From Semantics to Style: A Cross-Dataset Comparative Framework for Sentence Similarity Predictions

Yusuke Yamauchi[1] **Akiko Aizawa**[1,2†]
[1]The University of Tokyo, Tokyo, Japan
[2]National Institute of Informatics, Tokyo, Japan

Abstract

While Semantic Textual Similarity (STS) task serves as a cornerstone embedding task in natural language processing, the definition of similarity is inherently ambiguous and dataset-specific. Comprehensive cross-dataset analysis remains scarce, leaving it uncertain whether language models perceive diverse semantic and stylistic nuances as humans do. To address this, we propose a comparative framework utilizing lightweight poolers on a frozen encoder to conduct a unified analysis across STS, Paraphrase Identification (PI), and Triplet datasets. The STS and PI datasets encompass a diverse range of semantics, whereas the Triplet dataset includes styles of communication (e.g., poetic, offensive, or objective). Experimental results on 21 datasets indicate a high correlation of notions of semantic equivalence between STS and PI settings, while highlighting style as a distinct dimension necessitating explicit separation from semantics. Moreover, Procrustes, layer-wise and hierarchical clustering analyses elucidate the varying properties of the model's internal representation of meaning and the structural organization of the embedding space. These insights imply that treating semantics and style as separate components in embedding models is crucial for enhancing both interpretability and practical utility.[1]

1 Introduction

Semantic Textual Similarity (STS) (Agirre et al., 2012, 2013) represents the semantic closeness between two sentences, and serves as a fundamental component in a wide range of applications, such as information retrieval (Iida and Okazaki, 2021), document clustering (Rafi and Shaikh, 2013), and duplicate detection (Dammu and Alonso, 2024). While the STS task has long been used as a benchmark for evaluating the performance of text embedding models, it remains unclear to what extent these models capture sentence meaning akin to human perception. This uncertainty stems from the nature of semantic similarity, which remains inherently ambiguous and subjective, varying across datasets. Furthermore, language models can misalign with human judgment, treating trivial differences as meaningful. For instance, in the example shown in Figure 1, the embedding model assigns similarity scores to sentence pairs in a way that reflects a greater sensitivity to differences in vocabulary than to differences in the overall semantic meaning of the sentences. Although language models are capable of accurately capturing information in complex and long texts, such discrepancies in meaning perception persist even in simple sentences.

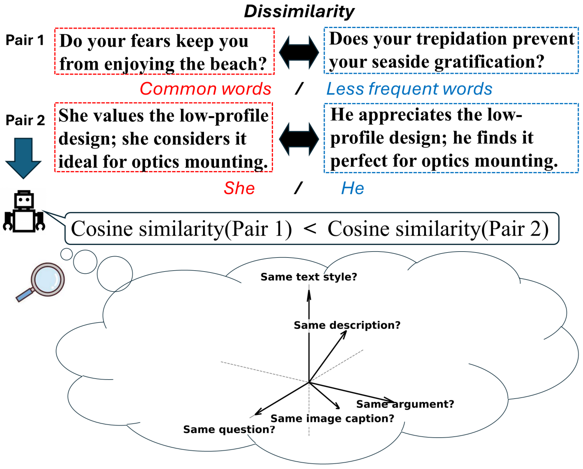

Dissimilarity
Pair 1
Do your fears keep you from enjoying the beach?
Does your trepidation prevent your seaside gratification?
Common words
/
Less frequent words
Pair 2
She values the low-profile design; she considers it ideal for optics mounting.
He appreciates the low-profile design; he finds it perfect for optics mounting.
She
/
He
Cosine similarity(Pair 1) < Cosine similarity(Pair 2)
Same text style?
Same description?
Same argument?
Same image caption?
Same question?

Figure 1: Pair 1 uses different vocabulary, whereas Pair 2 has a different subject. Human judges would perceive Pair 2 as having a different meaning. However, differences in vocabulary exert a greater influence on the model's similarity prediction than variations in subject. In this study, we analyze how the model captures semantic differences between sentences.

In recent years, there has been increasing interest in the interpretability and explainability of why an embedding model assigns high or low similarity to a given pair of sentences (Opitz et al.,

[1] Our code is available at https://github.com/yama11235/SemanticStylePooler.

Findings of the Association for Computational Linguistics: EACL 2026, pages 1848–1877
March 24-29, 2026 ©2026 Association for Computational Linguistics

2025). Existing approaches investigate the contribution of token pairs to the model's predicted semantic similarity (Lee et al., 2022; Moeller et al., 2024; Vasileiou and Eberle, 2024), or they aim to build explicitly explainable models by assigning an interpretable meaning to each dimension of the embeddings (Opitz and Frank, 2022; Sun et al., 2024). While these studies contribute to improving the transparency of the STS task, their scope remains limited, often restricted to datasets derived from image captions or surface-level semantics. Sentence differences encompass diverse elements beyond lexical and structural changes, such as numerical expressions and style. Therefore, analyzing how models capture these differences necessitates a comprehensive verification using datasets collected from a wide range of domains and tasks.

In this paper, we introduce a framework for cross-dataset comparison that spans tasks with distinct training and inference paradigms. By comparing embeddings produced by lightweight poolers attached to a frozen encoder, our approach places diverse aspects on a common cosine scale and enables practical, flexible analysis. To this end, we address the following research questions:

- Does the pooler distinguish diverse types of sentence differences as distinct properties?
- How does the model capture the relationship between semantics and style?

The key contributions of this work are as follows:

1. We collect 21 sentence-similarity datasets and propose a multi-perspective representation framework with separate lightweight poolers for semantics and style on a frozen encoder.

