of topics—each represented by a set of words that characterize the theme—along with a per-document breakdown indicating how much each topic contributes to that document's content.

Traditional methods, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), represent documents as bag-of-words counts and discover topics by modeling patterns of word co-occurrence. However, by operating on bag-of-words representations, they miss contextual and semantic nuances. Neural topic models (e.g., Bianchi et al., 2021a; Grootendorst, 2022; Wu et al., 2024) attempt to mitigate this limitation by leveraging pretrained embeddings to capture richer semantics. However, these models interpret topics as lists of words weighted by importance, which restricts their ability to articulate abstract concepts and nuanced semantic relationships. Even when probabilistic neural topic models incorporate pretrained embeddings, they still model the generation of word counts, implicitly retaining bag-of-words assumptions.

Independently, recent advances in *mechanistic interpretability* have shown that many high-level semantic concepts in large language models (LLMs) are encoded as linear directions within their internal activations (Mikolov et al., 2013; Elhage et al., 2022). Sparse autoencoders (SAEs) (Tamkin et al., 2024; Cunningham et al., 2024) are neural models that extract these interpretable features from LLM activations, each of which can be subsequently labeled with automatically generated textual descriptions (Bills et al., 2023; Paulo et al., 2025).

In this paper, we explore the use of these features for topic modeling. Unlike traditional bag-of-words representations, SAE features capture contextual and semantic concepts that extend beyond word co-occurrence patterns. Moreover, since these features can be labeled with textual descriptions, they enable discovering and describing topics at higher semantic abstraction levels.

We introduce Mechanistic Topic Models (MTMs), a family of topic models that adapt existing approaches to operate on SAE features rather than words. This adaptation enables MTMs to: (1) capture context and semantic nuance using pretrained LLM representations; (2) generate interpretable topic descriptions using SAE features that directly capture abstract concepts like style, tone,

[*] Equal contribution.
[1] github.com/blei-lab/mechanistic-topic-models

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2187–2212, 2026. https://doi.org/10.1162/TACL.a.797
Action Editor: Mark Johnson. Submission batch: 7/2025; Revision batch: 1/2026; Published 8/2026.

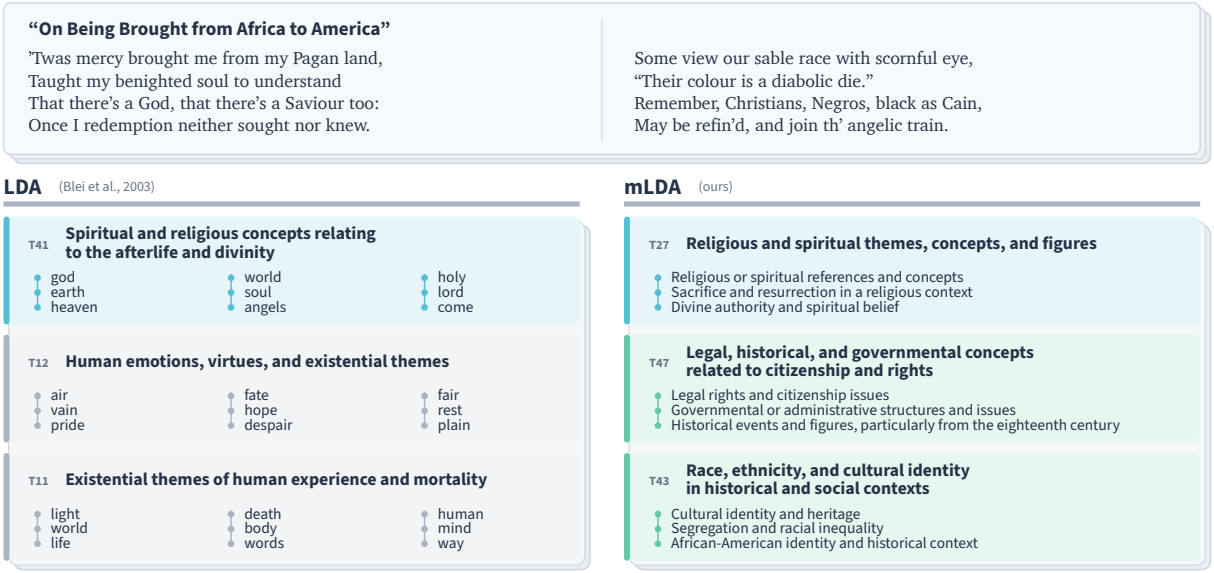

Figure 1: Comparison of LDA and mLDA topic outputs for Phillis Wheatley's "On Being Brought from Africa to America" from PoemSum. The top three topics from each model are shown with LLM-generated summaries. LDA topics are represented by their top words, while mLDA topics are represented by selected top SAE features—natural-language descriptions of learned directions in LLM activation space. Both models recover the poem's religious aspect (blue). LDA's remaining topics are relatively broad poetic themes (gray); mLDA instead identifies the poem's historical and racial context (green).

and discourse patterns; and (3) enable topic-based controlled generation through learned steering vectors. Figure 1 compares LDA and mLDA topic representations for a poem by Phillis Wheatley. Both models assign a religious topic; mLDA additionally assigns one topic on historical and civic concepts and another on race and cultural identity, while LDA's other topics are comparatively broad. This example illustrates how MTMs can represent abstract concepts that are difficult to capture with word lists alone.

We instantiate MTMs by adapting three standard topic models to SAEs: mechanistic LDA (mLDA) from Blei et al. (2003), mechanistic Embedded Topic Model (mETM) from Dieng et al. (2020), and mechanistic BERTopic (mBERTopic) from Grootendorst (2022).

We make three contributions. First, we introduce Mechanistic Topic Models (MTMs) alongside three variants and demonstrate that these models perform well on challenging corpora including abstract texts and short, context-limited documents. Second, we develop *topic judge*, a new evaluation method that uses LLM-based pairwise comparisons to assess how well topics describe documents, enabling fair cross-vocabulary evaluation while capturing semantic nuance. Third, we demonstrate that MTMs enable controllable text generation through topic-based steering vectors without sacrificing generation quality. These contributions demonstrate MTMs' utility as a new approach to topic modeling with distinct advantages on abstract and short-text datasets, and provide a case study in using interpretability tools for downstream tasks beyond model analysis.

2 Related Work

Mechanistic interpretability. We build on work establishing that many high-level concepts in large language models (LLMs) are encoded as recoverable linear directions (Mikolov et al., 2013; Elhage et al., 2022; Park et al., 2024), and that dictionary learning methods such as SAEs can extract these directions at scale (Yun et al., 2021; Bricken et al., 2023; Tamkin et al., 2024; Templeton et al., 2024; Cunningham et al., 2024; Gao et al., 2025a). Prior applications have primarily focused on model interpretability and control—using extracted directions as steering vectors or for ablation (Rimsky et al., 2024; Turner et al., 2023; Tan et al., 2024; Arditi et al., 2024)—with applications in refusal mitigation, enhancing truthfulness, reasoning correction, and style transfer, among others (Sakarvadia et al., 2023; Hernandez et al., 2024; Arditi et al., 2024;

O'Brien et al., 2025; Cao et al., 2024; Wang et al., 2025). We extend this line of research by applying SAE features beyond their original interpretability contexts and demonstrate their usefulness for discovering topics. Recent work has identified some limitations to SAEs, such as underperformance on downstream tasks (Smith et al., 2025; Wu et al., 2025) and challenges to the linear representation of concepts in LLMs (Engels et al., 2025). However, these concerns are less critical for our topic modeling application, which uses SAEs for semantic featurization and requires only that some high-level features are represented linearly.

Neural topic models. Neural topic models address limitations of purely probabilistic approaches (Blei et al., 2003). They generally fall into three distinct paradigms. The first paradigm involves probabilistic models aiming to reconstruct a word count matrix, often augmented with pretrained embeddings (Burkhardt and Kramer, 2019; Dieng et al., 2020; Bianchi et al., 2021a,b; Wu et al., 2024). The second paradigm frames topic discovery as a clustering task, leveraging embeddings derived from pretrained neural models (Angelov, 2020; Grootendorst, 2022; Zhang et al., 2022). The third paradigm employs LLMs directly, using prompt-based techniques to aggregate or define topics (Pham et al., 2024). Mechanistic Topic Models (MTMs) extend the first two paradigms by using SAEs instead of standard embeddings; concurrent work also connects SAEs to topic modeling by viewing the SAE objective itself as MAP inference in a continuous topic model (Girrbach and Akata, 2026). For MTMs, SAE features enable richer, context-aware topic descriptions and allow for controlled text generation through learned steering vectors. We do not directly compare MTMs against third-paradigm models such as TopicGPT (Pham et al., 2024); despite strong performance, their training requires many costly LLM API calls.

Topic model evaluation. Automated and human coherence metrics (Chang et al., 2009; Newman et al., 2010; Lau et al., 2014) have long been the standard for topic model evaluation, but are known to be imperfect proxies for human preferences (Hoyle et al., 2021; Doogan, 2022). Recent work has further shown that coherence metrics can miss important aspects of topic quality, such as document-space organization (Pereira et al., 2025), hierarchical consistency (Viegas et al., 2025), and the relationship between documents and the topics they are assigned to (Pereira et al., 2025). Topic judge is motivated by similar concerns: it goes beyond coherence by directly evaluating how well topics describe their assigned documents, providing a closer proxy for human preferences compared to topic quality. Recently, LLMs have demonstrated effectiveness as scalable evaluators across diverse language tasks (Naismith et al., 2023; Chiang and Lee, 2023; Stammbach et al., 2023; Li et al., 2025). Pairwise preference rankings by LLMs have proven particularly useful in contexts where relative comparisons are straightforward but eliciting global rankings or pointwise scores is challenging, such as in chatbot evaluation (Zheng et al., 2023; Li et al., 2024; Liu et al., 2024; Chiang et al., 2024; Gao et al., 2025b). Building on these insights, we introduce a tournament-style evaluation framework that leverages pairwise LLM judgments to systematically compare topic models.

3 Background

The *linear representation hypothesis* (Mikolov et al., 2013; Elhage et al., 2022; Park et al., 2024; Costa et al., 2025) suggests that LLMs encode many high-level features as linear directions in their activation spaces. It can be formalized as follows:[2]

Definition 1. ***Linear Representation Hypothesis (LRH):*** *Any activation vector $\mathbf{a} \in \mathbb{R}^H$, where H is the hidden dimension of the transformer, produced at a given layer can be decomposed as*

$$\mathbf{a} = \sum_{i=1}^{W} \alpha_i \mathbf{w}_i + \mathbf{b}, \tag{1}$$

where

- $\mathbf{b}$ *is an input-independent constant vector,*
- *the set* $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_W\}$ *consists of nearly orthogonal unit vectors (i.e.,* $\|\mathbf{w}_i\|_2 = 1$ *and* $|\langle \mathbf{w}_i, \mathbf{w}_j \rangle| < \epsilon$ *for* $i \neq j$*, with* ϵ *being a small positive constant),*
- *each vector* $\mathbf{w}_i$ *corresponds to a human-interpretable feature (e.g., semantic content, syntactic structure, or style),*
- *each scalar* α_i *represents the strength of feature* i *in the activation vector* $\mathbf{a}$*, with sparse activation (i.e.,* $|\{i : \alpha_i \neq 0\}| \ll W$*),*
- *the number of vectors* W *is typically much larger than their dimension* H*.*

[2] Several formalizations exist, but we adopt a definition close to that of Costa et al. (2025), as we think it adheres closely to its conventional usage.

This decomposition implies that (i) high-level semantic features that LLMs extract from text can be recovered from model activations, and (ii) we can construct *steering vectors* $\mathbf{s} = \sum_{i=1}^{W} \delta_i \mathbf{w}_i$ that when added to $\mathbf{a} = \sum_{i=1}^{W} \alpha_i \mathbf{w}_i + \mathbf{b}$, are equivalent to modifying its feature strengths $\mathbf{a} + \mathbf{s} = \sum_{i=1}^{W} (\alpha_i + \delta_i)\mathbf{w}_i + \mathbf{b}$. They can be used in generating text to modulate the expression of particular features, by setting $\delta_i > 0$ to increase the expression of feature i and $\delta_i < 0$ to decrease it.

