

Beyond Accuracy: Interpreting Topic Representation in Suicide Ideation Detection Models

Hamideh Ghanadian¹, Isar Nejadgholi², Hussein Al Osman³

^{1,2,3}University of Ottawa, Ottawa, Canada

²National Research Council Canada, Ottawa, Canada

^{1,3}{Hghan053, Hussein.alosman}@uottawa.ca

²isar.nejadgholi@nrc-cnrc.gc.ca

Abstract

Suicide ideation detection models are typically evaluated using aggregate performance metrics, yet little is known about how they internally represent psychologically meaningful risk factors. In high-stakes mental health applications, understanding these internal representations is essential for safety, transparency, and responsible deployment. In this work, we move beyond accuracy and analyze how suicide detection models trained on original and topic-augmented datasets encode psychological risk factors in their internal representation space. Using visualization and geometric analysis, we examine the coherence and separability of topic-related features. Our results show that topic-aware augmentation increases the clarity and distinctness of underrepresented psychosocial risk factors such as immigration, family issues, and financial crisis. These findings suggest that augmentation not only improves model performance but also leads to more structured and interpretable internal representations.

1 Introduction

Despite strong performance in mental health NLP, leading models remain largely uninterpretable. In high-stakes applications such as suicide ideation detection, interpretability is essential for identifying spurious correlations, auditing learned representations, and ensuring safe and accountable deployment (Rudin, 2019).

Studies show that some Generative Large Language Models (GLLMs) such as GPT models have superior performance in classifying suicidality of the text data compared to other examined GLLMs (Ghanadian et al., 2023). However, they still fall short when compared to classifiers tailored for this task through supervised training. Therefore, in this research, we focus on investigating the interpretability of an existing fine-tuned classifiers.

Some research showed that existing datasets lack topic diversity, and some topics are over-represented in datasets while others are absent or underrepresented

(Ghanadian et al., 2024). They demonstrated that classifiers may develop a false causal relationship between an over-represented concept and a given label, leading to excessive dependence on that concept and potentially compromising classification accuracy (Nejadgholi et al., 2023). Also, some classifiers might become under-sensitive to topics that are not well-represented in training sets (Nejadgholi et al., 2022). Therefore, it is crucial to develop methods that can facilitate the study of how topics are formed and used in trained classifiers.

In suicidal ideation detection, certain risk factors such as depression, anxiety, relationship problems, and hopelessness are strong predictors of suicidal ideation and appear frequently in training corpora (Ghanadian et al., 2024). These concepts are often over-represented in the suicidal class, making them susceptible to being learned as sufficient causes for suicidality (Ghanadian et al., 2025). Consequently, the classifier may over-rely on these concepts and ignore the broader context, leading to reduced generalizability. Conversely, topics that appear rarely in the dataset such as racism and immigration may be under-represented in the model’s internal features. Therefore, it is necessary to understand how under-representation of topics translates into feature coherence in the latent space. This insight is necessary for diagnosing gaps in topic coverage. These observations suggest that suicidal-ideation classifiers may not always form clear internal representations of psychological risk factors. Some concepts might appear as coherent and separable directions, while others may remain entangled with unrelated features.

Our two research questions regarding the interpretability of the suicidal detection models are as follows:

- **RQ1: How are psychological risk factors of suicidal ideation represented within the internal activation space of language models trained for suicide ideation detection?**
- **RQ2: Does topic-aware synthetic data augmentation improve the clarity, separability, and specialization of risk-factor representations in the model’s internal space?**

The main contributions of this study are as follows:

- We provide a mechanistic analysis of suicide detection models by examining their internal activation space rather than relying only on performance metrics. Using overcomplete sparse autoencoders, we extract interpretable feature directions and analyze how psychological risk factors are encoded in the model’s latent geometry.
- We investigate feature specialization and monosemanticity in learned sparse representations. Through analysis of top-activating samples and UMAP visualizations, we show that certain features align with specific psychosocial topics and form coherent clusters in representation space.
- We introduce a geometric separability framework to measure how clearly psychological concepts are represented. Using cosine distance between concept and non-concept means, we quantify representational clarity and demonstrate how topic-aware augmentation reshapes internal feature structure.

2 Background and Related Work

A common working assumption in mechanistic interpretability is the linear representation hypothesis, which states that meaningful features often approximate linear directions in a model’s activation space. Here, the activation space refers to the vector space formed by the model’s internal activations at a specific layer, where each vector corresponds to the model’s internal state for a given token or input. Under this view, probing the geometry of activations can reveal human-describable concepts that the model uses internally (Olah et al., 2020).

Modern transformer models achieve strong performance by representing many latent features within fixed-width activation spaces. Empirical evidence suggests that models encode far more features than there are embedding dimensions (Elhage et al., 2023). The apparent mismatch between the number of useful features and the dimensionality of the activation space motivates the superposition hypothesis. The superposition hypothesis states high-dimensional spaces contain many nearly orthogonal directions, allowing networks to store multiple features in overlapping subspaces (Bricken et al., 2023). When the number of useful features exceeds the available dimensions, features become entangled, producing polysemantic representations that respond to multiple concepts (Elhage et al., 2022).

In this context, a feature is defined as a direction in activation space that corresponds to a recurring internal pattern used by the model. A feature is called monosemantic if it consistently corresponds to a single interpretable concept across inputs; otherwise, it is polysemantic. Polysemanticity arises naturally under superposition and can obscure the internal logic behind model predictions. For high-stakes applications

such as suicide ideation detection, understanding whether psychological risk factors are represented in monosemantic or polysemantic forms is crucial for safety, transparency, and targeted intervention (Zhang et al., 2024).

Sparse autoencoders (SAEs) provide a principled way to uncover feature directions in a model’s activation space. They decompose internal activations into sparse combinations of learned basis directions, where each direction represents a feature and its activation indicates its presence in a given input. This sparsity makes SAEs well suited for identifying coherent psychological risk factors as geometric directions in representation space (Mudide et al., 2024).

Concept-based explanation methods aim to explain model behavior using human-aligned concepts rather than raw activations. Testing with Concept Activation Vectors (TCAV) defines a concept by gathering example instances, learns a linear separator in an internal activation space to obtain a concept direction, and then measures how sensitive a model’s prediction is to movement along that direction. Sensitivity is computed via directional derivatives and aggregated into a global score of concept importance for a class (Kim et al., 2018).