2. We show that semantic notions are largely shared across STS/Paraphrase identification datasets (especially in BERT, pairwise correlations > 0.9). Conversely, we show that style is a concept distinct from semantics, necessitating explicit separation from semantic information for effective extraction.

3. We conduct Procrustes, layer-wise, and clustering analyses, revealing that: (i) high similarity in pooler predictions is mirrored by the proximity of their weight parameters; (ii) semantic tasks peak in later layers, while many stylistic cues peak in earlier-to-mid layers; and (iii) semantic datasets form a tight cluster in our Mean Difference dendrogram, whereas style and sentiment remain substantially distinct.

2 Related Work

Task Formulations and Definitions of Semantic Similarity. Semantic Textual Similarity (STS) tasks aim to quantify the degree of meaning overlap between two sentences. The task was originally defined as: "Semantic Textual Similarity measures the degree of semantic equivalence between two sentences" (Agirre et al., 2012, 2013). The general definition has remained stable, but the practical implementation of similarity differs among datasets according to their target use cases. In most cases, semantic similarity is defined in a relative manner within each dataset, guided by short descriptions and a handful of examples. Due to the absence of unified linguistic standards, the notion of semantic similarity remains inherently dataset-specific and often reflects human intuitions rather than formal definitions.

Defining Style in Contrast to Semantics. A complementary line of work focuses on the notion of style especially in the domain of Text Style Transfer (TST). Style has been conceptualized from two primary perspectives (Hu et al., 2023; Jin et al., 2022). The linguistic perspective separates content and style by treating the propositional meaning or topic as semantic, while aspects of expression such as tone, formality, or humor are categorized as style (McDonald and Pustejovsky, 1985). While the distinction is intuitively clear, it remains challenging to define or implement in a precise manner. As a result, most deep learning research on TST adopts a data-driven perspective. In this approach, stylistic features are identified based on patterns of consistent variation across or within datasets, regardless of their potential impact on meaning. For example, emotional tone is often treated as a stylistic feature, even though it can influence the semantic content of a sentence.

Semantic Relatedness as a Broader Construct. Some recent datasets adopt the term "relatedness" rather than "similarity" to characterize semantic closeness (Abdalla et al., 2023; Ousidhoum et al., 2024). Relatedness is a broader concept that encompasses a wider range of connections between sentences. Two sentences may be related even if they are not paraphrases and do not exhibit entailment, as long as they refer to a shared context or topic. For example, "*There was a lemon tree next to* *the house.*" and "*The boy enjoyed reading under the lemon tree.*" do not convey the same proposition but are clearly linked by a common referent, and thus are highly related. This broader framing allows for the modeling of connections that go beyond strict semantic overlap. Although some studies argue that "relatedness" should be distinguished from "similarity" (Kumar and Kumar, 2024; Ranasinghe et al., 2025), it is unclear to what extent models are capable of discerning this distinction.

Feature Decomposition. Sentence similarity arises from multiple factors (e.g., content, style), yet much prior work reduces them to a single scalar score. Recent approaches move toward aspect-wise formulations: tasks that condition similarity on a specified criterion (Deshpande et al., 2023) or learn disentangled aspect representations (Schopf et al., 2023), and models that predict similarity along multiple-interpretable dimensions (Opitz and Frank, 2022; Sun et al., 2024; Tehenan et al., 2025).

These studies have enabled models to compute similarity based on multiple aspects and to make explicit what factors underlie their similarity judgments. However, the mechanisms of mainstream projection head architectures and their cross-dataset behaviors remain less understood. To address this gap, we propose a comparative framework for different pooler-based training paradigms. Using this framework, we analyze the extent to which these models capture diverse sentential aspects from the perspective of practical applications.

3 Methodology

We design our experimental framework to enable consistent comparison across datasets with different label scales. To this end, we adopt a bi-encoder (Reimers and Gurevych, 2019) architecture that embeds each sentence independently, keeping the encoder parameters frozen and training only a lightweight projection head (hereafter referred to as the pooler). Our approach shares the perspective of probing tasks in that it compares embedding representations from poolers trained on different datasets (Conneau et al., 2018), while extending the scope to sentence pairs.

3.1 Datasets

We collect 21 datasets spanning three different types: STS, Paraphrase identification (PI), and Triplet. Each dataset consists of sentence pairs or triplets designed to reflect varying notions of sentence similarity. Details and example instances are provided in Appendix A.

If a dataset is originally provided with predefined splits, we first merge all splits and then re-divide the data into 80% for training and 20% for testing, making sure that no identical sentence appears in both sets to prevent leakage from training. To ensure comparability across datasets, we cap the number of training samples at 5,000 by randomly subsampling when a dataset exceeds this size.

3.1.1 STS datasets

In this setting, the pooler predicts a continuous similarity score between a pair of sentences. Each dataset is represented as:

$$D_{i,\text{sts}} = \left\{ (s_1^{(k)}, s_2^{(k)}, y_k) \;\middle|\; y_k \in [0, 1] \right\}_{k=1}^{N},$$

where y_k denotes the similarity score of sentence pair $(s_1^{(k)}, s_2^{(k)})$. For each dataset, the similarity score is normalized so that it falls within the range of real values from 0 to 1. The following datasets fall under this category: STSB (Cer et al., 2017), SICK (Marelli et al., 2014), CxC (Parekh et al., 2021), STS3k (Fodor et al., 2025), Opusparcus (Creutz, 2018), AFS (Misra et al., 2016), BWS (Thakur et al., 2021), FinSTS (Yang et al., 2024), and SemRel2024 (Ousidhoum et al., 2024).