To identify the LRH feature directions $\{\mathbf{w}_i\}_{i=1}^{W}$, we can train *sparse autoencoders* (SAEs)—unsupervised models that learn to reconstruct LLM activations. SAEs are parameterized by a single-layer neural network,

$$\boldsymbol{\alpha}(\mathbf{a}) = \sigma(\mathbf{W}_{\text{in}}\mathbf{a} + \mathbf{b}_{\text{in}}), \tag{2}$$

$$\hat{\mathbf{a}}(\mathbf{a}) = \mathbf{W}_{\text{out}}\boldsymbol{\alpha}(\mathbf{a}) + \mathbf{b}_{\text{out}}, \tag{3}$$

where σ is an activation function (e.g., ReLU, JumpReLU), and the learnable parameters are $\mathbf{W}_{\text{in}} \in \mathbb{R}^{W \times H}$, $\mathbf{W}_{\text{out}} \in \mathbb{R}^{H \times W}$, $\mathbf{b}_{\text{in}} \in \mathbb{R}^{W}$, and $\mathbf{b}_{\text{out}} \in \mathbb{R}^{H}$. The network is trained to minimize a regularized reconstruction error,

$$\mathcal{L}(\mathbf{a}) = \frac{1}{2}\|\hat{\mathbf{a}}(\mathbf{a}) - \mathbf{a}\|_2^2 + \lambda\|\boldsymbol{\alpha}(\mathbf{a})\|_0. \tag{4}$$

Once trained, the feature directions $\{\mathbf{w}_i\}_{i=1}^{W}$ are identified with the columns of $\mathbf{W}_{\text{out}}$. The encoder network $\boldsymbol{\alpha}(\mathbf{a})$ can also be used to decompose any activation onto the feature directions and obtain its feature activation strengths α_i.

After training, the learned feature directions can be automatically interpreted into textual descriptions $\{\mathtt{d}_i\}_{i=1}^{W}$ through computational methods (Bills et al., 2023; Paulo et al., 2025; Templeton et al., 2024). A common approach involves computing feature activation strengths over a large corpus, selecting tokens with the highest activation strengths along with their surrounding context, and prompting an LLM to produce a short description of the underlying feature (Bills et al., 2023).

4 Mechanistic Topic Models

We introduce Mechanistic Topic Models (MTMs), which extend topic modeling by using SAE features. This shift provides three advantages: (1) semantic richness, as SAE features capture context-aware and semantically abstract concepts; (2) topic descriptions that can articulate complex themes that are hard to convey through word lists alone; and (3) topic steering vectors that can be used for topic-based controlled generation.

All MTMs share the same workflow. Given a corpus $\mathcal{D}$ of D documents and a desired number of topics K, we first transform documents into SAE feature counts (Section 4.1), then learn topic-feature weights $\boldsymbol{\beta}_k \in \mathbb{R}_+^W$ and document-topic distributions $\boldsymbol{\theta}_d \in \Delta^{K-1}$ (Section 4.2), generate interpretable topic descriptions $\mathtt{t}_k$ from learned features (Section 4.3), and construct steering vectors $\mathbf{s}_k$ for controllable generation (Section 4.3). We first describe the featurization process and then detail three specific MTM variants.

4.1 From Documents to SAE Features

MTMs represent documents as SAE feature counts rather than word counts or raw embeddings. This presents two challenges.

First, unlike words that either appear or not, SAE features have continuous activations at each token position. For a document d containing N_{tok} tokens, let $\mathbf{a}_{d,j} \in \mathbb{R}^H$ denote the LLM hidden-state activation at the chosen layer for token j. We use a thresholding approach to count how often each feature i activates strongly across $(\mathbf{a}_{d,1}, \ldots, \mathbf{a}_{d,N_{\text{tok}}})$,

$$\tilde{c}_{d,i} = \sum_{j=1}^{N_{\text{tok}}} \mathbb{1}\{\alpha_i(\mathbf{a}_{d,j}) > q_i\}, \tag{5}$$

where $\alpha_i(\mathbf{a}_{d,j})$ is feature i's activation on token j, and q_i is the 80th percentile of feature i's nonzero activation distribution on the original SAE training data, although any dataset of interest can be used. This approach produces interpretable counts, adapts to each feature's typical activation range, and prevents activation false positives.

Second, SAEs can learn spurious features with unclear meanings or mislabeled descriptions. We address this through an LLM-based preprocessing step that filters out likely spurious or topic-irrelevant features (e.g., low-level grammatical features), and an optional post-training refinement of topic descriptions. These quality control measures are detailed in Appendix A.

The feature vectors $\{\tilde{\mathbf{c}}_d\}_{d=1}^{D}$ serve as input to all MTMs described below. For the rest of the paper, we use W to denote the number of features after this filtering.

4.2 MTM Variants

Having transformed documents into SAE feature vectors $\{\tilde{\mathbf{c}}_d\}_{d=1}^{D}$, we now apply topic modeling

algorithms to these representations.

We propose three variants with different adaptation strategies: mechanistic LDA (mLDA) provides a straightforward extension of LDA (Blei et al., 2003) by treating features as words; mechanistic ETM (mETM) adapts ETM by representing topics as LLM activation vectors (Dieng et al., 2020); and mechanistic BERTopic (mBERTopic) takes a clustering approach using SAE feature directions to construct document embeddings (Grootendorst, 2022).

4.2.1 Mechanistic LDA (mLDA)

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a foundational probabilistic topic model that represents documents as mixtures of topics and topics as distributions over words. Mechanistic LDA (mLDA) adapts this model by replacing the topic-word distributions with distributions over SAE features. Following LDA's generative process, we assume

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}_W(\eta), \tag{6}$$
$$\boldsymbol{\theta}_d \sim \text{Dirichlet}_K(\alpha), \tag{7}$$

where $\boldsymbol{\beta}_k \in \Delta^{W-1}$ is a distribution over the W feature directions learned by the SAE and $\boldsymbol{\theta}_d \in \Delta^{K-1}$ is the document distribution over K topics.

The SAE feature counts are generated by

$$\tilde{\mathbf{c}}_d \sim \text{Multinomial}(\boldsymbol{\theta}_d \boldsymbol{\beta}, N_{\text{sae}}), \tag{8}$$

where $N_{\text{sae}} = \sum_i \tilde{c}_{d,i}$ is the total SAE feature count in the document. While the multinomial assumption has limitations (Section 4.2.2), here we retain it to leverage existing LDA inference algorithms. Depending on the dataset size, we use standard collapsed Gibbs sampling (Griffiths and Steyvers, 2004) or variational inference (Blei et al., 2003) to approximate posterior distributions over $\{\boldsymbol{\beta}_k\}$ and $\{\boldsymbol{\theta}_d\}$ (see Section C).

4.2.2 Mechanistic ETM (mETM)

The Embedded Topic Model (ETM) (Dieng et al., 2020) represents topics as vectors in word embedding space, leveraging these word embeddings to capture semantic relationships. Mechanistic ETM (mETM) extends this idea to the space of LLM activations, representing each topic k as a learned LLM activation $\boldsymbol{v}_k \in \mathbb{R}^H$.

The generative process first samples document-topic proportions from a logistic-normal,

$$\boldsymbol{\delta}_d \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{9}$$
$$\boldsymbol{\theta}_d = \text{softmax}(\boldsymbol{\delta}_d). \tag{10}$$

Each topic-feature distribution $\boldsymbol{\beta}_k \in [0,1]^W$ is obtained by transforming the learned activation $\boldsymbol{v}_k$ as in the SAE encoder,

$$\boldsymbol{\beta}_k = \sigma\left(\mathbf{W}_{\text{in}} \boldsymbol{v}_k + \mathbf{b}\right), \tag{11}$$

where $\mathbf{W}_{\text{in}}$ is initialized with the SAE encoder matrix and fixed during training, $\mathbf{b}$ is a learned bias vector, and σ is the sigmoid function. Here, $\beta_{k,i}$ represents the probability that feature i is active in a document token from topic k.

Unlike in mLDA, feature counts are drawn from a binomial distribution

$$\tilde{c}_{d,i} \sim \text{Binomial}([\boldsymbol{\theta}_d \boldsymbol{\beta}]_i, N_{\text{tok}}). \tag{12}$$

This distribution is chosen to respect the constraint that each feature activates at most once per token. Following Dieng et al. (2020), we use amortized variational inference with neural networks $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\sigma}_\phi$ to parameterize the variational distribution $q_\phi(\boldsymbol{\delta}_d) = \mathcal{N}(\boldsymbol{\mu}_\phi(\tilde{\mathbf{c}}_d), \boldsymbol{\sigma}_\phi(\tilde{\mathbf{c}}_d))$. The parameters ϕ, $\{\boldsymbol{v}_k\}$, and $\mathbf{b}$ are jointly optimized by maximizing the ELBO,

$$\sum_{d=1}^{D} \mathbb{E}_{q_\phi}\left[\log \frac{p(\tilde{\mathbf{c}}_d \mid \boldsymbol{\delta}_d, \{\boldsymbol{v}_k\}, \mathbf{b})\, p(\boldsymbol{\delta}_d)}{q_\phi(\boldsymbol{\delta}_d \mid \tilde{\mathbf{c}}_d)}\right]. \tag{13}$$

4.2.3 Mechanistic BERTopic (mBERTopic)

BERTopic (Grootendorst, 2022) frames topic modeling as a clustering problem in document embedding space. Mechanistic BERTopic (mBERTopic) takes a similar approach but forms document embeddings $\tilde{\mathbf{e}}_d$ from the SAE feature representation,

$$\tilde{\mathbf{e}}_d = \frac{1}{N_{\text{tok}}} \sum_{i=1}^{W} \tilde{c}_{d,i} \mathbf{w}_i, \tag{14}$$

where $\mathbf{w}_i$ is the i^{th} feature direction from the SAE decoder matrix.

Following Grootendorst (2022), we apply UMAP (McInnes et al., 2018) for dimensionality reduction followed by HDBSCAN clustering (McInnes et al., 2017) to the document embeddings $\{\tilde{\mathbf{e}}_d\}$. We then extract topic-feature distributions using class-based TF-IDF, which treats each cluster as a meta-document,

$$\beta_{k,i} \propto \text{tf}_{k,i} \cdot \log\left(1 + \frac{A}{\text{tf}_i}\right), \tag{15}$$

where $\text{tf}_{k,i} = \sum_{d \in \mathcal{D}_k} \tilde{c}_{d,i}$ is the count of feature i across all documents in cluster k, A is the average count of all features per cluster, and tf_i is the total count of feature i across all clusters.

We experimented with alternative embeddings in Equation (14), including using the SAE activations as opposed to counts, and skipping the pre-filtering step detailed in Section 4.1. However, the formulation in Equation (14) was consistently chosen by our hyperparameter optimization procedures, which we believe affirms the usefulness and importance of the steps detailed in Section 4.1.

4.3 Using MTMs

Once trained, all the MTMs above produce document-topic proportions $\boldsymbol{\theta}_d$ and topic-feature distributions $\boldsymbol{\beta}_k$ that can be used in downstream applications. Of particular significance to MTMs are topic interpretations and steering vectors.

Interpreting Topics. Topic interpretation in MTMs follows an approach similar to conventional topic models. Given a learned topic k, we identify the top n SAE features by selecting those with the highest weights in the topic-feature vector $\boldsymbol{\beta}_k$. We then construct the textual topic representation $\texttt{t}_k$ from the automatically generated descriptions $\texttt{d}_i$ corresponding to these n features.

The last step can be done in two ways. The TopFeatures (TF) approach directly uses the feature descriptions $\texttt{d}_i$ and concatenates them. The Summarization approach (Sum.) further processes the concatenated text by passing it through an LLM to convert it into a one-sentence summary. Summarization is beneficial, as it can also be applied to word-based models for standardization in evaluations. Figure 1 shows both approaches in use. Section F.2 provides the summarization prompt used in this paper.

Steering. One advantage of MTMs is their ability to steer text generation toward discovered topics. We achieve this by constructing a topic-specific steering vector $\mathbf{s}_k$ that we add to the LLM's activations to bias generation toward topic k.

Each steering vector $\mathbf{s}_k$ is constructed by weighting SAE feature directions $\{\mathbf{w}_i\}$ according to their importance in topic k as captured by the topic-feature weights $\boldsymbol{\beta}_k \in \mathbb{R}_+^W$,

$$\mathbf{s}_k = \frac{\sum_{i \in W} \boldsymbol{\beta}_{k,i} \mathbf{w}_i}{\left\| \sum_{i \in W} \boldsymbol{\beta}_{k,i} \mathbf{w}_i \right\|_2}. \tag{16}$$

Equation (16) is a unit vector that points in the direction most characteristic of topic k in the LLM's activation space.