In this paper, we adopt a core idea from concept-based explanation methods: defining a concept using example sets and contrasting it with a non-concept baseline. Rather than measuring a concept’s influence on final predictions, as in sensitivity-based approaches like TCAV, we assess whether the concept is geometrically well formed in the model’s internal representation. Sensitivity scores mix representational structure with decision-boundary effects and depend on task-specific gradients, limiting comparability across models and layers. In contrast, directional separability directly measures how distinctly a concept occupies activation space, independent of the classifier head. This makes separability a natural diagnostic for detecting entangled or underrepresented risk factors.

3 Datasets

This work analyzes internal representations learned by models trained on a real-world suicide risk dataset and its augmented variant.

3.1 University of Maryland Reddit Suicidality Dataset (UMD)

The University of Maryland Reddit Suicidality Dataset (UMD) (Zirikly et al., 2019; Shing et al., 2018) is a collection of Reddit posts and comments authored by individuals expressing varying levels of suicide risk. The data are drawn primarily from mental health-related subreddits, including *r/SuicideWatch*, and are annotated at the user level using a four-point ordinal scale: *No Risk*, *Low Risk*, *Moderate Risk*, and *High Risk*.

Following prior work (Ghanadian et al., 2023), we use the subset corresponding to Task A. In this setting,

each user typically contributes a small number of posts, often limited to activity within *r/SuicideWatch*. Because annotations are provided at the user level, all posts authored by a given user are concatenated into a single document, yielding one input instance per user.

This dataset serves as the base distribution for representation analysis in our study. One model is trained solely on UMD, while a second model is trained on an augmented version of the same data, enabling a controlled comparison of how data augmentation affects the geometry and sparsity of learned representations.

3.2 Augmented Dataset

The augmented dataset used in this study follows the topic-aware augmentation strategy introduced in (Ghanadian et al., 2024). In that work, synthetic suicide-related posts were generated using large language models, with ChatGPT (zero-shot setting) producing the highest-quality synthetic dataset. The final augmented training set was constructed by combining this synthetic dataset with progressively increasing subsets of the original UMD training data. Specifically, 10%, 20%, and 30% non-overlapping folds of UMD data were incrementally added to the synthetic corpus, and model performance was evaluated at each stage. The optimal balance was achieved at 30% real data augmentation, where the model matched or surpassed the baseline trained on the full UMD dataset. In the present paper, we utilize augmented dataset presented in (Ghanadian et al., 2024) to investigate how topic-aware augmentation reshapes internal feature representations.

4 Methodology

In this paper, we shift the focus from model outputs to the model’s internal representations, namely the intermediate activation vectors produced during processing. These activations live in the model’s residual-stream activation space and encode semantic and psychological structure implicitly.

To expose such feature directions, we apply dictionary learning via sparse autoencoders to the model’s residual-stream activations. The sparse autoencoder decomposes each activation vector into a sparse linear combination of learned basis directions. Each basis direction, represented by a decoder vector, defines a feature in activation space, while the corresponding scalar output of the encoder indicates the feature activation for a given token. The resulting set of learned feature directions serves as the foundational object of analysis in the remainder of the methodology, enabling both qualitative inspection through top-activating text examples and quantitative geometric analysis.

Once these feature directions are learned, we employ two complementary methods to assess their structure and interpretability. First, we use Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) as a non-linear dimensionality reduction

technique to visualize relationships between feature directions. The input to UMAP consists of the normalized decoder vectors of the sparse autoencoder, so each point in the visualization corresponds to a single learned feature rather than to an input sample. UMAP provides an intuitive view of local clustering, overlap, and separation among feature directions that would be difficult to interpret directly in the original high-dimensional activation space.

Second, we move beyond visualization and introduce a quantitative analysis based on cosine distance. Inspired by Concept-based explanation methods, for each psychological concept, we construct two sets of input texts: a concept set containing samples labeled with that topic, and a non-concept set consisting of unrelated Reddit posts. Each input text is represented by an activation vector extracted from the same residual-stream layer of the model. We compute an average representation for each set by taking the mean of the corresponding activation vectors, yielding one mean vector for the concept set and one for the non-concept set. These mean vectors live in the same residual-stream activation space and represent the average internal state associated with concept-related and unrelated content, respectively. We then compute the cosine distance between these two mean vectors, which measures their angular separation. Larger cosine distance indicates that the model encodes the concept in a direction that is more clearly separated from unrelated text.

4.1 Dictionary Learning on Residual Streams

The linear representation hypothesis proposes that neural networks encode meaningful concepts as approximately linear directions in activation space. In this context, we refer to such directions as features. The superposition hypothesis further suggests that high-dimensional activation spaces contain many nearly orthogonal directions, enabling models to represent more features than the ambient dimensionality by packing them into overlapping subspaces. When many features compete for limited representational capacity, they may become entangled, leading to polysemantic representations.

These hypotheses motivate the use of dictionary learning to recover a set of basis directions such that internal activations can be reconstructed as sparse linear combinations of feature directions. The goal is not only to reconstruct activations accurately, but also to investigate whether the recovered directions correspond to interpretable and semantically meaningful patterns that help explain model behavior (Olshausen and Field, 1997). Recent work has shown that this approach is effective for transformer models, where sparse autoencoders (SAEs) provide a scalable approximation for recovering interpretable and steerable feature directions from residual-stream activations (Bricken et al., 2023; Templeton et al., 2024; OpenAI, 2024). In this study, sparse autoencoders serve as the

core mechanism for dictionary learning and allow us to probe concept-like directions within the model’s residual-stream activation space.

4.1.1 Sparse Autoencoder (SAE)

Sparsity is a key constraint in sparse autoencoders. It enforces that only a small subset of latent units becomes active for any given input activation vector. This regularization prevents information from being diffusely distributed across many units and instead encourages each activation to be represented as a combination of a few distinct basis directions.

In the context of transformer residual streams, overcompleteness is essential for addressing superposition, where many independent features compete for limited representational capacity. By expanding the number of latent units beyond the dimensionality of the residual stream, the autoencoder gains sufficient capacity to separate overlapping directions into more specific and independent features. Overcomplete sparse autoencoders therefore provide a principled mechanism for uncovering monosemantic feature directions that would otherwise remain entangled within dense model activations.