3.1.2 PI datasets

In this setting, the pooler performs binary prediction to determine whether two sentences convey the same meaning. The dataset structure is formalized as:

$$D_{i,\text{pi}} = \left\{ (s_1^{(k)}, s_2^{(k)}, y_k) \;\middle|\; y_k \in \{0, 1\} \right\}_{k=1}^{N},$$

where $y_k = 1$ indicates that the sentence pair $(s_1^{(k)}, s_2^{(k)})$ is semantically equivalent (i.e., paraphrases), and $y_k = 0$ otherwise. This category includes APT (Nighojkar and Licato, 2021), PARADE (He et al., 2020), Webis-CPC-11 (Burrows et al., 2013), AskUbuntu (Lei et al., 2016), PAWS-Wiki (Zhang et al., 2019), and QQP (Quora, 2017).

While all STS and PI datasets target semantic equivalence, the operational definition of meaning varies. For instance, Opusparcus emphasizes equivalence in conversational phrasing and slang, whereas AskUbuntu focuses on whether forum questions are functionally equivalent.

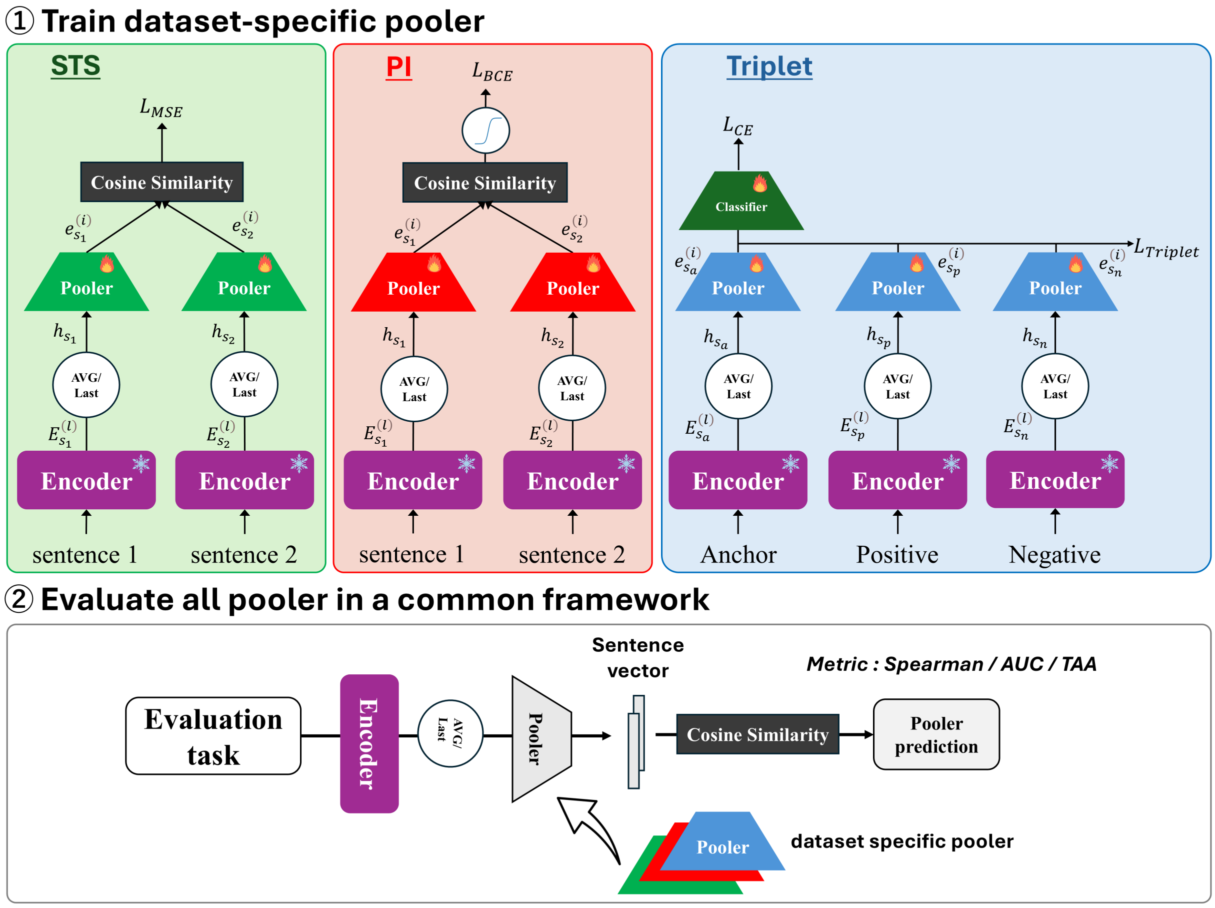


Figure 2: Training and inference paradigm of our framework. We freeze the encoder and train structurally identical poolers according to the supervision format of STS, PI and Triplet datasets. This setup allows us to replace only the pooler at inference time and apply cosine similarity based metrics, thereby enabling a unified evaluation across all poolers.