To control the intensity of topic steering, we use an intervention that first removes any existing topic signal before adding the desired amount. Consider centered activations $\bar{\mathbf{a}} = \mathbf{a} - \mathbf{b}$, where $\mathbf{b}$ is the bias from Equation (1). We decompose the activation into components parallel and perpendicular to the steering direction,

$$\bar{\mathbf{a}}_{\parallel} = (\bar{\mathbf{a}} \cdot \mathbf{s}_k)\mathbf{s}_k, \tag{17}$$

$$\bar{\mathbf{a}}_{\perp} = \bar{\mathbf{a}} - \bar{\mathbf{a}}_{\parallel}. \tag{18}$$

The steered activation replaces the parallel component with a scaled steering vector,

$$\mathbf{a}_{\text{steered}} = \bar{\mathbf{a}}_{\perp} + \lambda \mathbf{s}_k + \mathbf{b}, \tag{19}$$

where λ controls the steering strength, allowing for modulation of topic expression in generated text.

For example, one specific topic learned by mLDA on the Bills dataset (Section 5) places high weights on SAE features with descriptions: "specific legal terms and conditions related to immigration status", "references to government policies and legal regulations", and "references to labor conditions and economic structures". We then form a steering vector $\mathbf{s}_k$ from this topic using Equation (16). Applying this steering vector to the prompt "A text about" results in the continuation " a person who is not of the United States, but has been granted permission to enter the country. The term 'temporary resident' (TR) refers to people who have entered the U.S. and are allowed to stay in the US...", an example of successful topic-guided text generation.

5 Empirical Studies

We evaluate MTMs and baselines using standard topic modeling metrics (coherence, topic diversity, and alignment) and a novel metric, topic judge, which performs a series of comparisons between pairs of models, asking an LLM judge which set of topics they prefer with respect to a reference document from the corpus. We find that an MTM is preferred by topic judge in most datasets, while MTMs are largely comparable to baselines on standard metrics, with particular strengths on abstract and short-text datasets. Finally, we show that MTMs capture novel topics, and that their topics can be used to steer LLM generations.

Data		D-VAE	FAST	LDA Base	LDA MTM	ETM Base	ETM MTM	BERTopic Base	BERTopic MTM
20NG	TF	1094	1419	1580	**1613**	1602	1601	1544	1548
	Sum.	1322	1470	1514	**1614**	1510	1573	1484	1514
Bills	TF	1151	1316	**1763**	1571	1670	1584	1509	1436
	Sum.	1395	1447	1578	**1638**	1536	1630	1410	1366
Wiki	TF	1304	1288	1575	1641	1559	**1641**	1443	1549
	Sum.	1425	1424	1551	**1628**	1494	1620	1387	1471
Yelp	TF	1095	1271	1644	**1715**	1558	1675	1430	1611
	Sum.	1342	1383	**1627**	1621	1488	1609	1347	1585
AGN.	TF	1402	1462	1552	**1576**	1511	1537	1476	1484
	Sum.	1480	1491	1489	1590	1431	**1607**	1442	1469
GoEmo.	TF	1185	1431	1442	**1728**	1444	1664	1428	1678
	Sum.	1346	1419	1452	**1630**	1449	1630	1463	1611
Poem.	TF	1185	1306	1573	1662	1524	**1678**	1453	1619
	Sum.	1360	1412	1522	1536	1446	1597	1530	**1598**
WriPro.	TF	918	1263	1545	**1813**	1538	1808	1385	1731
	Sum.	1373	1375	1439	1674	1430	**1678**	1394	1637

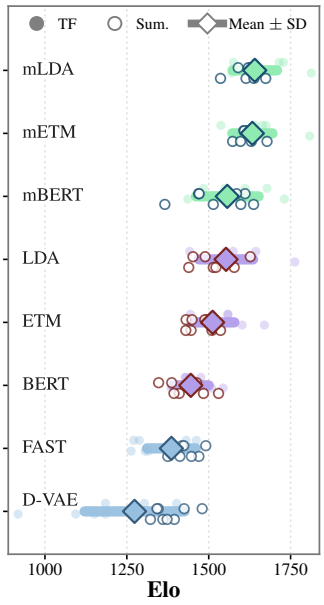

Table 1: Topic judge Elo scores across datasets, with a model-level summary at right. The first five datasets are standard benchmark datasets; the last three are hard (short and/or more abstract). **Bold** indicates the row best (bootstrap confidence interval, $p < 0.05$); underlining indicates $p > 0.05$ against bold. Green indicates an equal or higher score than the corresponding word-based model, and purple a lower score.

Datasets. We study eight datasets spanning a range of domains, document lengths, and thematic characteristics: online newsgroup posts (20NG; Lang 1995), bill summaries from the 110–114$^{\text{th}}$ U.S. congresses (Bills; Adler and Wilkerson 2018; Hoyle et al. 2022), Wikipedia articles labeled as "good" by editors (Wiki; Merity et al. 2017), short Reddit comments expressing emotion (GoEmotions; Demszky et al. 2020), a collection of poems (PoemSum; Mahbub et al. 2023), news articles across four categories (AGNews; Zhang et al. 2015), business reviews (Yelp Polarity; Zhang et al. 2015), and creative writing stories (WritingPrompts; Fan et al. 2018).

We expected GoEmotions to be more challenging due to its short documents, and PoemSum and WritingPrompts to be more challenging due to their abstract themes.

Section B contains dataset statistics, preprocessing steps for word-based models, and information on labels used for the topical alignment metric.

Models. We compare to the word-based counterparts of MTMs: LDA (Blei et al., 2003), ETM (Dieng et al., 2020), and BERTopic (Grootendorst, 2022). We also compare to two other neural models: Dirichlet VAE (D-VAE) (Burkhardt and Kramer, 2019), which is a VAE product-of-experts model; and FASTopic (Wu et al., 2024), which is a model using optimal transport alongside pretrained embeddings.

Setup. For MTMs, we use the GemmaScope family of SAEs trained on Gemma 2-9B LLM activations (Gemma Team et al., 2024; Lieberum et al., 2024). We use the 16k SAE from layer 10 throughout, as we found empirically that it provides the most useful features for topic modeling, though practitioners may prefer other layers depending on the characteristics of the dataset. The SAE feature metadata, including descriptions, are downloaded from Neuronpedia (Lin, 2023). The implementation details for the baseline models are in Section C.

For our experiments, we choose the number of topics to be $K = 50$ for AGNews, GoEmotions, and PoemSum, as these are the smallest datasets, and $K = 100$ for the remaining datasets. To select model hyperparameters, we use Bayesian optimization and optimize the topic quality metric proposed in Dieng et al. (2020) for each model-dataset pair. Section D contains details on this procedure.

5.1 Topic Judge

Topic model evaluation is challenging. Existing metrics have limitations: topical alignment (Hoyle et al., 2022) requires labeled data for attributes of interest and does not assess topic description quality; coherence metrics like ratings or intrusion

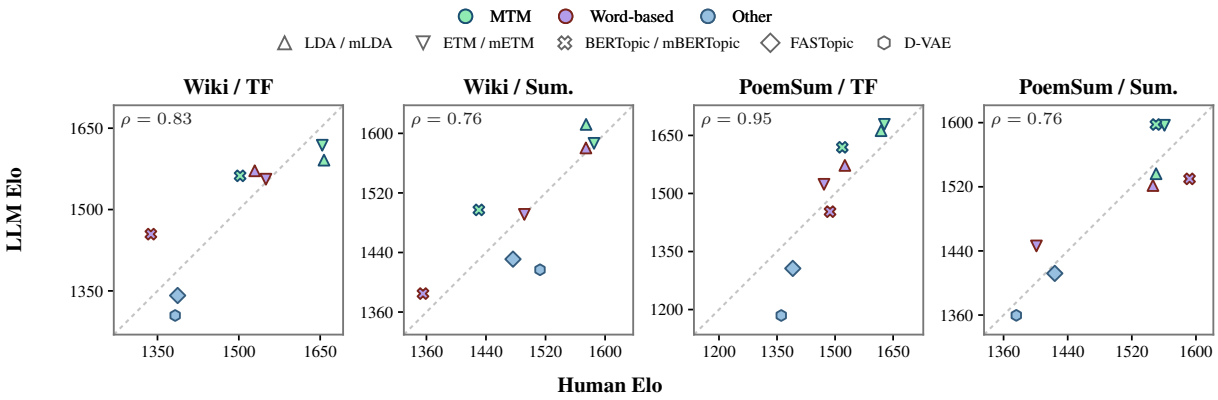

Figure 2: Human vs. LLM topic-judge Elo scores on Wiki and PoemSum. Each panel compares the Elo scores assigned by the human study and the GPT-4.1 judge for one dataset/representation condition. The dashed diagonal indicates perfect agreement, and each panel reports Spearman's ρ. The panels and corresponding ρ values show strong agreement between human and LLM Elo scores.

(Chang et al., 2009; Newman et al., 2010) measure feature relatedness in topics but ignore their relationship to documents; and automated metrics like C_{NPMI} may correlate poorly with human judgments (Hoyle et al., 2021) and cannot compare models with different vocabularies. These limitations motivated us to develop *topic judge*, a new evaluation framework for topic models.

Methodology. Topic judge is inspired by Chatbot Arena-style rankings (Chiang et al., 2024). It evaluates topic models using pairwise comparisons, where an LLM judge determines which model's topic assignments better capture a given document's content. The comparison results are then aggregated via a Bradley-Terry model (Bradley and Terry, 1952). The assumption is that superior topic models should produce topics that are more descriptive of the documents they are assigned to.

In detail, the method works by performing pairwise comparisons between all model pairs (m, m') to score M models. For each of T comparisons per pair, topic judge: (1) randomly samples a document d; (2) identifies the set of top topics for each model according to $\boldsymbol{\theta}_d$—either the top q topics or all topics with cumulative mass below a threshold p, whichever yields fewer topics; (3) creates text representations for each topic set $\mathtt{t}_k$ using either approach (TopFeatures or Summarization); and (4) prompts an LLM judge to select which representation better captures the document.

After collecting all pairwise comparisons, topic judge aggregates the results using a Bradley-Terry model to compute final scores. This model assumes that model m defeats model m' with probability $\sigma(\pi_m - \pi_{m'})$, where π_m represents model m's strength and σ is the sigmoid function. We infer these parameters via maximum likelihood and convert them to Elo scores (normalized to average 1500) for interpretability. Elo scores are a rescaling of the Bradley-Terry strength parameters such that $P(m \succ m') = 1/(1 + 10^{(E_{m'} - E_m)/400})$, where E_m and $E_{m'}$ are the Elo scores of models m and m'.

Topic judge addresses limitations of existing metrics: it evaluates topic-document relationships directly without requiring labeled data, uses pairwise comparisons that capture relative quality differences across different vocabularies (words vs. SAE features), and leverages an LLM judge to assess semantic content directly rather than through proxies like C_{NPMI}. For the specific prompt and examples with judge responses, see Section G.

Results. We perform 100 comparisons for each model pair (2800 comparisons per dataset) using GPT-4.1 (OpenAI, 2025) with temperature 0, prompting the LLM to choose the set of topics that best captures the general meaning of the document.

Table 1 shows that an MTM achieves the highest Elo score in 14 of 16 dataset–representation rows, and MTMs outperform their word-based counterparts in 42 of 48 pairwise comparisons. Of all 48 comparisons, 71% show an MTM advantage of at least 50 points (∼57% win probability), 50% exceed 100 points (∼64%), and 21% exceed 200 points (∼76%). The main exception is Bills with top features, where all three word-based models outscore their MTM counterparts; we attribute this to the judge finding MTM's SAE feature descrip-

	20NG		Bills		Wiki		Yelp		AGNews		GoEmo.		PoemSum		WriPro.	
Model	R	I	R	I	R	I	R	I	R	I	R	I	R	I	R	I
D-VAE	2.09	0.55	2.58	0.66	2.93	**0.91**	1.75	0.46	2.79	0.81	1.77	0.38	1.76	0.37	1.77	0.44
FASTopic	2.43	0.71	2.70	0.74	2.89	0.87	2.22	0.59	2.82	0.85	2.70	0.81	1.97	0.43	2.33	0.66
LDA	2.65	0.75	**2.96**	0.71	2.83	0.77	2.35	0.61	2.83	0.75	2.20	0.56	2.30	0.55	2.63	0.79
ETM	2.40	0.74	2.81	0.72	2.72	0.84	2.32	0.65	2.43	0.60	1.90	0.37	2.17	0.51	2.55	0.81
BERTopic	**2.87**	0.80	2.93	0.78	**2.97**	0.86	**2.84**	0.74	**2.92**	0.81	2.76	0.80	2.36	0.51	2.53	0.65
MTM (w/ LDA)	2.63	0.83	2.88	**0.81**	2.77	0.85	2.77	0.78	2.71	0.80	2.77	0.80	**2.61**	0.73	**2.78**	0.80
MTM (w/ ETM)	2.57	0.80	2.87	0.79	2.79	0.85	2.60	0.73	2.46	0.73	2.70	0.76	2.54	0.70	2.65	0.76
MTM (w/ BERTopic)	2.58	**0.83**	2.75	0.81	2.76	0.85	2.73	0.77	2.74	0.83	**2.80**	**0.88**	2.28	0.53	2.61	0.71

Table 2: Ratings (R) and intrusion (I) across all datasets, averaged over five runs. **Bold** indicates the column best (two-sided t-test, $p < 0.05$); underlining indicates $p > 0.05$ against bold. Color coding: for ratings, purple (< 2.0) to green (> 2.8); for intrusion, purple (< 0.5) to green in 0.1 increments.

tions less specifically relevant than the precise keywords learned by word-based models (Section 4.1). Switching to the summarization representation recovers mLDA and mETM on Bills, though mBERTopic remains lower. The advantages are largest on the challenging datasets—GoEmotions, PoemSum, and WritingPrompts—where MTMs win by an average of 195 Elo points, suggesting that SAE features capture semantic nuances beyond word co-occurrence patterns.