Let $x \in \mathbb{R}^D$ denote a model activation vector extracted from the residual stream, normalized so that the average squared ℓ_2 norm equals the model dimension D . A sparse autoencoder learns F latent features with $F \gg D$, yielding an overcomplete dictionary. The encoder maps the input activation vector to sparse feature activations via

$$f_i(x) = \text{ReLU}(\langle W_{:,i}^{\text{enc}}, x \rangle + b_i^{\text{enc}}), \quad (1)$$

where $W^{\text{enc}} \in \mathbb{R}^{F \times D}$ and $b_{\text{enc}} \in \mathbb{R}^F$ are learned parameters. The scalar quantity $f_i(x)$ represents the activation of feature i for input x .

The decoder reconstructs the original activation vector as a linear combination of learned feature directions:

$$\hat{x} = b_{\text{dec}} + \sum_{i=1}^F f_i(x) W_{:,i}^{\text{dec}}, \quad (2)$$

where $W^{\text{dec}} \in \mathbb{R}^{D \times F}$ and $b_{\text{dec}} \in \mathbb{R}^D$. Each column $W_{:,i}^{\text{dec}}$ defines a feature direction in residual-stream activation space.

The model is trained to minimize a loss function that combines reconstruction error with a sparsity penalty:

$$\mathcal{L} = \mathbb{E}_x \left[\|x - \hat{x}\|_2^2 + \lambda \sum_{i=1}^F f_i(x) \|W_{:,i}^{\text{dec}}\|_2 \right], \quad (3)$$

where λ controls the strength of the sparsity constraint.

Including $\|W_{:,i}^{\text{dec}}\|_2$ inside the penalty allows the decoder vectors to be interpreted as normalized feature directions,

$$\tilde{W}_{:,i}^{\text{dec}} = \frac{W_{:,i}^{\text{dec}}}{\|W_{:,i}^{\text{dec}}\|_2}, \quad (4)$$

with corresponding rescaled feature activations

$$a_i(x) = f_i(x) \|W_{:,i}^{\text{dec}}\|_2. \quad (5)$$

Under this decomposition, each token-level activation vector is explained by a small number of nonzero feature activations drawn from a large dictionary of feature directions. This parts-based representation aligns with the goal of identifying interpretable and potentially monosemantic directions within the model’s residual-stream activation space (Bricken et al., 2023; Templeton et al., 2024).

Overcompleteness ($F > D$) provides sufficient capacity to separate overlapping directions that arise under superposition, while sparsity ensures that only a small number of features activate for any given token. Empirically, increasing F and appropriately tuning the sparsity coefficient improves feature specificity, reduces feature interference, and increases the effectiveness of causal interventions such as feature clamping (Templeton et al., 2024; OpenAI, 2024).

4.2 Feature–Topic Association

To relate learned feature directions to psychological risk factors, we explicitly define a mapping between sparse autoencoder features and annotated topics.

Let i index a learned feature direction, and let $a_i(x)$ denote the rescaled activation of feature i for input sample x as defined in Equation 5. Each sample x is associated with a psychological topic label $t \in \mathcal{T}$, where \mathcal{T} denotes the set of all annotated topics.

For each feature i and each topic t , we compute the average feature activation over all samples belonging to that topic:

$$\bar{a}_{i,t} = \mathbb{E}_{x \sim \mathcal{D}_t} [a_i(x)], \quad (6)$$

where \mathcal{D}_t denotes the set of samples labeled with topic t . This quantity measures how strongly feature i responds, on average, to content associated with topic t .

We define the *dominant topic* of feature i as the topic for which this average activation is maximal:

$$t^*(i) = \arg \max_{t \in \mathcal{T}} \bar{a}_{i,t}. \quad (7)$$

This assignment yields a many-to-one mapping from feature directions to psychological topics. Importantly, this mapping is descriptive rather than exclusive: a feature may respond to multiple topics, but the dominant topic captures the strongest association and provides a principled basis for visualization and analysis.

4.3 UMAP for Latent Feature Visualization

While sparse autoencoders allow us to extract interpretable feature directions from the model’s residual stream activation space, these directions live in a high-dimensional space that is difficult to inspect directly. To gain intuition about the organization of the learned feature directions, we employ Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensionality reduction technique designed

to preserve local neighborhood structure while revealing coarse geometric patterns (McInnes et al., 2018).

In this work, UMAP is applied to the set of learned feature directions, represented by the normalized decoder vectors $\tilde{W}_{:,i}^{\text{dec}}$. Each feature direction is therefore treated as a point in residual-stream activation space, and UMAP provides a two-dimensional embedding that reflects similarities between feature directions. UMAP serves a primarily exploratory and qualitative role in our methodology. It allows us to visually assess whether feature directions associated with similar psychological topics occupy nearby regions of activation space, indicating structured and selective representations, or whether they are dispersed, suggesting stronger superposition and polysemantic encoding. By comparing UMAP embeddings across models trained on different datasets, we can further examine how data augmentation affects the geometric organization of risk-related feature directions. Importantly, UMAP is not used to define or enforce feature structure, but rather to reveal patterns that emerge naturally from the learned representations.

4.4 Cosine Distance for Quantitative Separability Analysis

Cosine distance is well suited to linear representation settings because it measures angular differences between mean activation vectors. If internal representations align with approximately linear directions, larger cosine distance indicates clearer directional separation between concept and non-concept sets. This makes separability easier to interpret and directly relatable to feature directions recovered by the sparse autoencoder. UMAP provides an intuitive visualization of feature relationships but does not quantify how clearly psychological concepts are separated. Cosine distance addresses this by measuring directional differences while ignoring magnitude, allowing us to assess whether a concept is encoded distinctly from a non-concept baseline.

To complement the qualitative insights from visualization, we introduce a quantitative analysis based on cosine distance computed in the sparse autoencoder (SAE) feature space. For each psychological concept, we define a concept set consisting of posts associated with that concept and a corresponding non-concept set drawn from unrelated Reddit content. For each post, token-level residual-stream activations are first mapped into the SAE latent space, yielding a vector of feature activations.