3.1.3 Triplet datasets

These datasets are composed of sentence triplets in the form of:

$$D_{i,\text{trip}} = \left\{ \left(s_a^{(k)}, s_p^{(k)}, s_n^{(k)}\right) \right\}_{k=1}^{N},$$

Each triplet consists of an anchor sentence $s_a^{(k)}$, a positive sentence $s_p^{(k)}$, and a negative sentence $s_n^{(k)}$. In our collection, the triplets are constructed such that (s_a, s_p) share stylistic features while differing in content, and (s_a, s_n) share content but differ in style. An exception is found in the ELSA-Emotion dataset, where (s_a, s_n) differ in both content and style. The following datasets belong to this category: ELSA (Gandhi and Gandhi, 2025), MTST (Mukherjee et al., 2023), Paradetox (Logacheva et al., 2022), APPDIA (Atwell et al., 2022), WNC (Pryzant et al., 2020), and StyleDistance (Patel et al., 2025). Since these datasets were originally constructed as parallel corpora, the (s_a, s_n) pairs were designed with the anchor's attribute fixed. For our purposes, however, we reshuffle the data so that the attributes of the anchors are evenly distributed, and we randomly sample a sentence with the same attribute as the anchor to serve as positive. In addition, although StyleDistance contains 40 types of stylistic features, each feature is represented by only 100 sentence pairs; therefore, we do not use it for training purposes.

3.2 Dataset-specific Pooler Mechanism

Figure 2 illustrates the three training paradigms: **(1) STS (2) PI (3) Triplet** and a common evaluation paradigm supported by our framework. When training the pooler on each respective dataset, the learning process is conducted in accordance with the format of the gold labels. However, during evaluation, a common methodology is employed regardless of the specific dataset used for training. In all cases, input sentences are first passed through a shared encoder to obtain token-level representations. The l-th encoder layer produces token em-

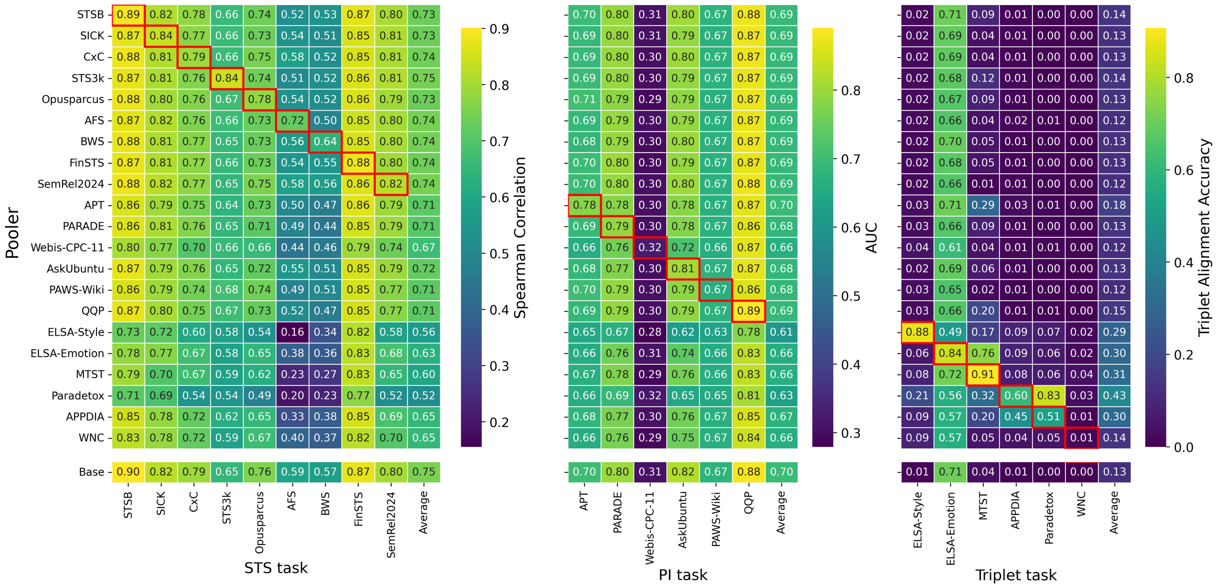


Figure 3: Performance of each pooler on STS, PI, and Triplet tasks. Each task applies a different metric, and the values within the cells represent the scores of each pooler (higher is better). ***Base*** denotes using the frozen encoder's embeddings without a learned pooler. The results of poolers trained and evaluated on the same dataset are highlighted with red boxes.

beddings for sentence s as $E_s^{(l)} \in \mathbb{R}^{T \times d_{\text{emb}}}$, where T is the number of tokens and d_{emb} is the embedding dimension.

To obtain a fixed-size sentence vector $\mathbf{h}_s \in \mathbb{R}^{d_{\text{emb}}}$, we apply either *average pooling* or *last token pooling* over $E_s^{(l)}$. This vector is then transformed by a **dataset-specific pooler** into a common-dimensional representation space:

$$\mathbf{e}_s^{(i)} = W_i\,\mathbf{h}_s + \mathbf{b}_i, \quad W_i \in \mathbb{R}^{d_p \times d_{\text{emb}}},\ \mathbf{b}_i \in \mathbb{R}^{d_p}, \tag{1}$$

where i indexes the dataset, and d_p is the projected embedding dimension. Thus, each dataset D_i has its own pooler parameters $(W_i, \mathbf{b}_i)$, allowing the pooler to learn dataset-specific representations.

(1) STS For each sentence pair $\left(s_1^{(k)}, s_2^{(k)}\right) \in D_{i,\text{sts}}$ with a target similarity score $y_k \in [0, 1]$, the predicted similarity is given by:

$$\hat{y}_k = \cos\left(\mathbf{e}_{s_1}^{(k)}, \mathbf{e}_{s_2}^{(k)}\right). \tag{2}$$

The pooler is trained using the mean squared error loss L_{MSE}. Evaluation is conducted using Spearman's rank correlation.

(2) PI For sentence pairs $\left(s_1^{(k)}, s_2^{(k)}\right) \in D_{i,\text{pi}}$ with binary label $y_k \in \{0, 1\}$, we apply a sigmoid activation to the cosine similarity:

$$\hat{y}_k = \sigma\left(\cos\left(\mathbf{e}_{s_1}^{(k)}, \mathbf{e}_{s_2}^{(k)}\right)\right), \tag{3}$$

with binary cross-entropy loss L_{BCE}. Pooler performance is evaluated by computing AUC (Area Under the ROC Curve).