To validate the LLM judge, we conducted a human evaluation study on Wiki and PoemSum. We recruited 68 participants from an adult population with at least a bachelor's-level English education. Each participant was shown a document alongside topic representations from multiple models and asked to judge which topics best capture the document's meaning. For statistical efficiency, each participant ranked 5 randomly sampled models per document across 10 tasks; we then decomposed rankings into pairwise comparisons and fit a Bradley-Terry model to obtain Elo ratings. Further details are provided in Section H.

Table 8 reports the resulting human Elo scores. MTMs achieve the highest scores in three of four conditions; the exception is PoemSum with summaries, where BERTopic ranks first but is not significantly different from MTM (w/ ETM). To assess agreement between human and LLM evaluators, we plot their Elo scores against each other in Figure 2 and compute Spearman's ρ for each condition. The points cluster closely around the diagonal, and the scores are strongly correlated (ρ between 0.76 and 0.95), with agreement holding for both MTMs and word-based models. This confirms that the LLM judge is a reliable proxy for human preferences.

5.2 Standard Evaluation Metrics

While topic judge is a holistic measure of model performance, we also report two complementary metrics: *coherence*—the semantic relatedness of a topic's top features—and *topic diversity*—the distinctiveness of topics.

We measure coherence in two ways. First, we measure the average rating assigned to each topic on a 1–3 scale based on how semantically related its top features are (R) (Newman et al., 2010). Second, we measure how accurately an evaluator can identify an "intruder" feature from another topic when it's mixed with the target topic's features (I) (Chang et al., 2009). Following Stammbach et al. (2023), we use GPT-4.1 as the rater and evaluator in both tasks (see Section F.3 for the prompts). For ratings, we present the top 10 features per topic and report the average rating across all topics. For intrusion, we perform 25 trials with 5 true features and 1 intruder per trial and report the average accuracy. For both, we use zero temperature to sample.

To measure topic diversity (TD), we use the metric by Dieng et al. (2020) described in Section D.3.

Results in Table 2 show that MTMs achieve 2.5–2.9 coherence on benchmark datasets, although word-based models outperform MTMs in some cases. On the challenging datasets, MTMs perform well, indicating that SAE directions remain interpretable even for very short texts (GoEmotions) or abstract themes (PoemSum, WritingPrompts). In topic diversity, mLDA and mETM are comparable to word-based models (Table 7). However, mBERTopic learns some topics with substantial overlap on PoemSum and WritingPrompts (0.38 and 0.36 diversity) and has a lower coherence score of 2.28 on PoemSum. Further work is needed to improve mBERTopic's robustness across datasets.

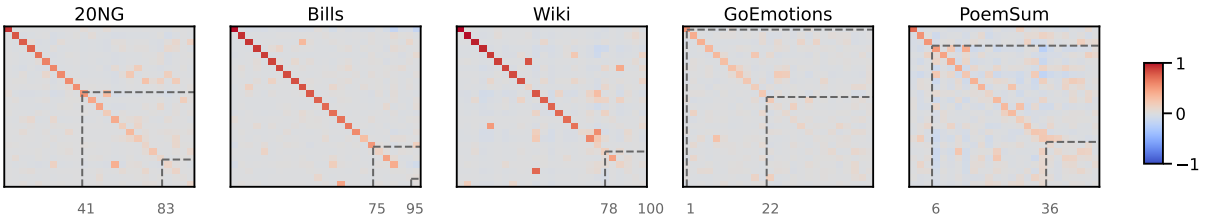

Figure 3: Heatmap representations of how similar mLDA topics (columns) and LDA topics (rows) are in terms of their proportions across documents; topics are aligned greedily. The two models learn similar topics on Bills and Wiki, but mLDA finds new topics on 20NG, GoEmotions, and PoemSum. Dashed gray boxes show submatrices: within, the two models' topics have medium (≤ 0.5) or low (≤ 0.2) correlations.

Finally, we report topical alignment (Table 13) and model stability (Table 11) in the Appendix. MTMs achieve comparable alignment compared to word-based models on benchmark datasets, and learn topics that are as or more stable than their word-based counterparts under the stability metric of Hoyle et al. (2022).

5.3 Topic Novelty Analysis

Having established that MTMs are competitive on standard metrics and preferred by topic judge, we now investigate whether they discover new topics or instead redescribe topics that word-based models already find. To address this, we analyze how similarly topics from different models weigh documents—if two topics assign similar weights across documents, they likely capture the same concepts regardless of description.

We formalize this idea by computing correlations between document-topic distributions. Consider trained models m and m' with document-topic matrices $\boldsymbol{\theta}^m, \boldsymbol{\theta}^{m'} \in \mathbb{R}_+^{D \times K}$, where row d contains document d's topic weights and column $\boldsymbol{\theta}^m_{:,i}$ is topic i's prevalence across documents. We compute the cross-correlation matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$, where each entry is $C_{i,j} = \text{corr}(\boldsymbol{\theta}^m_{:,i}, \boldsymbol{\theta}^{m'}_{:,j})$.

Figure 3 shows these correlations after greedy alignment: we iteratively select the unpaired column with the strongest correlation in the entire matrix and pair it from the remaining rows, placing matches consecutively along the diagonal (see Section J for details). Dashed boxes mark regions where correlations fall below 0.5 or 0.2, highlighting where MTMs discover different topics.

On Bills and Wiki, over 70% of mLDA topics correlate at least 0.5 with LDA topics, suggesting that the topics represent similar themes. On GoEmotions and PoemSum, nearly all topics fall below 0.5 correlation (many below 0.2), indicating they are new. 20NG lies between these extremes—we think this is due to MTMs capturing stylistic qualities like argumentation that word co-occurrence misses. ETM and BERTopic pairs show similar patterns. The three additional datasets (Figure 7) also vary: AGNews shows moderate overlap, while Yelp and WritingPrompts topics are largely novel. We provide examples of MTM topics, including novel topics, in Figure 10.

5.4 Steering Evaluation

MTMs enable text generation via steering vectors formed from discovered topics (Section 4.3). To evaluate this capability, we conduct three experiments. We measure two criteria for steering: *topic relevance*—steering increases the expression of the target topic in text; and *fluency*—steering preserves the coherence and naturalness of the generated text.

Throughout, we use $p(\cdot)$ and $p(\cdot; \lambda \mathbf{s}_k)$ for unsteered and steered LLM probabilities, respectively (steering vector $\mathbf{s}_k$, magnitude λ), $\mathbf{x} = (x_1, \ldots, x_{N_{\text{gen}}})$ for sampled token sequences, and $\mathbf{x}_d = (x_{d,1}, \ldots, x_{d,N_{\text{tok}}})$ for document d sequences. We describe each experiment below, with additional details in Section E.

Topic Relevance Win Rate. We first verify whether MTM steering vectors guide LLM-generated text toward exhibiting specific topics. We take a "best-of-$2L$" approach: for each topic k, we sample L steered texts with different steering strengths and L unsteered texts using the same prompt. An LLM judge then selects the text most representative of the topic according to its summarized description $\texttt{t}_k$. We record a win if any steered text is chosen.

We repeat this procedure R times for each topic. We define the topic relevance win rate (TWR) as the fraction of comparisons where the judge selects one of the L steered samples, averaged across the

Topic Summary: "Achievements, statistics, and aspirations in college and professional sports, focusing on tournaments, records, player performance, and team accomplishments."	
$\lambda = 10$	"A text about the history of a place, its people and their customs.\n\nThe book is written in an easy language, with many illustrations to make it more interesting for children. It also contains some words in the local dialect [...]"
$\lambda = 20$	"A text about the 2019 novel coronavirus (COVID-19) outbreak in China has been circulating on WhatsApp. The message claims that a new strain of the virus, which is more deadly than SARS and MERS, has emerged from [...]"
$\lambda = 30$	"A text about the 2018 NBA draft has been circulating on social media, and it's a doozy.\n\nThe text is from an anonymous source who claims to have inside information that says Zion Williamson will be drafted by [...]"
$\lambda = 40$	"A text about the 2014-15 NBA All Star Game was announced on Thursday.\n\nThe game will be held at Madison Square Garden in New York City, and it will feature a team of all stars from both leagues. The game is scheduled."
$\lambda = 50$	"A text about the best-selling player in college basketball, a 10th place finish and an All-American selection.\n\nThose are just some of the accomplishments that have been achieved by Baylor's senior forward John Wall."

Table 3: Examples of generated text at various steering strengths (λ) for a sports-related topic found by mLDA on the Wiki dataset. Higher λ values result in increasingly topic-focused content.

KR trials. When computed for a single topic, a TWR greater than 0.5 indicates that steering biased text generation toward the specified topic.

Topic Likelihood Difference. We next assess if steering vectors capture topic semantics by comparing their effect on the likelihood of different documents from the training corpus. Intuitively, if $\mathbf{s}_k$ represents topic k, increasing the steering strength λ should increase the likelihood of documents about topic k more than other documents from the corpus.

Formally, let $\mathcal{D}_k$ contain documents highly associated with topic k and $\mathcal{D}_{-k}$ be an equally sized random sample from the set of documents highly associated with another topic (see Section E.3). We measure the relative log likelihood difference when steering as

$$\Delta\ell_k(\lambda) = \frac{1}{S}\sum_{i=1}^{S} \log \frac{p(\mathbf{x}_{d_i^+}; \lambda\mathbf{s}_k)}{p(\mathbf{x}_{d_i^-}; \lambda\mathbf{s}_k)}, \qquad (20)$$

where $d_i^+ \in \mathcal{D}_k$ and $d_i^- \in \mathcal{D}_{-k}$. A positive value for $\Delta\ell_k(\lambda)$ when $\lambda \gg 0$ indicates effective topical steering, while a negative value when $\lambda = 0$ indicates effective topic ablation. We let $\Delta\ell(\lambda)$ without subscript denote the average over all topics.

We also use this metric to evaluate partial steering vectors $\mathbf{s}_k^{(n)}$, constructed from only the top n features in the topic, to verify the advantage of using the full steering vector.

Perplexity. Finally, we evaluate whether steering maintains natural-sounding generations. We compute the perplexity (PPL) of the sampled steered generations $\mathbf{x}' \sim p(\mathbf{x}'; \lambda\mathbf{s}_k)$ under the original model $p(\mathbf{x})$. We report the perplexity for 10 generations per topic under five values of λ, and the perplexity of unsteered generations as a baseline. Values close to the baseline indicate that fluency is comparable to that of the unsteered texts.

	20NG	Bills	Wiki	Yelp	AGN	GoEmo	Poem	WP
TWR (%)	88.7	98.9	95.1	92.6	96.6	84.4	85.8	87.7
$\text{PPL}_{\text{control}}$	6.23	6.24	6.24	6.23	6.24	6.25	6.23	6.22
$\text{PPL}_{\lambda=10}$	6.36	6.27	6.43	6.40	6.34	6.31	6.46	6.33
$\text{PPL}_{\lambda=20}$	6.93	6.83	7.04	7.14	6.72	6.61	7.19	7.00
$\text{PPL}_{\lambda=30}$	7.97	8.04	7.72	8.62	7.49	7.83	8.89	8.24
$\text{PPL}_{\lambda=40}$	9.82	9.42	9.23	11.14	8.61	10.68	11.55	10.82
$\text{PPL}_{\lambda=50}$	14.52	12.49	13.03	17.11	11.02	20.13	16.86	16.95

Table 4: Steering metrics for mLDA across all 8 datasets. TWR: fraction of times steered text better matches target topic. PPL: perplexity of steered generations under the original model at various steering strengths λ.