Representations are extracted from the same model layer for both concept and non-concept sets and mapped into the sparse autoencoder feature space. Within this space, we compute mean feature activation vectors for the concept set and the non-concept set. To obtain post-level feature activation representations, we aggregate token-level feature activations within each post using a max-pooling operation. For each feature, the post-

level activation is defined as the maximum activation across all tokens in the post. This choice reflects the assumption that the presence of a psychologically meaningful feature anywhere in a post is sufficient to indicate its relevance and avoids dilution of sparse but salient signals that would otherwise occur with averaging across tokens.

For each psychological topic, we then compute a mean post-level feature activation vector by averaging these post-level representations across all posts in the concept set, yielding μ_{concept} . An analogous mean feature activation vector μ_{non} is computed from the non-concept baseline. Larger cosine distances indicate that concept-related and non-concept content activate more distinct sets of interpretable features, reflecting clearer separation in the model’s internal representation. Separation between the two is quantified using cosine distance,

$$d_{\text{cos}} = 1 - \frac{\mu_{\text{concept}} \cdot \mu_{\text{non}}}{\|\mu_{\text{concept}}\|_2 \|\mu_{\text{non}}\|_2}. \quad (8)$$

Despite TCAV methods that measure sensitivity of model outputs to these directions, we use cosine distance to quantify their geometric separability within the model’s internal activation space. Prior work in representation analysis and concept-based interpretability has similarly relied on cosine similarity or distance to assess alignment between learned representations and semantic concepts (Kim et al., 2018; Bricken et al., 2023).

Larger values indicate greater separation between the concept and the non-concept baseline. We compute d_{cos} for both the UMD fine-tuned model and the AUG fine-tuned model, each evaluated on the same mixed synthetic test set, and repeat this procedure for every topic to obtain a cosine distance per topic.

4.5 Experimental Setup

In our Experimentation we utilize ALBERT model(Albert-base-V2)¹, a lightweight transformer model that reduces parameter redundancy through cross-layer parameter sharing and factorized embedding parameterization. We fine-tune ALBERT separately on the UMD dataset and the augmented dataset (Ghanadian et al., 2025). ALBERT is particularly suitable for our interpretability analysis because its compact architecture and shared layers produce stable residual-stream representations, enabling clearer examination of learned feature directions through sparse autoencoders.

We train a sparse autoencoder on token-level residual-stream activations from a fine-tuned ALBERT classifier in order to recover an overcomplete set of feature directions that can be inspected and analyzed. We focus on the final-layer residual stream because it captures high-level semantic representations that are most relevant to the downstream suicide ideation

¹ALBERT Model

detection task, while remaining upstream of the classification head. This choice allows us to analyze psychologically meaningful directions that the model uses for decision-making, without conflating them with task-specific output weights. Prior work in mechanistic interpretability has similarly shown that late-layer residual streams contain rich, interpretable features aligned with abstract concepts, making them suitable targets for sparse dictionary learning (Bricken et al., 2023; Templeton, 2024).

We adopt the encoder–decoder architecture described in Section 4.1. An overcomplete feature dictionary is used, with $F = 4096$ learned feature directions for input activation vectors of dimension $D = 768$. This expansion increases the representational capacity of the dictionary and improves the separation of overlapping directions that arise due to superposition, allowing more specific and interpretable feature directions to emerge. The token dimension is flattened and filtered using the attention mask so that only non-padding token activations are fed to the sparse autoencoder. The encoder consists of a linear map followed by a ReLU nonlinearity, yielding non-negative and selective feature activations. The decoder is a linear map that reconstructs the original activation vector from a sparse linear combination of feature directions. Under this formulation, the decoder columns represent candidate feature directions, while the encoder outputs indicate how strongly each feature direction is activated for a given token.

Optimization is performed using the AdamW optimizer, which provides stable convergence and improved control over model capacity by decoupling weight decay from gradient updates. Early stopping is applied based on validation reconstruction loss to prevent overfitting once learning plateaus. After training, feature activations are computed for all tokens in the dataset and aggregated at the sample (post) level by taking the maximum activation per feature across tokens within each text. This aggregation reflects the assumption that a feature is present in a sample if it activates strongly for at least one token.

For each learned feature direction, we then select the top 100 samples with the highest feature activations. This procedure follows established practice in sparse autoencoder interpretability work, where a fixed number of top-activating examples has been shown to be sufficient for reliably characterizing feature semantics while remaining computationally tractable (Templeton, 2024; Bricken et al., 2023).

5 Results

5.1 Analysis I: Latent Feature Visualization

A central question in this analysis is whether suicide-related psychological risk factors such as hopelessness, anxiety, and social isolation form coherent geometric structure in the model’s learned feature directions. We

also ask whether a topic-aware data augmentation improves the coherence of these features.

For each model—ALBERT fine-tuned on the UMD dataset and ALBERT fine-tuned on the augmented dataset—we assign a dominant psychological topic to each learned feature direction using the following procedure. First, for a given feature, we collect its activation values across all posts in the dataset. Second, posts are grouped according to their annotated psychological topic. For each topic, we compute the average feature activation over all posts belonging to that topic. The dominant topic of the feature is defined as the topic for which this average activation is highest. This assignment provides a coarse but interpretable mapping between learned feature directions and psychological risk factors.