(3) Triplet For each triplet $(s_a, s_p, s_n) \in D_{i,\text{trip}}$, we apply a margin-based triplet loss (Zhang et al., 2022):

$$L_{\text{Triplet}} = \max\big(0,\ m + \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_n}) - \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_p})\big), \tag{4}$$

where m is a margin hyperparameter. L_{Triplet} encourages the distance between the anchor and the positive to become smaller while pushing the anchor and the negative further apart. In addition, the projected anchor vector $\mathbf{e}_{s_a}$ is used to predict a style label via a softmax classifier:

$$\hat{y} = \text{softmax}(W_i'\mathbf{e}_{s_a} + \mathbf{b}_i'), \tag{5}$$

with corresponding cross-entropy loss L_{CE}. The total loss combines triplet loss and classification loss: $L_{\text{CE}} + \alpha\, L_{\text{Triplet}}$ where α is a balancing weight. Evaluation is based on the Triplet alignment accuracy (TAA) (Feng et al., 2025), which computes the proportion of $\cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_p}) > \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_n})$ samples in the test set.

As shown in Figure 2, the same pooler structure is used across task types but trained separately per dataset D_i. Once trained, the pooler W_i can be reused to extract representations for each task, enabling flexible comparison among the poolers.

3.3 Basic setup

We experiment with two backbones: the BERT-large-based model[2] and the Qwen3-based model[3]. We use average pooling for BERT and last-token pooling for Qwen3. The pooler's embedding dimension is set to $d_p = 256$. Although the BERT and Qwen3 models differ significantly in architecture and parameter size, no notable differences in trends were observed. Therefore, we primarily report the results of the BERT model in the main text. For the margin m and the weighting parameter α, we evaluate several combinations to identify effective settings. Unless otherwise specified in tables or figures, we use the last layer, $m = 0.2$, $\alpha = 1$, as the default values. A list of hyperparameters is provided in Appendix B.1. As a result, we obtained 21 types of poolers that enable a unified comparison based on cosine similarity.

4 Experiments

In this section, we leverage the proposed framework to analyze the following two questions:

- Does the pooler distinguish diverse types of sentence differences as distinct properties (Section 4.1)?
- How does the model capture the relationship between semantics and style (Section 4.2)?

4.1 Evaluation results under the proposed framework

Each pooler shares a common notion of semantic similarity across datasets. We begin by asking whether datasets that all aim to judge if two sentences have the same meaning actually define semantic similarity in different ways. Figure 3 compares STS, PI, and Triplet task performance for each pooler alongside the raw embeddings from the encoder model in Figure 2. The fact that the encoder model already performs well indicates that a basic concept of semantic difference is shared across these datasets. Although the performance gaps reveal that certain types of semantic distinctions are more difficult for the model to learn (e.g., AFS, BWS, Webis-CPC-11), poolers trained on STS and PI datasets generally achieved consistent scores even on evaluation datasets they were not explicitly trained for. To explore this further, we computed cosine similarity scores for every sentence pair in each dataset and measured the correlations between poolers. For poolers trained with STS and PI datasets, all pairwise correlations exceed 0.9 in BERT (Figure 9). These results suggest that while diverse semantic differences perceived by humans are captured by the model as universal and closely related concepts, the poolers for some datasets fail to sufficiently learn dataset-specific semantic distinctions, relying instead on a common criterion of semantic similarity for prediction. Such a tendency is consistent with the observation that embedding models tend to capture surface-level information (e.g., word overlap) rather than the implicit semantics of sentences (Sun et al., 2025). To enable the separation and representation of dataset-specific semantic differences, more delicate and sophisticated approaches are likely required.

Style is a concept clearly distinct from semantics. On the right side of Figure 3, corresponding to the Triplet task, the results differ from those of the STS and PI tasks. In general, poolers trained on datasets other than the evaluation dataset itself achieved substantially lower scores. This indicates that each style represents a concept distinct from both semantics and other styles. We further conducted analyses by varying the margin and the weighting parameter α in the triplet loss (see Appendix C.2 for details). Through these experiments, we found that disentangling style from semantics contributes to improved classification accuracy, and that even without the triplet loss, learning to capture style progressively drives the model's predictions away from those of the original. This finding suggests that the embedding space of the base model primarily encodes semantic information, and that substantial transformations of this space are required to extract stylistic representations effectively.

4.2 Semantic-Style Relation

So far, we have treated semantics and style separately. Here, we look more closely at how they differ and interact.

4.2.1 Orthogonal Procrustes Distance

In analyzing the differences in tendencies across poolers, evaluation based on predicted similarity provides an intuitive and interpretable perspective. However, such analyses remain correlation-based and cannot directly assess causal effects. To address this limitation, we instead conduct a more direct examination by focusing on the weights and biases of the linear transformations within each pooler. Specifically, we measure the distance be-

[2] We use `mxbai-embed-large-v1` (Lee et al., 2024).

[3] We use `Qwen3-Embedding-8B` (Zhang et al., 2025).