5.4.1 Results

Table 3 shows representative examples of text steered toward a sports-related topic. This topic was discovered by mLDA trained on the Wiki dataset, and the table illustrates outputs generated using a range of steering strengths (λ). At lower values of λ, generated texts remain generic and off-topic, while at higher values, the content aligns with the sports theme. Additional generated texts for other models, datasets, and steering strengths are provided in Table 6.

Quantitatively, steering improves topic relevance. As shown in Table 4, the topic relevance win rate (TWR) exceeds 84% across all datasets, reaching 99% on Bills. This indicates that steering guides text generation toward intended topics. Table 10

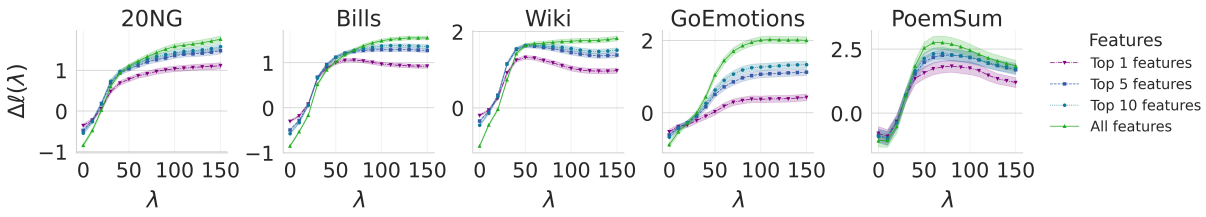

Figure 4: Average document log likelihood differences $\Delta\ell(\lambda)$ between on-topic and off-topic documents for mLDA. Lines show $\Delta\ell(\lambda)$ with steering vectors $\mathbf{s}_k^{(n)}$ constructed using n = 1, 5, 10, or all SAE features. Negative $\Delta\ell$ at $\lambda = 0$ shows that the topics are ablated; positive $\Delta\ell$ at higher λ demonstrates steering toward target topics, with better steering when using all features.

shows that higher steering strengths win more often: $\lambda \in \{30, 40, 50\}$ accounts for 80% of mLDA wins, 77% of mETM wins, and 86% of mBERTopic wins. Figure 4 further illustrates this effect by showing how relative document log likelihood changes as steering strength increases. At baseline ($\lambda = 0$), ablation reduces log likelihood for on-topic documents. Moreover, positive steering ($\lambda > 0$) increases the likelihood of on-topic documents relative to off-topic ones. Steering vectors that use all topic features produce the largest differences, indicating that the full feature set effectively characterizes topic expression. Figure 9 shows log likelihood plots for mLDA on Yelp, AGNews, and WritingPrompts; Figure 8 shows the corresponding mETM and mBERTopic plots.

Finally, the perplexities of steered outputs (Table 4) remain close to baseline values, showing that steering preserves the fluency of generated text. The steering metrics for mBERTopic and mETM are reported in Tables 12 and 14.

6 Limitations

MTMs trade the simplicity of word-based topic models for richer semantic representations, but this tradeoff introduces two practical constraints.

The first constraint is a dependency on SAEs and their ecosystem. Pretrained SAEs are currently available for only a handful of LLMs, and topic quality is sensitive to the choice of SAE layer, the dataset used to train the SAE, and the accuracy of automatically generated feature descriptions. Hence, if MTMs use an SAE that is not sufficiently relevant or specific for the dataset at hand, or if the labels for the features are inaccurate, the resulting topic quality may be low. We believe training or finetuning SAEs directly on target corpora, as in Movva et al. (2025), is a promising direction to mitigate these issues, as features learned from the data of interest are more likely to be relevant and their auto-generated labels more accurate.

The second constraint is a heavier preprocessing pipeline. Although the underlying algorithms (LDA, ETM, BERTopic) scale comparably to their word-based counterparts once given a feature matrix, the pipeline adds upfront featurization through an LLM and SAE, and feature filtering via another LLM (GPT-4o mini in our experiments, at a total cost of $0.33 for the eight datasets). Featurization requires a GPU with enough memory to fit the LLM and SAE. Depending on the dataset, this can represent a significant additional cost compared to simpler alternatives; Table 9 reports per-dataset throughput to provide an idea of additional runtime. While these additional demands are not prohibitive and need not preclude practitioners from using MTMs, they should be weighed when deciding whether richer topic descriptions, topic steering, and more abstract topics justify the overhead.

7 Conclusion

We introduced Mechanistic Topic Models (MTMs), a family of topic models that operate on SAE activation patterns rather than word counts or raw text embeddings. Our empirical evaluation shows MTMs are comparable with baselines on standard benchmarks, with significant advantages on abstract and short-text datasets as measured by topic judge. MTMs enable controlled text generation, allowing researchers to create new texts with specific topic compositions.

MTMs suggest how, despite some recent negative results, interpretability tools like SAEs can be successfully repurposed for downstream tasks.

Acknowledgments

We thank the reviewers and action editor for their thoughtful feedback, which has substantially improved the paper. David M. Blei was supported by NSF IIS-2127869, NSF DMS-2311108, ONR N000142412243, and the Simons Foundation.

References

Edward S. Adler and John D. Wilkerson. 2018. Congressional bills project: 1995-2018.

Dimo Angelov. 2020. Top2Vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 759–766.

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 European Chapter of the Association for Computational Linguistics*, pages 1676–1683.

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. Technical blog post.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

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

Ralph A. Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Technical blog post.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *Advances in Neural Information Processing Systems*, volume 37, pages 49519–49551.

Jonathan D. Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22, pages 288–296.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the Association for Computational Linguistics*, pages 15607–15631.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the International Conference on Machine Learning*, volume 235, pages 8359–8388.

Valérie Costa, Thomas Fel, Ekdeep S. Lubana, Bahareh Tolooshams, and Demba Ba. 2025. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neu-*

ral Information Processing Systems, volume 38, pages 34381–34424.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the International Conference on Learning Representations*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the Association for Computational Linguistics*, pages 4040–4054.

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.

Caitlin Doogan. 2022. *A Topic Is Not a Theme: Towards a Contextualised Approach to Topic Modelling*. Ph.D. thesis, Monash University, Clayton, Vic, Australia.

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. *arXiv preprint arXiv:2209.10652*.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are one-dimensionally linear. In *Proceedings of the International Conference on Learning Representations*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the Association for Computational Linguistics*, pages 889–898.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025a. Scaling and evaluating sparse autoencoders. In *Proceedings of the International Conference on Learning Representations*.

Mingqi Gao, Yixin Liu, Xinyu Hu, Xiaojun Wan, Jonathan Bragg, and Arman Cohan. 2025b. Re-evaluating automatic LLM system ranking for alignment with human preference. In *Findings of the Association for Computational Linguistics: North American Chapter of the Association for Computational Linguistics*, pages 4605–4629.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Os-

car Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, David Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Leander Girrbach and Zeynep Akata. 2026. Sparse autoencoders are topic models. In *Proceedings of the International Conference on Machine Learning*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

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

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In *Proceedings of the Conference on Language Modeling*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Alexander M. Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan L. Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033.

Alexander M. Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing*, pages 5321–5344.

Ken Lang. 1995. NewsWeeder: Learning to filter netnews. In *Proceedings of the International Conference on Machine Learning*, pages 331–339.

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 European Chapter of the Association for Computational Linguistics*, pages 530–539.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2025. Exploring the reliability of large language models as customized evaluators for diverse NLP tasks. In *Proceedings of the International Conference on Computational Linguistics*, pages 10325–10344.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300.

Johnny Lin. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement:

The role of pairwise preference in large language model evaluators. In *Proceedings of the Conference on Language Modeling*, pages 1–20.

Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In *Empirical Methods in Natural Language Processing*, pages 14878–14886.

Andrew K. McCallum. 2002. MALLET: A machine learning for language toolkit.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of the International Conference on Learning Representations*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 746–751.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. Sparse autoencoders for hypothesis generation. In *Proceedings of the International Conference on Machine Learning*, volume 267, pages 44997–45023.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 394–403.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangde. 2025. Steering language model refusal with sparse autoencoders. In *Proceedings of the Workshop on Reliable and Responsible Foundation Models*.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence.

OpenAI. 2025. Introducing GPT-4.1 in the API.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the International Conference on Machine Learning*, volume 235, pages 39643–39666.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. 2025. Automatically interpreting millions of features in large language models. In *Proceedings of the International Conference on Machine Learning*, volume 267, pages 48393–48421.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Antônio Pereira, Felipe Viegas, Diego Roberto Colombo Dias, Elisa Tuler, Ana Cláudia Machado, Guilherme Fonseca, Marcos André Gonçalves, and Leonardo Rocha. 2025. "Are the current topic modeling evaluation metrics enough?" Mitigating the limitations of topic modeling evaluation metrics using a multi-perspective game theoretic approach. *Knowledge-Based Systems*, 320:113634.

Chau Minh Pham, Alexander M. Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 2956–2984.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner.

2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the Association for Computational Linguistics*, pages 15504–15522.

Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2023. Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models. In *Proceedings of the BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 342–356.

Lewis Smith, Senthooran Rajamanoharan, Arthur Conmy, Callum McDougall, Tom Lieberum, János Kramár, Rohin Shah, and Neel Nanda. 2025. Negative results for SAEs on downstream tasks and deprioritising SAE research (GDM mech interp team progress update #2). Technical blog post.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Empirical Methods in Natural Language Processing*, pages 9348–9357.

Alex Tamkin, Mohammad Taufeeque, and Noah D. Goodman. 2024. Codebook features: Sparse and discrete interpretability for neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 235, pages 47535–47563.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, pages 139179–139212.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Christopher Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Technical blog post.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Felipe Viegas, Antonio Pereira, Washington Cunha, Celso França, Claudio Andrade, Elisa Tuler, Leonardo Rocha, and Marcos André Gonçalves. 2025. Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics. *Computational Linguistics*, 51(3):843–883.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free LLM truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the Association for Computing Machinery Web Conference*, pages 2562–2578.

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024. FASTopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *Advances in Neural Information Processing Systems*, volume 37, pages 84447–84481.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the International Conference on Machine Learning*, volume 267, pages 67035–67080.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: Contextualized embedding as a linear superposition of transformer factors. In *Proceedings of the Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? An empirical study on clustering with contextual embeddings for topics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 3886–3893.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A MTM Implementation Details

Section A.1 and Section A.2 detail the quality control measures used to address challenges in using SAE features in MTMs.

A.1 Preprocessing: Feature Filtering

Before training MTMs, we filter out SAE features that are likely spurious or irrelevant for topic modeling. We remove features in the following categories:

- Features with textual descriptions about programming, math, grammar, text formatting, or stop words.
- Features that activate in more than 1% of the SAE training data.
- Features that appear in more than 90% of corpus documents (analogous to removing high-frequency words in traditional topic modeling).

This filtering step is crucial. Without it, the resulting models do poorly in our benchmarks. To remove the features with irrelevant textual descriptions, we use the OpenAI GPT-4o mini model (OpenAI, 2024, version 2024-07-18).

A.2 Post-training: Topic Refinement

Even after preprocessing, SAE feature descriptions can occasionally be mislabeled. These mislabelings often become apparent when examining the full textual representation $\mathtt{t}_k$ of a topic (see Section 4.3). An example of this type of mislabeling is provided in Section I.

To address this, we apply an automated post-training refinement step. This step is optional and can easily be done in practice by inspection.

1. For each topic k, retrieve the top $n + m$ features by weight (where m is small, e.g., 2).
2. Prompt an LLM to identify and remove up to m features that appear irrelevant or mislabeled relative to the other features.
3. Retain the resulting top n features as the final topic representation.

Automated refinement is cost-effective since the number of topics and features per topic is typically small. The specific prompt used for this task is provided in Section F.1. We set $n = 10$ and $m = 2$ in our experiments.

To ensure a fair comparison, and avoid biasing the coherence metrics in favor of MTMs, we apply the exact same post-training refinement to all models in our experiments, including word-based baselines (see Section 5).

B Datasets

The dataset statistics are in Table 5. For the word-based models, the documents are preprocessed into word counts using the `soup-nuts` package (Hoyle et al., 2021) with the following settings: words are lowercased and named entities are automatically detected and merged together using spaCy (Honnibal et al., 2020) (e.g., "New York" becomes "New_York"); words must match the regex `^[\w-]*[a-zA-Z][\w-]*$`, contain at least 2 characters, and appear in less than 90% of documents. The vocabulary size is set to 5k for the smaller datasets (AGNews, GoEmotions, PoemSum) and 15k for the remaining datasets. After preprocessing, documents with less than 5 words are removed from the training corpus. We randomly subset to 10,000 training documents for AGNews, Yelp Polarity, and WritingPrompts.