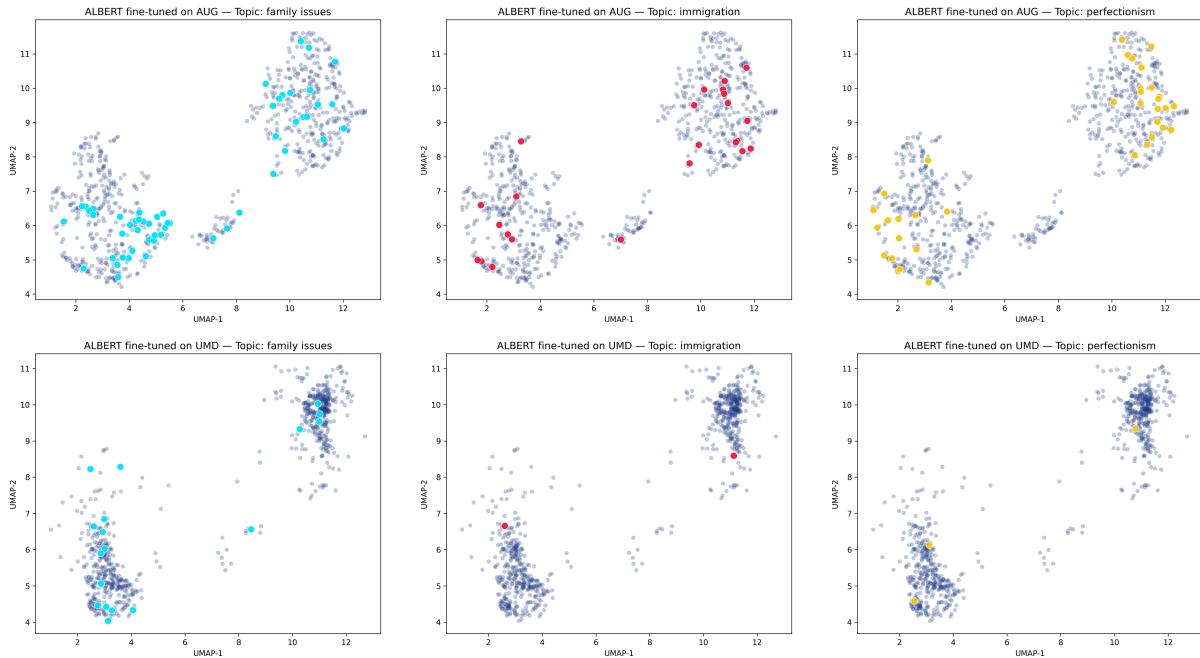
Throughout this paper, sparse autoencoders are trained on token-level residual-stream activations in order to capture fine-grained linguistic features. However, psychological topics are defined at the post level rather than at the token level. To bridge this gap, token-level feature activations are aggregated within each post using a max-pooling operation, yielding a post-level feature activation vector. All topic assignments, visualizations, and quantitative analyses reported in this paper are therefore performed at the post level, unless explicitly stated otherwise.

After assigning dominant topics, the full set of feature directions (decoder vectors) is projected into two dimensions using UMAP. In Figure 1, we visualize the geometry of the feature space one topic at a time by highlighting features whose dominant topic matches the topic of interest, while rendering all remaining feature directions as a low-opacity background. Compact highlighted regions indicate feature directions that respond selectively to the same psychological concept, suggesting movement toward monosemantic structure. In contrast, dispersed highlighted features suggest shared or overlapping representations, consistent with superposition and polysemantic encoding. For each topic, we present paired UMAP panels, with the AUG-trained model shown on the left and the UMD-trained model on the right, enabling direct comparison of how data augmentation affects the geometric organization of risk-related feature directions.

In Figure 1, each point corresponds to a learned feature direction, represented by its decoder vector in the model’s residual-stream space. The colored points indicate features whose dominant activations are associated with the indicated psychological topic. In the UMD fine-tuned model, topics such as perfectionism and immigration activate only a small number of scattered feature directions, suggesting limited topic coverage and stronger superposition, as the same directions appear to respond to multiple phenomena.

In contrast, the augmented fine-tuned model exhibits a larger number of highlighted feature directions for the same topics, which often form compact regions within

Figure 1: UMAP projections of features highlighting three selected topics. Left column: ALBERT fine-tuned on AUG (the augmented dataset with high topic coverage). Right column: ALBERT fine-tuned on UMD. Colored points mark features whose dominant responses align with the indicated topic; dark blue points show all other features.



the two macro-clusters of the embedding. These regions indicate sets of directions that respond more selectively to the target topic. The family issues topic follows a similar pattern, with sparse and dispersed feature directions in the UMD model and denser, more coherent groupings in the AUG model. This shift suggests that data augmentation improves both the availability and the consistency of topic-specific directions in the learned representation space.

Monosemanticity is valuable because a compact, well-separated group of topic-selective features supports clear explanations, stable linear steering, and predictable causal edits. However, polysemanticity is not inherently undesirable. Distributed highlights can reflect useful sharing of linguistic substrates, overlapping risk factors, or genuine subtopics that co-occur in real text. Such sharing can improve generalization and parameter efficiency when concepts partially overlap. The practical concern is interference rather than polysemanticity itself; if the same features fire for unrelated content, the representation becomes noisy and separation weakens. The visualizations indicate that augmentation moved several topics toward monosemanticity by increasing the number and coherence of topic-selective directions, while still preserving beneficial sharing where concepts naturally overlap.

5.2 Analysis II: Quantifying Concept Separability

The goal of this section is to quantify how clearly each psychological topic is separated from generic,

non-risk-related text in the model’s internal activation space. We use cosine distance as a measure of separability because it captures directional differences between mean activation vectors associated with concept-related and non-concept-related content. If the mean vectors corresponding to two sets of points in noticeably different directions, this indicates that the model encodes them as distinct internal representations. Because cosine distance depends only on direction, it is robust to changes in vector magnitude caused by scaling, preprocessing, or dataset-specific norm shifts.

For post-level analysis, token-level activation vectors within each post are averaged to obtain a single activation vector per post. We evaluate two models that are fine-tuned separately: one model fine-tuned on the UMD dataset and one model fine-tuned on the augmented dataset. All measurements are computed on a held-out test set consisting of a mixture of synthetic posts annotated with psychological topic labels. No training or validation data are used in these analyses.

In addition to the synthetic dataset with topic labels, we use a separate Reddit dataset, which contains content unrelated to suicidal ideation. This Reddit set is treated as the non-concept baseline and is collected from unrelated subreddits to suicidal ideation. We merge the synthetic topic-labeled posts and the Reddit posts to form the final test set used for analysis. For each psychological topic, we define the **concept set** as all test posts labeled with that topic, and the **non-concept set** as the Reddit test set.

Table 1 summarizes topic separation on the synthetic

Table 1: Cosine distance between topic means and a general Reddit baseline for models fine-tuned on AUG and UMD, evaluated on the synthetic test set. Distances reflect angular separation in representation space (higher is better). Δ is AUG-UMD and summarizes the change due to augmentation.

Topic	Cosine Distance in AUG	Cosine Distance in UMD	$\Delta(\text{AUG-UMD})$
Relationship Problems	0.513	0.178	0.334
Financial Crisis	0.524	0.213	0.310
Family Issues	0.466	0.199	0.266
Anxiety	0.268	0.0492	0.218
Racism	0.239	0.024	0.215
Anger	0.237	0.033	0.204
Bullying	0.206	0.019	0.186
Immigration	0.199	0.015	0.184
Perfectionism	0.153	0.041	0.113
Death of Closed One	0.212	0.108	0.105
Education	0.125	0.077	0.048
Hopelessness	0.200	0.217	-0.017
Unemployment	0.824	0.903	-0.079

test set by showing the cosine distance between each topic’s mean SAE feature activation vector and a non-concept Reddit baseline. Higher cosine distance indicates stronger directional separation of a topic from general Reddit content, reflecting more selective internal representations.