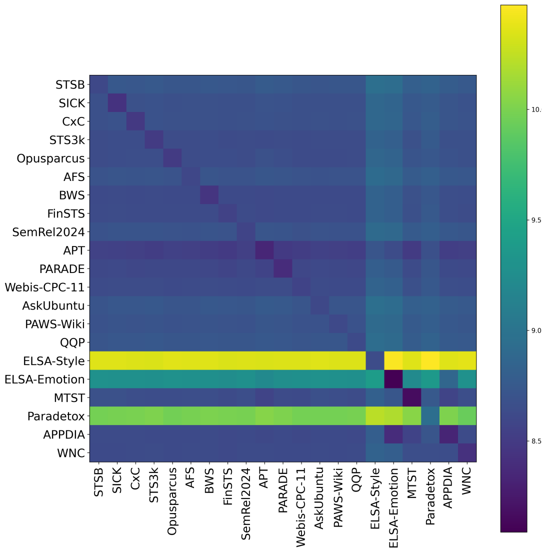


Figure 4: Orthogonal Procrustes distance between BERT poolers.

tween the mappings of the pooler i and j using the Orthogonal Procrustes distance, defined as follows:

$$d_{\mathrm{Proc}}\big((W_i, b_i), (W_j, b_j)\big) \\ := \min_{P \in \mathcal{O}(d_p)\ a>0} \big(|W_i - aPW_j|_F^2 + |b_i - aPb_j|_2^2\big) \,, \tag{6}$$

where P denotes a d_p-dimensional orthogonal matrix and a represents a positive scalar. This metric computes the rotation- and scaling-invariant difference between two linear transformations; a larger distance implies a greater divergence in the predicted cosine similarity of the transformed embeddings. Note that P and a are analytically and uniquely determined for the parameters of each pooler pair (the proof is provided in Appendix A). Using Equation (6), we compute the distances between the linear transformations of different BERT poolers. For each pooler, we conducted five independent training runs with different random seeds, and then measured the average orthogonal Procrustes distance both within the same dataset and across different datasets. The results are presented in Figure 4.

We observe that poolers trained on STS datasets and those trained on PI datasets exhibit nearly the same orthogonal Procrustes distances when compared across different random seeds within the same dataset and across different datasets. These results indicate that, despite being trained on different datasets, the obtained parameters exhibit negligible differences. In contrast, poolers trained on the Triplet dataset possess mutually distinct mappings, demonstrating the uniqueness of their linear transformations.

4.2.2 Layer-wise analysis

Dataset	Best		Worst	
	Layer	Score	Layer	Score
Metric: Spearmanr				
STSB	23	0.891	0	0.760
SICK	23	0.838	7	0.741
CxC	23	0.787	13	0.689
STS3k	21	0.813	3	0.642
Opusparcus	20	0.783	0	0.657
AFS	23	0.721	0	0.607
BWS	22	0.645	13	0.557
FinSTS	23	0.884	0	0.841
SemRel2024	23	0.823	0	0.733
Metric: AUC				
APT	23	0.778	0	0.748
PARADE	23	0.794	3	0.761
Webis-CPC-11	17	0.389	2	0.311
AskUbuntu	23	0.809	13	0.637
PAWS-Wiki	23	0.699	0	0.488
QQP	23	0.885	0	0.777
Metric: TAA				
ELSA-Emotion	23	0.841	0	0.709
ELSA-Style	15	0.948	23	0.879
MTST	20	0.918	0	0.578
Paradetox	9	0.917	23	0.827
APPDIA	12	0.635	0	0.406
WNC	17	0.162	23	0.010

Table 1: Best and worst layer performance across datasets. See Appendix C.3 for more detailed results.

Prior research has established that embedding models encode distinct information at different layers (Ma et al., 2019; Oh et al., 2022). To investigate this, we trained a pooler on every layer of the encoder model and empirically found that the optimal-performing layer varies across datasets. Table 1 reports the best-performing and worst-performing layers for each dataset.

For datasets related to semantic similarity, the highest performance is typically found in the later layers, while the worst performance occurs in the early to middle layers. We hypothesize that this outcome arises because semantic closeness between sentences depends on contextual information. Prior studies have shown that the early layers of BERT capture syntactic structure, whereas its later layers encode context (Tenney et al., 2019; Turton et al., 2021). Therefore, we consider that using embeddings from those later layers allowed us to most effectively detect semantic closeness between sentences.

In contrast, style-related datasets show a different trend. For example, in sentiment analysis