Labels. Below is the label information for each dataset, along with some examples of labels.

20NG (20 categories): *talk.politics.guns, comp.graphics, misc.forsale*.

Bills (114 subtopics): *drug coverage and cost, water resources, insurance, postal service*.

Wiki (279 subcategories): *architecture buildings, 1990–1999 songs, fungi, warships of Russia*.

GoEmotions (28 annotated emotions): *anger, disappointment, optimism, neutral*.

C Baseline Implementations

We implement the baselines as follows. For LDA, we default to using the `Mallet` package (McCallum, 2002), which implements Gibbs sampling for inference. For the Wiki dataset, due to the large number of words/tokens, we instead use the LDA `scikit-learn` implementation (Pedregosa et al., 2011) with coordinate ascent variational inference. For the four neural models, we use the authors' provided code when it is available—we use the `bertopic` and `fastopic` PyPI packages, and for ETM and D-VAE, we use the authors' provided model code and re-implement the model training in our codebase using PyTorch Lightning.

For the baseline models, we use the default settings from either the paper or codebase (if unspecified in the paper), unless detailed as follows.

For LDA, we use the default settings in the `Mallet` package.

		Word-based			Mechanistic		
Corpus	#Docs	#Words	$\|V\|$	Avg. Len.	#Tokens	$\|V\|$	Avg. Len.
20NG	10,496	780,825	15k	74.4	3,023,135	8,750	288.0
Bills	32,659	3,442,488	15k	105.4	7,729,695	8,750	236.7
Wiki	14,290	14,037,490	15k	982.3	38,065,712	8,746	2663.8
GoEmotions	19,264	131,293	5k	6.8	767,694	8,750	39.9
PoemSum	2,398	184,695	5k	77.0	707,794	8,743	295.2
AGNews	10,000	162,414	5k	16.2	528,485	8,750	52.8
Yelp Polarity	10,000	465,280	15k	46.5	1,698,692	8,749	169.9
WritingPrompts	10,000	2,063,889	15k	206.4	6,924,757	8,749	692.5

Table 5: Training corpus statistics after preprocessing for the word-based and mechanistic topic models.

For ETM, we specifically use the Labeled ETM variant (i.e., we train skip-gram embeddings on the dataset and initialize the word embeddings with them, which are frozen during training), as we found that it outperformed regular ETM.

For BERTopic and FASTopic, we use the all-mpnet-base-v2 embedding model. For BERTopic, we set the top n words for c-TF-IDF to 25.

We allow ETM and D-VAE to use a separate validation set for early stopping. For mechanistic ETM, we instead form a validation set from 10% of the training set.

D Bayesian Optimization

We run Bayesian optimization for 25 iterations for each model-dataset pair. The first subsection lists the hyperparameter search space for each model. The next subsections provide mathematical formulations of the topic quality metrics used in the optimization objective. Following Dieng et al. (2020), we set the optimization objective to be the product of NPMI coherence and topic diversity.

D.1 Hyperparameter Ranges

Here, $[\![a, b]\!]$ denotes the set of integers from a to b.

D-VAE

`topic density` $\in [0.01, 5.0]$ (log-uniform)
`learning rate` $\in [10^{-3}, 10^{-1}]$ (log-uniform)
`n KL divergence warmup epochs` $\in [\![100, 200]\!]$
`use RSVI` $\in \{\texttt{true}, \texttt{false}\}$

FASTopic

`DT alpha` $\in [\![1, 25]\!]$
`TW alpha` $\in [\![1, 25]\!]$
`n epochs` $\in [\![100, 400]\!]$
`Sinkhorn threshold` $\in [10^{-7}, 0.05]$ (log-uniform)

LDA & mLDA

`topic density` $\in [0.01, 5.0]$ (log-uniform)
`word/feature density` $\in [0.01, 0.1]$ (log-uniform)

ETM

`learning rate` $\in [10^{-4}, 10^{-2}]$ (log-uniform)
`weight decay` $\in [10^{-7}, 10^{-5}]$ (log-uniform)
`use doc. completion validation` $\in \{\texttt{true}, \texttt{false}\}$

mETM

`learning rate` $\in [10^{-3}, 10^{-2}]$ (log-uniform)
`weight decay` $\in [10^{-7}, 10^{-5}]$ (log-uniform)
`dropout` $\in [0, 0.1]$
`decoder dropout` $\in [0, 0.1]$
`scheduler type` $\in \{\texttt{none}, \texttt{cosine}\}$

BERTopic

`n UMAP neighbors` $\in [\![5, 50]\!]$
`n UMAP components` $\in [\![5, 50]\!]$
`min topic size` $\in [\![5, 15]\!]$

mBERTopic

`n UMAP neighbors` $\in [\![5, 50]\!]$
`n UMAP components` $\in [\![5, 50]\!]$
`min topic size` $\in [\![5, 15]\!]$
`use SAE count embeddings` $\in \{\texttt{true}, \texttt{false}\}$
`use Sentence Transformer embs.` $\in \{\texttt{true}, \texttt{false}\}$[3]

D.2 NPMI Coherence

NPMI (Normalized Pointwise Mutual Information) coherence, or C_{NPMI}, measures the semantic relatedness of the top words within each topic based on their co-occurrence patterns in a reference corpus. For a given topic, NPMI coherence is defined as

$$C_{\text{NPMI}} = \frac{1}{\binom{10}{2}} \sum_{j=2}^{10} \sum_{i=1}^{j-1} \frac{\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-\log p(w_i, w_j)}, \quad (21)$$

where $\{w_i\}_{i=1}^{10}$ are the top 10 words in a topic, and $p(w_i, w_j)$ is estimated using word co-occurrences in a sliding context window across documents (Lau et al., 2014). In our experiments, we set the reference corpus to be the training corpus, and the entire document is used as the context

[3]This option was always chosen to be false.

window. The probabilities $p(w_i)$ and $p(w_j)$ represent the individual word frequencies in the corpus.

D.3 Topic Diversity

Topic diversity measures how distinct topics are from each other by computing the fraction of unique features among the top features across all topics. Formally, topic diversity is calculated as

$$\mathrm{TD} = \frac{\left|\bigcup_{k=1}^{K} \mathrm{TopFeatures}_k^{(n)}\right|}{Kn}.$$

Let $\mathrm{TopFeatures}_k^{(n)}$ denote the n highest-weight features of topic k, where features are words for word-based models and SAE features for mechanistic models. Following Dieng et al. (2020), we use $n = 25$.

D.4 Limitations for Cross-Model Comparison

While topic diversity is roughly comparable across mechanistic and word-based topic models, NPMI coherence is not directly comparable between these model types due to their differing vocabulary distributions. Mechanistic models operate on SAE feature spaces, while word-based models use traditional word vocabularies, making direct coherence comparisons problematic.

E Steering Experiment Details

E.1 Steering Intervention

For all steering, we perform the topic ablation and subsequent topic addition interventions across all layers and all token positions using a pre-hook on the forward activations.

E.2 Topic Relevance Win Rate

We run $R = 10$ trials per topic with temperature set to 0.3 for both the steered and unsteered models. We generate a maximum of 50 tokens. We use GPT-4o mini (OpenAI, 2024) as the LLM judge, and the prompt is provided in Section F.4.

E.3 Topic Likelihood Difference

We select the most relevant documents for each topic using a threshold-based approach. For threshold $\tau = 0.5$, we define $\widetilde{\mathcal{D}}_k = \{d : \boldsymbol{\theta}_{d,k} \geq \tau\}$ and set

$$\mathcal{D}_k = \begin{cases} \widetilde{\mathcal{D}}_k, & 3 \leq |\widetilde{\mathcal{D}}_k| \leq 10 \\ \mathrm{Top}_{10}(\widetilde{\mathcal{D}}_k), & |\widetilde{\mathcal{D}}_k| > 10 \\ \mathrm{Top}_3(\{1, \ldots, \mathcal{D}\}). & |\widetilde{\mathcal{D}}_k| < 3 \end{cases}$$

This ensures 3–10 documents per topic. We use all threshold-exceeding documents when feasible, capping at 10 for topics with many relevant documents. When fewer than 3 documents meet the threshold, we select the top 3 from the corpus.

F Prompts

We provide the prompt templates used for post-training topic refinement (Section A.2), topic summarization (Section 4.3), and the LLM-based evaluations in Section 5. For the topic judge prompt and examples, see Section G.

The prompted LLM is GPT-4.1 (OpenAI, 2025, version 2025-04-14), except for topic relevance win rate (see Section E.2).

F.1 Topic Refinement

System prompt. The goal of this task is to evaluate a list of features produced by an automatic method. We call this list a "topic". Given a topic, you'll be answering the question: "Which [*word* | *feature*](s) don't belong in this list?" For each topic, choose the [*word* | *feature*](s) whose meaning does not match with what the list seems to be about.

Here is an example: [example]

Here is another example: [example]

Reply with a brief reasoning for your choice, and up to two numbers corresponding to the [*word* | *feature*]s that don't belong (or -1 if there are less than two).

Important: Prioritize identifying [*word* | *feature*](s) that are oddly specific and/or clearly out of place. Use your world knowledge in considering whether a [*word* | *feature*] belongs. If you are not sure, do not choose the [*word* | *feature*].

User prompt. Topic: [topic]

F.2 Topic Summarization

System prompt. You are a helpful assistant specializing in topic summarization. You will be given a list of either topic keywords (some of them may be several words concatenated with "_") or descriptions of text generated by an automated method. Your task is to summarize these keywords or descriptions into no more than 1 sentence describing the topic's central theme.

Provide a concise and informative summary that captures the topic's essence. Here are some specific guidelines:

1. Do not use a full sentence or a complete thought.
2. Use your world knowledge to help you decide what the topic is about.
3. The summary should be general, capturing the commonalities of the items as a single main theme. In particular, do not rely on lists in your response, or include specifics that only pertain to a few items in the topic.
4. If unsure, err on the side of being more general rather than too specific in your summary.

User prompt. Topic: [topic]

F.3 Ratings and Intrusion

System prompt. The goal of this task is to evaluate a list of [*word* | *feature*]s produced by an automatic method. We call this list a "topic". Given a topic, you will determine how related its [*word* | *feature*]s are on a 3-point scale. The rating options are: (1) Not Very Related, (2) Somewhat Related, (3) Very Related. A helpful question to ask yourself is: "What is this group of [*word* | *feature*]s about?" If you can answer easily, then the [*word* | *feature*]s are probably related. Use your world knowledge and the context provided by the other [*word* | *feature*]s to help determine your rating. Here is some guidance and examples on how to apply these ratings.

Very Related - Most of the [*word* | *feature*]s are clearly related to each other, and it would be easy to describe how they are related.

Example 1: [example]

Example 2: [example]

Somewhat Related - The [*word* | *feature*]s are loosely related to each other, but there may be a few that are ambiguous, generic, or unrelated.

Example 1: [example]

Example 2: [example]

Not Very Related - The [*word* | *feature*]s do not share any obvious relationship to each other. It would be difficult to describe how the [*word* | *feature*]s are related to each other.

Example: [example]

Reply with a brief reasoning for your choice and a single number, indicating the overall relatedness of the [*word* | *feature*]s in that topic.
User prompt. Topic: [topic]

System prompt. The goal of this task is to evaluate a list of [*word* | *feature*]s produced by an automatic method. We call this list a "topic". Given a topic, you'll be answering the question: "Which [*word* | *feature*] doesn't belong in this list?" For each topic, choose the [*word* | *feature*] with the meaning or usage that is most different from the others. If you feel that multiple [*word* | *feature*]s do not belong, choose the one that you feel is most out of place.
Here are some examples: [example]
Here is another, harder example: [example]
You might be given multiple topics. For each topic, reply with a brief reasoning for your choice and the number of the [*word* | *feature*] that doesn't belong.
User prompt. Topic: [topic]

F.4 Topic Relevance Win Rate (TWR)

System prompt. You are an expert evaluator of text relevance to topics. You will be given a topic summary and a list of text samples. Your task is to determine which text sample is most relevant to the given topic.
User prompt. Topic summary: [topic summary]
Text samples: [texts]
Which text sample (by index number) is most relevant to the topic? Provide the index (starting from 0) and a brief explanation.

G Topic Judge

In our experiments, we set $q = 2$, $p = 0.75$, and $T = 100$. For TopFeatures, we always take the top 10 words for baselines, and take either the top 10 (for documents with 1 topic) or top 5 (for documents with 2 topics) features for MTMs.

We provide the prompt and example inputs and responses for topic judge below.