Across most topics, augmentation leads to a clear increase in directional separation. The largest gains are observed for relationship problems (+0.334), financial crisis (+0.311), and family issues (+0.266), indicating that augmentation substantially sharpens the representation of these psychosocial factors relative to general Reddit content. Additional consistent improvements are visible for anxiety, racism, anger, bullying, and immigration (approximately +0.18 to +0.22), suggesting that topic-aware augmentation broadly improves the model’s ability to isolate diverse risk-related themes in its internal geometry.

More moderate gains are observed for perfectionism (+0.113), death of a close one (+0.105), depression (+0.092), and education (+0.048). These smaller increases indicate partial improvement in separability, consistent with topics that are either less lexically distinctive or more diffusely expressed across posts.

Two topics deviate from the overall trend. Hopelessness shows a slight decrease (−0.017), likely reflecting its strong semantic overlap with other risk factors such as depression and anxiety, which limits how distinctly it can be separated as an independent direction. Unemployment exhibits a larger decrease (−0.079); however, this topic already displays very high separation in both models (approximately 0.82–0.90). Because its baseline separation is near the upper bound of the scale, there is limited room for further improvement, and small fluctuations appear as a decrease despite the topic remaining strongly isolated from the non-concept baseline.

6 Discussions and Conclusion

This paper examined how suicide ideation detection models encode psychologically meaningful topics and how these representations change under topic-aware augmentation. Grounded in the linear representation and superposition hypotheses, we used overcomplete sparse autoencoders to recover feature directions from residual activations and analyzed their structure using UMAP visualizations and cosine-based separability. Together, these methods allowed us to evaluate not only which psychological topics are encoded in the model’s latent space, but also how selectively and distinctly they are represented.

Three findings emerge. First, UMAP visualizations show that topic-aware augmentation increases the number and compactness of topic-selective feature clusters, indicating a shift toward more structured and monosemantic representations. Second, cosine analysis confirms this effect quantitatively: augmentation generally increases directional separation between topic means and a general Reddit baseline, with the largest gains observed for relationship problems, financial crisis, and family issues. Third, some topics behave differently. Hopelessness shows little change, likely due to semantic overlap with other risk factors, while unemployment decreases slightly from an already high baseline, suggesting a ceiling effect rather than representational degradation.

These results refine our understanding of interpretability. While monosemantic features support explanation and intervention, some polysemantic sharing may reflect natural topic overlap and aid generalization. The key concern is not superposition itself, but harmful interference that blurs concept boundaries. Our findings suggest that topic-aware augmentation reduces such interference for many risk factors while preserving meaningful overlap, thereby improving both coverage and separability of psychological topics in the model’s internal geometry without imposing overly rigid representations.

7 Future Works

This paper explored suicide ideation detection from multiple complementary perspectives, including the use of large language models for classification, synthetic data generation to address dataset limitations, and representation-level analysis to better understand learned decision signals. While the results demonstrate the effectiveness of topic-aware augmentation and highlight the potential of both generative and discriminative models, several important directions remain for future research.

This paper provides insight into how topic-aware augmentation influences internal representations. Future work could build on this analysis by introducing causal interventions on learned features, such as clamping or amplifying sparse autoencoder activations,

to assess their direct impact on model predictions. This would strengthen the connection between descriptive interpretability and functional model behavior, helping distinguish features that are merely correlated with suicidality from those that actively influence decisions.

The analyses in this paper focused on individual posts treated as independent samples. However, suicidal ideation often manifests through evolving language patterns over time. Future research could extend both classification and representation analyses to longitudinal user-level data, examining how risk-related signals accumulate or change across sequences of posts. This may enable earlier detection of emerging risk and provide a more realistic modeling framework for real-world applications.

Finally, future work should examine how both classifiers and synthetic-data-driven models perform across different populations, platforms, and linguistic styles. Assessing robustness to domain shifts and potential demographic biases is particularly important in mental health applications. Incorporating fairness-aware evaluation alongside performance and interpretability analyses would help ensure that improvements in accuracy do not come at the expense of equitable and responsible deployment.

8 Limitations

Despite providing insight into internal representations, the interpretability analysis in this thesis has several limitations. First, sparse autoencoders approximate the model’s internal structure but do not guarantee recovery of true underlying computational features. Learned features may still reflect mixtures of signals or residual polysemanticity. Second, cosine-based separability measures capture directional differences but do not establish causal importance for model predictions. Third, UMAP visualizations, while useful for understanding structure, depend on hyperparameters and dimensionality reduction choices that may influence observed cluster patterns. Moreover, the analysis was conducted on a specific model architecture and dataset configuration, and therefore the findings may not fully generalize to other architectures, scales, or domains.

References

Michal Aharon, Michael Elad, and Alfred Bruckstein. 2006. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322.

Johan Anderberg, Mats Bogren, Cecilia Mattisson, and Louise Brådvik. 2016. Long-term suicide risk in anxiety—the lundby study 1947–2011. *Archives of Suicide Research*, 20(3):463–475.

Ella Arensman, Celine Larkin, J McCarthy, Sara Leitao, Paul Corcoran, Eileen Williamson, C McAuliffe,

Ivan J Perry, Eve Griffin, EM Cassidy, et al. 2019. Psychosocial, psychiatric and work-related risk factors associated with suicide in ireland: optimised methodological approach of a case-control psychological autopsy study. *BMC psychiatry*, 19:1–11.

Hilary Arksey and Lisa O’Malley. 2005. Scoping studies: towards a methodological framework. *International journal of social research methodology*, 8(1):19–32.

Rebecca Armstrong, Belinda J Hall, Jodie Doyle, and Elizabeth Waters. 2011. ‘scoping the scope’ of a cochrane review. *Journal of public health*, 33(1):147–150.