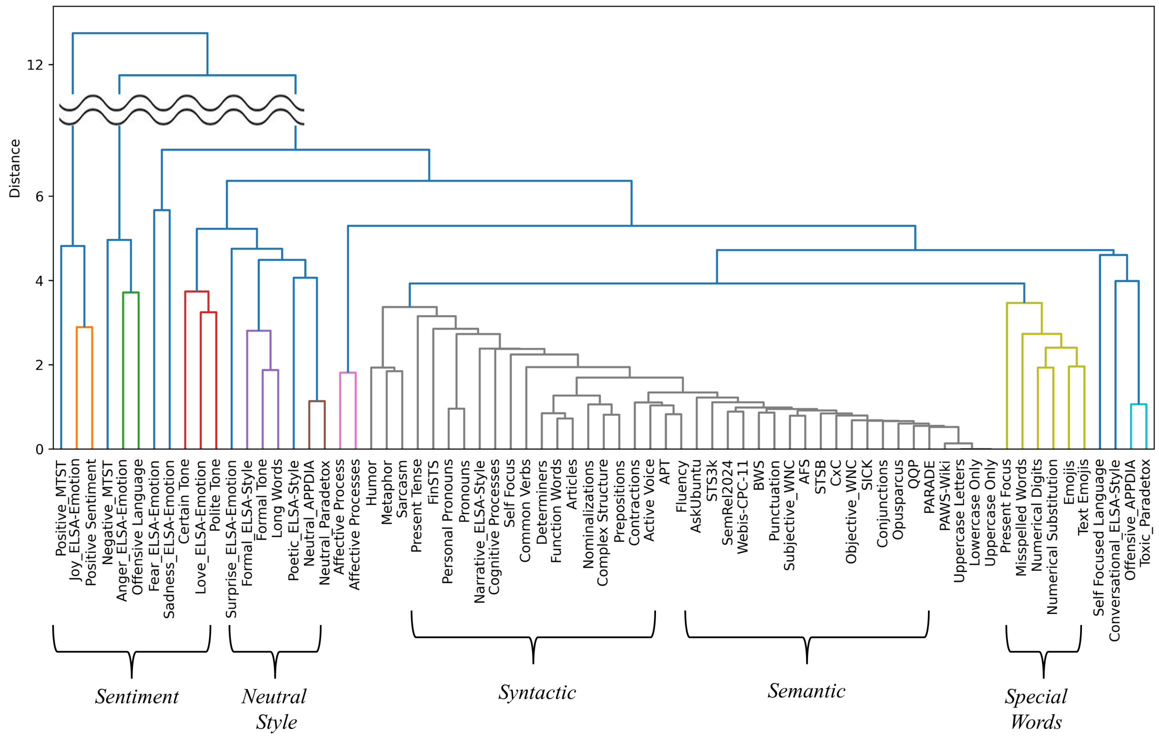


Figure 5: Dendrogram of the hierarchical clustering of Mean Difference vectors computed by the encoder model. The merged clusters exhibit clear commonalities, and the clustering appears to reflect broader, more abstract linguistic properties rather than specific types of semantic differences.

datasets like ELSA-Emotion and MTST, the later layers still produce the best results. However, when it comes to other types of stylistic features such as toxicity, the way of speaking, and subjectivity, the final layer often performs the worst. We interpret this result not as evidence that the middle layers capture stylistic features particularly well, but rather that such features become more difficult to extract in the later layers, where information is increasingly contextualized. Since this contextualization process is inherently unavoidable, the finding suggests the importance of optimizing layer selection according to the specific characteristics of each task.

4.2.3 Aspect dendrogram

Prior research has shown that language models organize knowledge in ways that differ from human understanding (Dalvi et al., 2022). However, how these models encode *abstract* or *stylistic* concept differences remains unclear.

To explore this, we analyze the relationships between aspects using collected datasets. Our goal is to extract a representative vector for each dataset that captures its characteristic conceptual differences.

Step 1: Sampling Dissimilar Sentence Pairs

We extract K sentence pairs from each dataset that differ in meaning, using different strategies depending on the dataset type. For STS datasets, we select the K least similar pairs as $D^{-}_{i,\text{sts}} = \{(s_1^{p(k)}, s_2^{p(k)})\}_{k=1}^{K}$. For PI datasets, we randomly sample K sentence pairs labeled as dissimilar, i.e., $D^{-}_{i,\text{pi}} = \{(s_1^{(k)}, s_2^{(k)}) \mid y_k = 0\}_{k=1}^{K}$. For Triplet datasets, we randomly select K Anchor–Negative pairs: $D^{-}_{i,\text{trip}} = \{(s_a^{(k)}, s_n^{(k)})\}_{k=1}^{K}$. The combined set of dissimilar pairs is denoted as $D^{-}_{i} = D^{-}_{i,\text{sts}} \cup D^{-}_{i,\text{pi}} \cup D^{-}_{i,\text{trip}}$. We used $K = 100$ for StyleDistance, while $K = 1000$ was used for the other datasets.

Step 2: Computing Mean Difference Vectors

To extract concepts, we adopt the *Mean Difference* (MD) approach, which is commonly used in language model steering (Rimsky et al., 2024; Wu et al., 2025). We compute a MD vector for each dataset using sentence embeddings from the final layer of the pretrained model:

$$h_{\text{MD}_i} := \frac{1}{K} \sum_{n \in D^{-}_{i}} \left(h_{s_1}^{(n)} - h_{s_2}^{(n)} \right). \qquad (7)$$

Intuitively, this vector represents the aspect of semantic or stylistic difference that is consistently expressed within the dataset. Other unrelated elements are expected to be canceled out by averaging. Since this vector is asymmetric, swapping s_1 and s_2 changes the result. For triplet datasets except for StyleDistance, we compute separate vectors for each anchor direction accordingly.

Step 3: Hierarchical Clustering

Finally, we apply hierarchical clustering to all computed MD vectors using Ward's method (Murtagh and Legendre, 2014). The resulting dendrogram visualizes the relationships among different types of conceptual differences captured by the model.

Figure 5 presents the dendrogram of the clustering results. An enlarged version of the graph is provided in Appendix C.4. Each cluster represents a group of MD vectors that share similar attributes, and merging occurs in order of increasing distance between vectors. In the dendrogram, vectors capturing semantic differences are merged first, followed by those representing syntactic, special word, style, and finally sentiment differences. They indicate that while semantic differences span a wide range of perspectives, they form a relatively cohesive cluster from the model's point of view. This implies that semantic differences are conceptually close within the embedding space. In contrast, differences in style and sentiment are captured as concepts that are substantially distinct from semantics. Given this behavior of encoder models, it may be more effective to treat meaning, style, and emotion as separate components. Their conceptual independence should be taken into account when designing model architectures.

5 Discussion and Conclusion

We focus on semantic and stylistic properties of sentences, training separate poolers using a diverse set of datasets and analyzing the embedding performance and interrelations of each pooler. In addition, our experiments reveal that there is substantial room for improvement in tasks that involve assessing the similarity of claims and opinions on social issues (AFS and BWS), long passages (Webis-CPC-11) and classification of subjectivity and objectivity (WNC). Since these tasks are not supported by MTEB (Muennighoff et al., 2023), a representative benchmark for evaluating embedding models, there are relatively few studies on them, indicating the need for more detailed analyses in future work.