G.1 Prompt

System prompt. In this task, you will be presented with a document, a criteria, and two sets of "topics". [*A given topic is a list of either single words (or occasionally, instead of a single word, several words concatenated with "_") or descriptions of text about 5-15 words long.* | *A given topic will be shown as a summary description no more than 1 sentence long.*] Each set of topics includes 1-2 topics total. The task is to choose which of the two sets of topics is better suited to the document based on the provided criteria. Reply with "A" if the first set of topics is better or "B" if the second set of topics is better. If you think that the two sets of topics are equally good, please reply with "tie". Only use "tie" if the two sets of topics are very similar and you cannot choose one over the other. Before making your choice, provide a brief reasoning for your decision.
User prompt. Document: [document]
Criteria: Consider how well each topic captures the general meaning of the document. Consider all types of meaning, including the text's subject matter and the text's affect, emotive content, and style. If Set of Topics A presents a better overall summary of the document compared to Set of Topics B, then "A" should be chosen, and vice versa.
Set of Topics A:
Topic: [topic]
Set of Topics B:
Topic: [topic]

G.2 Examples

Example 1: Bills with Summarization
Document:
Amends part A of title I of the Elementary and Secondary Education Act of 1965 to alter the percentage of school improvement funds that are reserved for outlying areas and Indian education. Reserves 0.5% of those funds for outlying areas and 0.75% for Indian education. (Currently, 1% of those funds are reserved for outlying areas and Indian education.)
Set of Topics A:
- Educational policies, administration, and support strategies
- Financial management, funding, and policy in public and organizational contexts
Set of Topics B:
- Primary and secondary education systems and funding
- Geographical regions and territories in the United States
Choice: B
Reasoning: Set B is more specific to the content of the document, which discusses changes to funding allocations within the Elementary and Secondary Education Act, specifically for outlying areas and Indian education. 'Primary and secondary education systems and funding' directly addresses the subject matter, and 'Geographical regions and territories in the United States' captures the focus on outlying areas and Indian education. Set A is broader and less directly tied to the specifics of the document.
Example 2: PoemSum with TopFeatures
Document:
I saw in Louisiana a live-oak growing,
All alone stood it and the moss hung down from the branches,
Without any companion it grew there uttering joyous leaves of dark green,
And its look, rude, unbending, lusty, made me think of myself,
But I wonder'd how it could utter joyous leaves standing alone there without its friend near, for I knew I could not,
And I broke off a twig with a certain number of leaves upon it, and twined around it a little moss,
And brought it away, and I have placed it in sight in my room,
It is not needed to remind me as of my own dear friends,
(For I believe lately I think of little else than of them,)
Yet it remains to me a curious token, it makes me think of manly love;
For all that, and though the live-oak glistens there in Louisiana solitary in a wide flat space,
Uttering joyous leaves all its life without a friend a lover near,
I know very well I could not.
Set of Topics A:
Topic: "tree", "plants", "leaves", "trees", "green", "earth", "boughs", "leaf", "life", "growth"
Topic: "spring", "summer", "autumn", "winter", "fields", "heart", "long", "year", "beauty", "snow"
Set of Topics B:
Topic:
- "literary elements and themes related to poetry"
- "specific types of flowering plants and their characteristics"
- "descriptions and characteristics of autumn foliage"
- "descriptions of agricultural practices and related processes"
- "references to specific plants and their biochemical properties"
Topic:
- "emotional connections and feelings of companionship"
- "themes of longing and escape"
- "instances of humor or comedic elements in the text"
- "sentences that imply existential questions or dilemmas regarding trust and character integrity"
- "expressions related to doubt or uncertainty"
Choice: B
Reasoning: Set A focuses on nature, trees, and the cycle of seasons, which are present in the poem, but it misses the deeper emotional and thematic content about solitude, companionship, and longing. Set B, while a bit broad and including some irrelevant elements (like humor and agricultural practices), does include topics about emotional connections, companionship, and existential questions, which are central to the poem's meaning. Despite some extraneous topics, Set B better captures the poem's affective and thematic depth.

H Human Evaluation Study

Figure 5 shows the evaluation interface used in the human topic judge study described in Section 5.1. Below we provide the instructions shown to participants.

H.1 Instructions Shown to Participants

This survey asks you to rank the outputs of computer programs – called "topics" – that attempt to summarize a given document. In different questions, you will see the topics in one of two formats, "Top Features" or "Summary". "Top Features" topics are lists of either related words or short descriptions. "Summary" topics have summarized the list of words/descriptions into a single sentence.

Your task is to rank the options (computer program outputs) based on how well they describe the provided document. Each option is a "topic set", which will have 1–2 topics. Your criteria for judging each topic set is the following:

Consider how well each topic captures the general meaning of the document. Consider all types of meaning, including the text's subject matter and the text's affect, emotive content, and style.

Please read as much of the document as necessary to confidently determine your ranking.

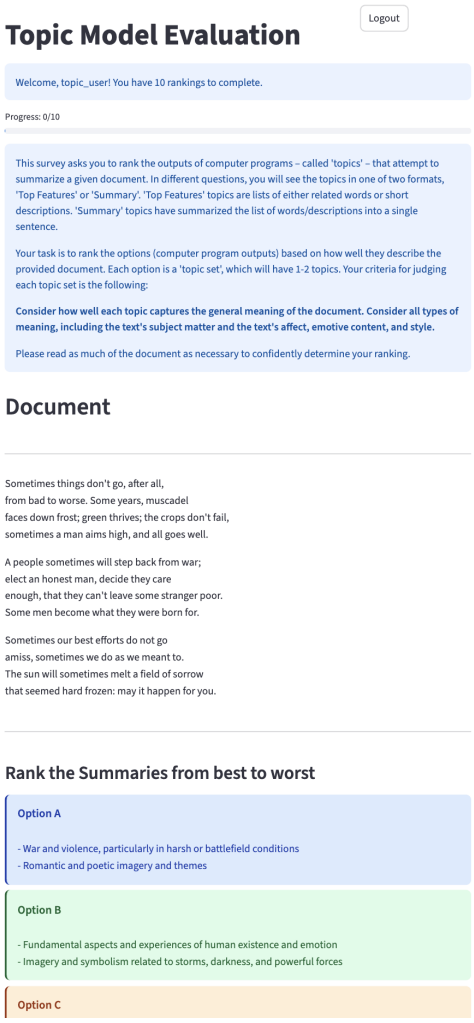

Figure 5: Screenshot of the evaluation interface shown to participants. Each task presents a document alongside topic set options produced by different models. Participants rank the options from best to worst based on how well each topic set captures the document's meaning.

I Topic Refinement Example

– *references to Azerbaijani cultural elements, particularly music and instruments*
– references to protests and related incidents
– events or actions involving protests and their consequences
– topics related to the Holocaust and atrocities committed during wartime
– references to combat and military actions
– topics related to military actions and warfare
– references to military involvement or actions related to Russia
– keywords related to violence and its victims
– references to violent or aggressive actions
– instances of violence or conflict
– *phrases related to archaeological discoveries and remains*
– mentions of military divisions or actions

Figure 6: The features identified as spurious in topic refinement for this topic are italicized. The documents associated with this topic discuss the Armenian genocide. The two identified feature descriptions have an incorrect focus (music and instruments) or context (archaeological discoveries).

J Topic Alignment Procedure for Cross-Correlation Heatmaps

To create the cross-correlation heatmaps in Figure 3, we use greedy alignment:

1. **Compute correlations:** For models m and m' with K topics each, compute the $K \times K$ correlation matrix $\mathbf{C}$ where $C_{i,j} = \text{corr}(\theta^{m}_{:,i}, \theta^{m'}_{:,j})$.
2. **Greedy matching:** Starting with all topics unmatched, iteratively:
   - Find the column j^* with highest maximum across all rows: $j^* = \arg\max_{j \in \text{remaining}} \max_i C_{i,j}$
   - Among remaining rows, find the best match for this column: $i^* = \arg\max_{i \in \text{remaining}} C_{i,j^*}$
   - Add pair (i^*, j^*) to alignment and remove from consideration
3. **Visualize:** Reorder topics by alignment order to create the heatmap, with strongest column matches appearing first along the diagonal.

High correlations along the diagonal indicate shared concepts between models, while low-correlation regions (dashed boxes) reveal novel topics unique to one model.

Model: mETM | **Dataset:** Bills **Topic Summary:** "Legal terminology and concepts surrounding property rights, ownership, environmental conservation, and financial transactions in real estate and land management."	
$\lambda = 10$	"A text about the history of a house in New York City, with an emphasis on its architecture and interior design. The author explores how the house has changed over time, from its original construction to its current state [...]"
$\lambda = 20$	"A text about the history of a 10-acre plot in San Bernardino County, California. The property is located near the town of Twentynine Palms and was owned by John W. Smith from 1924 until his death in 1"
$\lambda = 30$	"A text about the 1980s of a parcel in the Town of Silverton, Colorado. The property is located at approximately 35°24'N and 67°W on the west side of Main Street between E.S"
$\lambda = 40$	"A text about 1.02 acres in the town of East Helena, Montana is for sale by sealed bid No. 49-3758 (the "Property") located at approximately 6 miles north and west from the city limits of East"
$\lambda = 50$	"A text about 10 acres in the City of San Bernardino, County of Maricoa.\n\nThe property is located on a road that was granted by the State of Arizona and United States for the purpose of which is to be used as an irrigation [...]"
Model: mBERTopic | **Dataset:** PoemSum **Topic Summary:** "Exploration of agriculture, plant varieties, propagation techniques, and their interactions with environmental factors and literary themes."	
$\lambda = 10$	"A text about the history of a house, its inhabitants and their memories.\n\nThe book is divided into three parts: 1) The story of the house; 2) A series of short stories by different authors from around the world that are inspired [...]"
$\lambda = 20$	"A text about a woman who is in love with her best friend.\n\nI've been thinking of you all day,\nAnd I can't help but wonder if you feel the same way?\nYou are my best friend and I know that we will"
$\lambda = 30$	"A text about a girl who was born on the 4th of July, and grew up in the shadow of her mother's love. Her father had died when she was young, but he left behind him a memory that never faded from her heart; for"
$\lambda = 40$	"A text about the beauty of a flower,\nThe fragrance that it gives to air;\nAnd how its petals fall in showers,\nLike stars from heaven's starry hair.\n\nBut soon they fade and die away,\nAs if their hearts were broken"
$\lambda = 50$	"A text about the tree is a beautiful sight to behold.\n\nThe branches of this tree are full of fruit, and it'ens with blossoms that bloom in the spring. The leaves of this plant are green and red, and they grow on the ground. And"
Unsteered examples (control)	
"A text about a new study on the role of women in science, technology and innovation (STI) was published by UN Women.\n\nThe report "Women's participation in STI: A review of evidence" is based on an analysis of 41 studies from"	
"A text about the new *Star Wars* movie, which is set to be released in December of this year.\n\n<blockquote>The film will follow a group of young heroes and villains who are on the run from an evil empire that has taken over their"	
"A text about the 1970s, and how it was a time of great change for women.\n\nThe author talks about her own experience as a young woman in the 70's, when she had to fight against sexism and discrimination."	

Table 6: Representative examples of steered and unsteered text generations for different λ values.

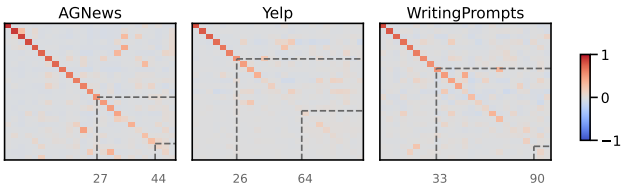

Figure 7: Topic cross-correlation heatmaps (mLDA vs. LDA) for the three additional datasets. Conventions as in Figure 3.

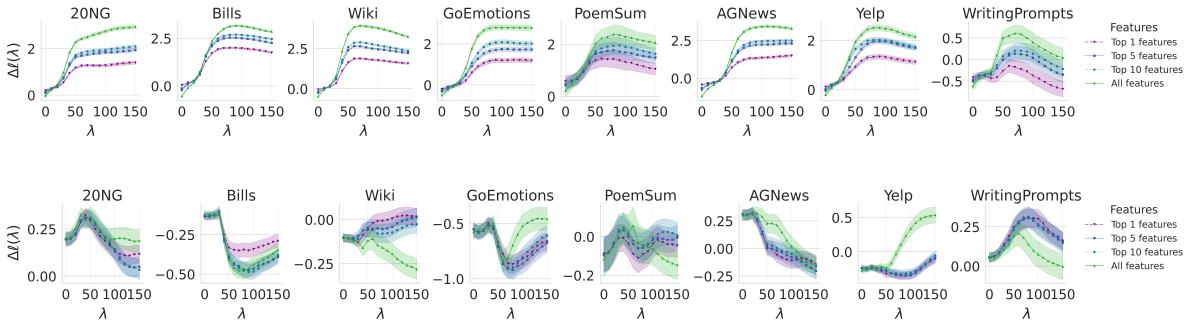

Figure 8: Document log-likelihood difference $\Delta\ell(\lambda)$ for mETM (above) and mBERTopic (below). For mETM, $\Delta\ell$ generally increases as λ increases, confirming bias toward the target topic. In contrast, mBERTopic does not always show positive $\Delta\ell$ shifts. We think this may be because it is the only model that uses class-based TF-IDF to learn its topic-vocabulary weights, but further investigation is needed. Nevertheless, mBERTopic performs comparably to the other MTMs on metrics measuring the quality of its generated text (Section 5.4), indicating that its steering vectors are still generally effective.