Alison Athey, James Overholser, Courtney Bagge, Lesa Dieter, Eric Vallender, and Craig A Stockmeier. 2018. Risk-taking behaviors and stressors differentially predict suicidal preparation, non-fatal suicide attempts, and suicide deaths. *Psychiatry research*, 270:160–167.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2022a. Challenges in applying explainability methods to improve the fairness of nlp models. *arXiv preprint arXiv:2206.03945*.

Esma Balkir, Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. 2022b. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. *arXiv preprint arXiv:2205.03302*.

Luke T Bayliss, Steven Christensen, Andrea Lamont-Mills, and Carol du Plessis. 2022. Suicide capability within the ideation-to-action framework: A systematic scoping review. *PloS one*, 17(10):e0276070.

Chris M Bishop. 1995. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116.

Panchanit Boonyarat, Di Jie Liew, and Yung-Chun Chang. 2024. Leveraging enhanced bert models for detecting suicidal ideation in thai social media content amidst covid-19. *Information Processing & Management*, 61(4):103706.

M Dolores Braquehais, Maria A Oquendo, Enrique Baca-García, and Leo Sher. 2010. Is impulsivity a link between childhood abuse and suicide? *Comprehensive psychiatry*, 51(2):121–129.

Trent Bricken, Lawrence Chan, Nelson Elhage Qiu, and Chris Olah. 2023. [Towards monosemanticity: Decomposing language models with dictionary learning](#). *Transformer Circuits Thread*.

Teresa Brockie, Michelle Kahn-John, Laura Mata Lopez, Eleesha Bell, Truth Brockie, Terry Brockie, Ellie Decker, Nancy Glass, Hannah Has Eagle, Kenneth Helgeson, et al. 2024. A mixed-methods study protocol on factors contributing to suicide clusters among native american youth in

- a northern plains reservation. *Frontiers in public health*, 11:1281109.
- Beth S Brodsky, Kevin M Malone, Steven P Ellis, Rebecca A Dulit, and J John Mann. 1997. Characteristics of borderline personality disorder associated with suicidal behavior. *American Journal of Psychiatry*, 154(12):1715–1719.
- Silje Støle Brokke, Thomas Bjerregaard Bertelsen, Nils Inge Landrø, and Vegard Øksendal Haaland. 2022. The effect of sexual abuse and dissociation on suicide attempt. *BMC psychiatry*, 22:1–8.
- Chaofan Chen, Oscar Li, Chaoxu Tao, Antonia Barnett, Jie Su, and Cynthia Rudin. 2020. [Concept whitening for interpretable image recognition](#). In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1135–1145.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Karen M Devries, Joelle YT Mak, Jennifer C Child, Gail Falder, Loraine J Bacchus, Jill Astbury, and Charlotte H Watts. 2014. Childhood sexual abuse and suicidal behavior: a meta-analysis. *Pediatrics*, 133(5):e1331–e1344.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shan Carter, Tom Brown, and Chris Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. <https://transformer-circuits.pub/2023/monosemanticity/>. Anthropic.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.
- Louis Favril, Rongqin Yu, John R Geddes, and Seena Fazel. 2023. Individual-level risk factors for suicide mortality in the general population: an umbrella review. *The Lancet Public Health*, 8(11):e868–e877.
- David Finkelhor. 1982. Sexual abuse: A sociological perspective. *Child abuse & neglect*, 6(1):95–102.
- Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.
- Lin Gan, Tao Yang, Yifan Huang, Boxiong Yang, Yami Yanwen Luo, Lui Wing Cheung Richard, and Dabo Guo. 2023. Experimental comparison of three topic modeling methods with lda, top2vec and bertopic. In *International Symposium on Artificial Intelligence and Robotics*, pages 376–391. Springer.
- Charles Merriam George Merriam. 1831. [Merriam-webster dictionaries](#).
- Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2024. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access*.
- Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2025. Improving suicidal ideation detection in social media posts: Topic modeling and synthetic data augmentation approach. *JMIR Formative Research*, 9:e63272.
- Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2023. Chatgpt for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials and limitations. *arXiv preprint arXiv:2306.09390*.
- Giancarlo Giupponi, Marco Innamorati, Ross J Baldessarini, Diego De Leo, Francesca de Giovannelli, Roger Pycha, Andreas Conca, Paolo Girardi, and Maurizio Pompili. 2018. Factors associated with suicide: Case-control study in south tyrol. *Comprehensive psychiatry*, 80:150–154.
- Tina R Goldstein, Wonho Ha, David A Axelson, Benjamin I Goldstein, Fangzi Liao, Mary Kay Gill, Neal D Ryan, Shirley Yen, Jeffrey Hunt, Heather Hower, et al. 2012. Predictors of prospectively examined suicide attempts among youth with bipolar disorder. *Archives of general psychiatry*, 69(11):1113–1122.
- Frederick K Goodwin and Kay Redfield Jamison. 2007. *Manic-depressive illness: bipolar disorders and recurrent depression*, volume 2. Oxford university press.
- Cameron Grover, Jacqueline Huber, Matthew Brewer, Ashna Basu, and Matthew Large. 2023. Meta-analysis of clinical risk factors for suicide among people presenting to emergency departments and general hospitals with suicidal thoughts and behaviours. *Acta Psychiatrica Scandinavica*, 148(6):491–524.
- E Clare Harris and Brian Barraclough. 1997. Suicide as an outcome for mental disorders. a meta-analysis. *British journal of psychiatry*, 170(3):205–228.
- Keith Hawton, Lesley Sutton, Camilla Haw, Julia Sinclair, and Jonathan J Deeks. 2005. Schizophrenia and suicide: systematic review of risk factors. *The British Journal of Psychiatry*, 187(1):9–20.
- Stephanie Homan, Marion Gabi, Nina Klee, Sandro Bachmann, Ann-Marie Moser, Sofia Michel, Anna-Marie Bertram, Anke Maatz, Guido Seiler, Elisabeth Stark, et al. 2022. Linguistic features of suicidal thoughts and behaviors: A systematic review. *Clinical psychology review*, 95:102161.
- Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yihan Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, and Andrew Beam. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.