	20NG	Bills	Wiki	Yelp	AGN.	GoE.	Poem.	WriP.
D-VAE	.90	.90	.86	.92	.98	.96	.82	.81
FASTopic	**.99**	**1.0**	**.98**	**.98**	**1.0**	**1.0**	**.99**	**.96**
LDA	.61	.43	.63	.75	.61	.54	.47	.58
ETM	.62	.55	.71	.60	.81	.79	.71	.61
BERTopic	.75	.69	.65	.72	.80	.78	.63	.60
mLDA	.71	.53	.68	.53	.69	.83	.64	.58
mETM	.65	.50	.61	.54	.65	.81	.62	.61
mBERTopic	.64	.58	.60	.40	.72	.85	.38	.36

Table 7: Topic diversity across all datasets, averaged over five runs. Conventions as in Table 2.

	Control	$\lambda = 10$	$\lambda = 20$	$\lambda = 30$	$\lambda = 40$	$\lambda = 50$
mLDA	8.2	2.8	8.4	17.9	28.0	34.5
mETM	10.6	3.6	8.8	16.3	27.7	33.0
mBERTopic	4.8	2.4	7.4	18.2	31.5	35.8

Table 10: Distribution of TWR wins (%) by steering magnitude, pooled across all datasets. Each row sums to 100%. Higher magnitudes account for the majority of steered wins across all models.

	Normal		Hard	
Model	$\mathbf{B}$	$\mathbf{\Theta}$	$\mathbf{B}$	$\mathbf{\Theta}$
D-VAE	0.76	0.85	0.89	0.93
FASTopic	0.70	0.79	0.83	0.86
LDA	0.45	0.74	0.51	0.77
ETM	0.53	0.73	0.64	0.79
BERTopic	0.33	0.72	0.43	0.68
MTM (w/ LDA)	**0.29**	**0.48**	**0.32**	**0.49**
MTM (w/ ETM)	0.35	0.55	0.40	0.54
MTM (w/ BERTopic)	0.30	0.74	0.35	0.69

Table 11: Averaged model stability for normal (20NG, Bills, Wiki, Yelp, AGN) and hard (GoEmo., PoemSum, WriPro.) datasets. Each model is trained 5 times with different random seeds. Distance between topic pairs is defined as the Rank-Biased Overlap distance ($p = 0.9$) of the top 500 features ($\mathbf{B}$) or documents ($\mathbf{\Theta}$); topics are then matched with the Hungarian algorithm (Hoyle et al., 2022). MTMs are overall more stable than their word-based counterparts.

	20NG		Bills		Wiki		GoEmotions	
Model	P_1	NMI	P_1	NMI	P_1	NMI	P_1	NMI
D-VAE	0.27	0.33	0.33	0.38	0.45	0.71	0.28	0.05
FASTopic	0.58	0.47	0.44	0.48	0.41	0.65	0.29	0.06
LDA	0.54	0.44	0.50	0.51	0.44	**0.72**	0.28	0.05
ETM	0.54	0.45	0.46	0.50	0.36	0.69	0.30	0.05
BERTopic	**0.65**	**0.59**	0.49	**0.56**	0.42	0.69	0.27	0.04
MTM (w/ LDA)	0.49	0.42	**0.52**	0.53	0.42	0.71	**0.31**	**0.08**
MTM (w/ ETM)	0.53	0.44	0.51	0.53	**0.45**	0.71	0.31	0.08
MTM (w/ BERTopic)	0.56	0.51	0.47	0.53	0.37	0.68	0.28	0.07

Table 13: Alignment metrics Purity (P_1) and NMI for labeled datasets, averaged over five runs. Conventions as in Table 2. Purity quantifies single-category clusters; NMI measures topic-label mutual information (Hoyle et al., 2022). MTMs outperform word-based counterparts on GoEmotions, consisting of short texts labeled with one of 28 emotions.

	Wiki		PoemSum	
Model	TF	Sum.	TF	Sum.
D-VAE	1383	1513	1361	1376
FASTopic	1387	1476	1391	1424
LDA	1529	1574	1525	1546
ETM	1550	1492	1471	1401
BERTopic	1338	1355	1487	**1592**
MTM (w/ LDA)	**1657**	1574	1619	1550
MTM (w/ ETM)	1654	**1585**	**1627**	1561
MTM (w/ BERTopic)	1502	1430	1519	1550

Table 8: Human topic judge Elo scores on Wiki and PoemSum. Elo conventions as in Table 1. Human–LLM agreement is visualized in Figure 2.

20NG	Bills	Wiki	GoEmo.	Poem.	AGN.	Yelp	WriPro.
59.5	80.1	7.6	735.0	64.8	355.3	118.3	30.0

Table 9: Featurization throughput (docs/s) per dataset. We used Gemma 2-9B on an NVIDIA H200 with a batch size of 64 with documents sorted by length. Computations were done on bfloat16 and we only passed the document through the first 10 layers of the LLM as this is the layer used in our experiments.

	20NG	Bills	Wiki	Yelp	AGN	GoEmo	Poem	WP
TWR (%)	93.9	99.2	96.5	98.1	99.2	93.7	97.0	85.3
$\text{PPL}_{\text{control}}$	6.25	6.21	6.22	6.21	6.19	6.16	6.24	6.24
$\text{PPL}_{\lambda=10}$	6.41	6.34	6.53	6.50	6.13	6.25	6.23	6.22
$\text{PPL}_{\lambda=20}$	7.00	6.86	6.91	7.39	6.73	6.76	7.10	7.00
$\text{PPL}_{\lambda=30}$	8.10	8.02	7.76	8.96	7.42	7.81	8.93	8.29
$\text{PPL}_{\lambda=40}$	10.02	9.75	9.24	11.11	8.47	10.46	11.46	11.25
$\text{PPL}_{\lambda=50}$	15.01	12.73	12.43	15.37	11.30	23.52	16.48	25.34

Table 12: Steering metrics for mBERTopic across all 8 datasets. Conventions as in Table 4.

	20NG	Bills	Wiki	Yelp	AGN	GoEmo	Poem	WP
TWR (%)	88.1	99.2	93.1	88.7	93.0	76.0	85.8	84.9
$\text{PPL}_{\text{control}}$	6.20	6.20	6.25	6.21	6.27	6.20	6.22	6.21
$\text{PPL}_{\lambda=10}$	6.41	6.17	6.41	6.44	6.18	6.27	6.31	6.30
$\text{PPL}_{\lambda=20}$	6.83	6.85	7.03	7.19	6.66	6.75	7.05	6.84
$\text{PPL}_{\lambda=30}$	7.93	8.02	7.74	8.74	7.70	7.80	8.66	8.10
$\text{PPL}_{\lambda=40}$	10.04	9.66	9.26	11.11	8.72	10.83	11.51	10.81
$\text{PPL}_{\lambda=50}$	14.40	12.87	12.88	16.93	11.83	21.06	17.33	17.02

Table 14: Steering metrics for mETM across all 8 datasets. Conventions as in Table 4.

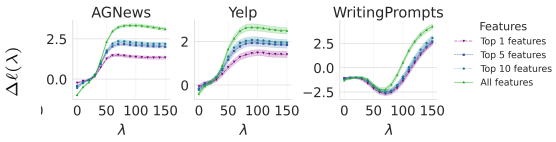

Figure 9: Document log-likelihood difference $\Delta\ell(\lambda)$ for mLDA on AGNews, Yelp, and WritingPrompts. As in Figure 4, $\Delta\ell$ generally becomes more positive as steering strengthens, with the full steering vector producing the largest shift.

species, shark, cap, stem, sharks, brown, fungus, shaped, smooth, fruit, spores, fins, bodies, surface, genus, typically, thick, distribution, habitat, fin

- scientific classifications and descriptions of plant species
- terms related to flora and plant characteristics
- details about plant characteristics and descriptions
- terms related to cryptic species and their ecological studies
- specific colors and descriptions related to plants and their characteristics
- terms related to biological and anatomical features

(a) [Wiki] mETM uncovers high-level botanical taxonomy concepts, robust to varied word choice.

people, law, government, right, rights, laws, like, case, think, state, public, crime, use, time, person

- negative sentiments directed towards authority and governance
- references to freedom of speech and expression
- references to extremist ideologies and discriminatory language
- names of political figures and references to political actions
- themes related to political criticism and party dynamics
- phrases related to political accountability and ethics

(b) [20NG] mLDA captures both critical tone and content in political discourse, whereas LDA captures only the broader political theme.

- expressions of disbelief or surprise
- expressions of unexpectedness or surprise in various contexts
- words expressing strangeness or oddity
- expressions of emotions and connections between people
- expressions of doubt, reflection, and introspection
- expressions of surprise or shock in personal experiences

(c) [GoEmotions] mLDA identifies complex emotional states (surprise, disbelief) as a coherent topic.

- references to relationship dynamics and communication issues
- references to workplace disputes and grievances
- familial relationships and conflicts
- references to psychological distress and coping mechanisms
- phrases and terms related to emotional support and effective communication in caregiving
- emotions related to anger and frustration

(d) [20NG] mBERTopic reveals nuanced interpersonal conflict and coping themes.

- expressions related to rudeness and offensive behavior
- expressions of anger and strong negative emotions
- terms associated with negative character traits or behavior
- expressions of criticism and critical comments
- themes related to feelings of embarrassment and shame
- references to feedback, accountability, and constructive criticism

(e) [GoEmotions] mLDA captures sentiment and tone in a unified topic, identifying negative sentiment and social disapproval themes.

government, political, minister, party, military, prime, president, national, economic, leader, country, opposition, policy, parliament, foreign, secretary, leadership, independence, general, leaders

- references to historical or political movements and conflicts
- references to political dynamics, power struggles, and social discrimination
- references to specific groups or organizations
- references to U.S.-backed interventions and coups in foreign nations
- references to political factions or militia groups involved in conflicts
- references to historical events involving the Soviet Union

(f) [Wiki] mETM provides rich contextual descriptions of historical conflicts versus ETM's generic political terms.

- references to characters or individuals from Greek mythology
- nouns associated with historical events and figures
- instances of significant literary expressions or metaphors
- mythological figures and their related narratives
- religious or spiritual references and concepts
- references to mythological gods and their interactions

(g) [PoemSum] mETM discovers novel Greek mythology theme (max correlation 0.12 with any ETM topic).

beauty, woman, greece, maid, thy, praise, eyes, white, lovely, thee, world, skill, face, darkened, love, wan, fleet, deeds, lip, art

- concepts related to existential purpose and the divine
- themes related to existential questions and the nature of God
- instances of significant literary expressions or metaphors
- literary elements and themes related to poetry
- concepts related to spirituality and eternal life
- descriptive phrases related to physical characteristics or appearances

(h) [PoemSum] mBERTopic captures existential and religious motifs beyond surface-level poetic vocabulary.

- literary elements and themes related to poetry
- information about an artist's biography and professional background
- emotional connections and feelings of companionship
- references to poetry and poets
- instances of humor or comedic elements in the text
- sentences that imply existential questions or dilemmas regarding trust and character integrity
- keywords or phrases related to judicial or legal processes
- mentions of struggles and hardships faced by individuals

(i) [PoemSum] mLDA demonstrates a potential MTM failure mode: while the features correctly represent a poetic theme, they're too broad to be useful, and spurious features irrelevant to the topic's top documents persist despite post-training refinement (Section A.2).

Figure 10: Illustrative examples of topic sets from MTMs (mLDA, mETM, mBERTopic) across diverse datasets. Each mechanistic topic is represented by its top 6–8 SAE features. When appropriate, mechanistic topics are contrasted with their most similar word-based topic (top 20 words), selected based on correlation of document-topic distributions (see Section 5.3). The word-based topic is shown above the corresponding mechanistic topic.