- Jin-Mahn Jang, Jong-Il Park, Keun-Young Oh, Keon-Hak Lee, Myung Sig Kim, Myeong-Sook Yoon, Sung-Hee Ko, Hye-Chung Cho, and Young-Chul Chung. 2014. Predictors of suicidal ideation in a community sample: roles of anger, self-esteem, and depression. *Psychiatry research*, 216(1):74–81.
- Smriti Jha, Gerry Chan, and Rita Orji. 2023. Identification of risk factors for suicide and insights for developing suicide prevention technologies: A systematic review and meta-analysis. *Human Behavior and Emerging Technologies*, 2023(1):3923097.
- Been Kim, Martin Wattenberhorban, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.
- Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Gregory Valiant, and Percy Liang. 2020. **Concept bottleneck models**. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5338–5348.
- Vernon M Langford. 2023. Risk factors for suicide in men. *Nursing Clinics*, 58(4):513–524.
- Aurélie M Lasserre, Helena Marti-Soler, Marie-Pierre F Strippoli, Julien Vaucher, Jennifer Glaus, Caroline L Vandeleur, Enrique Castelao, Pedro Marques-Vidal, Gérard Waeber, Peter Vollenweider, et al. 2016. Clinical and course characteristics of depression and all-cause mortality: a prospective population-based study. *Journal of affective disorders*, 189:17–24.
- Klaus Lieb, Mary C Zanarini, Christian Schmahl, Marsha M Linehan, and Martin Bohus. 2004. Borderline personality disorder. *The Lancet*, 364(9432):453–461.
- Jutta Lindert, Ondine S von Ehrenstein, Rachel Grashow, Gilad Gal, Elmar Braehler, and Marc G Weisskopf. 2014. Sexual and physical abuse in childhood is associated with depression and anxiety over the life course: systematic review and meta-analysis. *International journal of public health*, 59:359–372.
- Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Modhurima Moitra, Damian Santomauro, Louisa Degenhardt, Pamela Y Collins, Harvey Whiteford, Theo Vos, and Alize Ferrari. 2021. Estimating the risk of suicide associated with mental disorders: A systematic review and meta-regression analysis. *Journal of psychiatric research*, 137:242–249.
- Anish Mudide, Joshua Engels, Eric J Michaud, Max Tegmark, and Christian Schroeder de Witt. 2024. Efficient dictionary learning with switch sparse autoencoders. *arXiv preprint arXiv:2410.08201*.
- Zachary Munn, Micah DJ Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, 18:1–7.
- Isar Nejadgholi, Esma Balkır, Kathleen C Fraser, and Svetlana Kiritchenko. 2022. Towards procedural fairness: Uncovering biases in how a toxic language classifier uses sentiment information. *arXiv preprint arXiv:2210.10689*.
- Isar Nejadgholi, Svetlana Kiritchenko, Kathleen C Fraser, and Esma Balkır. 2023. Concept-based explanations to test for false causal relationships learned by abusive language classifiers. *arXiv preprint arXiv:2307.01900*.
- Rory C O’Connor and Matthew K Nock. 2014. The psychology of suicidal behaviour. *The Lancet Psychiatry*, 1(1):73–85.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.
- Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill*, 2(11):e7.
- Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill*, 3(3):e10.
- Bruno A Olshausen and David J Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325.
- OpenAI. 2024. **Sparse autoencoders for interpretability**. *OpenAI Research*.
- Andrew Page, Stephen Morrell, Coletta Hobbs, Greg Carter, Michael Dudley, Johan Dufflou, and Richard Taylor. 2014. Suicide in young adults: psychiatric and socio-economic factors from a case–control study. *BMC psychiatry*, 14:1–9.
- Joel Paris. 2002. Chronic suicidality among patients with borderline personality disorder. *Psychiatric services*, 53(6):738–742.
- Andres Pemau, Carolina Marin-Martin, Marina Diaz-Marsa, Alejandro de la Torre-Luque, Wala Ayad-Ahmed, Ana Gonzalez-Pinto, Nathalia Garrido-Torres, Lucia Garrido-Sanchez, Natalia Roberto, Purificación Lopez-Peña, et al. 2024. Risk factors for suicide reattempt: a systematic review and meta-analysis. *Psychological medicine*, pages 1–8.
- Micah DJ Peters, Christina M Godfrey, Hanan Khalil, Patricia McInerney, Deborah Parker, and Cassia Baldini Soares. 2015. Guidance for conducting systematic scoping reviews. *JBIM Evidence Implementation*, 13(3):141–146.

- Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2023. Concept-based explainable artificial intelligence: A survey. *ACM Computing Surveys*.
- Mélanie Racine. 2018. Chronic pain and suicide risk: a comprehensive review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 87:269–280.
- Vikram V Ramaswamy, Sunnie SY Kim, Ruth Fong, and Olga Russakovsky. 2023. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10932–10941.
- Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB endowment. International conference on very large data bases*, volume 11, page 269. NIH Public Access.
- Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.
- Stephen H Schultz, Stephen W North, and Cleveland G Shields. 2007. Schizophrenia: a review. *American family physician*, 75(12):1821–1829.
- Leo Sher and René S Kahn. 2019. Suicide in schizophrenia: an educational overview. *Medicina*, 55(7):361.
- Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenber, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.
- Loitongbam Gyanendro Singh, Junyu Mao, Rudra Mutalik, and Stuart E Middleton. 2024. Extraction and summarization of suicidal ideation evidence in social media content using large language models. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics*.
- Adly Templeton. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.
- John Templeton, Jesse Mu, Andrew Jones, Adrien Floreano, Sabri Eyuboglu, et al. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. *Anthropic Research*.
- Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. *Explainable depression detection using large language models on social media data*. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126, St. Julians, Malta. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Kristin L Wolfe, Paul A Nakonezny, Victoria J Owen, Katherine V Rial, Alexandra P Moorehead, Beth D Kennard, and Graham J Emslie. 2019. Hopelessness as a predictor of suicide ideation in depressed male and female adolescent youth. *Suicide and Life-Threatening Behavior*, 49(1):253–263.
- Chih-Kuan Yeh, Cho-Jui Hsieh, Arun Suggala, David I. Inouye, and Pradeep Ravikumar. 2020. *On concept-based explanations in deep neural networks*. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10278–10289.
- Mary C Zanarini, Frances R Frankenburg, John Hennen, D Bradford Reich, and Kenneth R Silk. 2006. Prediction of the 10-year course of borderline personality disorder. *American Journal of Psychiatry*, 163(5):827–832.
- Qi Zhang, Yifei Wang, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, and Yisen Wang. 2024. Beyond interpretability: The gains of feature monosemanticity on model robustness. *arXiv preprint arXiv:2410.21331*.
- Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.