The high correlation observed between the predictions of poolers trained on the STS and PI datasets in this experiment may be attributed to the anisotropy issues inherent in embedding models (Gao et al., 2019; Ethayarajh, 2019). Because the model constructs an anisotropic embedding space, semantic distinctions can only be represented within very small distances. Consequently, regardless of the dataset used for training, the model tends to learn nearly identical weights, leading to highly similar predictions. Indeed, as shown in Section 4.2.1, our analysis using orthogonal Procrustes distance demonstrates that differences across datasets are scarcely reflected in changes to the pooler weights.

Focusing on fine-grained aspects of sentence meaning and style offers not only improvements in interpretability and transparency, but also practical benefits. These include the potential to enhance model performance by considering interactions among different aspects, and to measure dataset diversity in terms of fine-grained linguistic properties, as discussed in Appendix E. We hope that our study contributes to the advancement of both the utility and interpretability of embedding models.

Limitations In this study, we employed only two types of encoder models, which limits the scope of our analysis regarding model architectures. While we believe that our findings and insights are likely model-agnostic, experiments with a broader range of models are necessary for more rigorous verification. Moreover, all datasets used in this study are in English. How differences in language are embedded by the model, and to what extent such differences interact with other factors, remains an open question for future investigation.

Ethical Considerations To analyze how embedding models capture stylistic features of text, we utilized APPDIA and ParaDetox, two datasets that contain toxic language collected from Reddit. As the embedding models used in this study are not capable of text generation, the models trained in our experiments do not pose a direct risk of producing harmful content. Clarifying how such concepts are embedded contributes to a better understanding and control of model behavior, which can support the development of safer and more transparent language technologies. We used AI tools for proofreading and mathematical typesetting in the preparation of this paper. All content has been thoroughly reviewed and verified by the human authors.

Acknowledgments

This work was supported by Research and Development Center for Large Language Model/National Institute of Informatics, and JSPS KAKENHI Grant Number 24K03231.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase Acquisition via Crowdsourcing and Machine Learning. *Transactions on Intelligent Systems and Technology (ACM TIST)*, 4(3):43:1–43:21.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*.

Preetam Prabhu Srikar Dammu and Omar Alonso. 2024. Near-duplicate question detection. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 493–496, New York, NY, USA. Association for Computing Machinery.

Ameet Deshpande, Carlos Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik Narasimhan. 2023. C-STS: Conditional semantic textual similarity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5669–5690, Singapore. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Yingchaojie Feng, Yiqun Sun, Yandong Sun, Minfeng Zhu, Qiang Huang, Anthony Kum Hoe Tung, and Wei Chen. 2025. Don't reinvent the wheel: Efficient instruction-following text embedding based on guided space transformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24511–24525, Vienna, Austria. Association for Computational Linguistics.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2025. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190.

Vishal Gandhi and Sagar Gandhi. 2025. Elsa: A style aligned dataset for emotionally intelligent language generation. *Preprint*, arXiv:2504.08281.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7572–7582, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2023. Text style transfer: A review and experimental evaluation. *Preprint*, arXiv:2010.12742.

Hiroki Iida and Naoaki Okazaki. 2021. Incorporating semantic textual similarity and lexical matching for information retrieval. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 582–591, Shanghai, China. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. *Preprint*, arXiv:2004.11362.

Anand Kumar and Hemanth Kumar. 2024. scaLAR SemEval-2024 task 1: Semantic textual relatednes for English. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 902–906, Mexico City, Mexico. Association for Computational Linguistics.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embedding model.

Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5969–5979, Dublin, Ireland. Association for Computational Linguistics.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1279–1289, San Diego, California. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from bert: An empirical study. *arXiv preprint arXiv:1910.07973*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

David D. McDonald and James D. Pustejovsky. 1985. A computational theory of prose style for natural language generation. In *Second Conference of the European Chapter of the Association for Computational Linguistics*, Geneva, Switzerland. Association for Computational Linguistics.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Lucas Moeller, Dmitry Nikolaev, and Sebastian Padó. 2024. Approximate attributions for off-the-shelf Siamese transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2059–2071, St. Julian's, Malta. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Sourabrata Mukherjee, Akanksha Bansal, Pritha Majumdar, Atul Kr. Ojha, and Ondřej Dušek. 2023. Low-resource text style transfer for Bangla: Data & models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 34–47, Singapore. Association for Computational Linguistics.

Fionn Murtagh and Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7106–7116, Online. Association for Computational Linguistics.

Dongsuk Oh, Yejin Kim, Hodong Lee, H Howie Huang, and Heuiseok Lim. 2022. Don't judge a language model by its last layer: Contrastive learning with layer-wise attention pooling. *arXiv preprint arXiv:2209.05972*.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Juri Opitz, Lucas Möller, Andrianos Michail, and Simon Clematide. 2025. Interpretable text embeddings and text similarity explanation: A primer. *Preprint*, arXiv:2502.14862.

Nedjma Ousidhoum, Shamsuddeen Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Ahmad, Sanchit Ahuja, Alham Aji, Vladimir Araujo, Abinew Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine Kock, Genet Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, and 8 others. 2024. SemRel2024: A collection of semantic textual relatedness datasets for 13 languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2512–2530, Bangkok, Thailand. Association for Computational Linguistics.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2855–2870, Online. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Quora. 2017. Quora question pairs. `https://www.kaggle.com/datasets/quora/question-pairs-dataset`.

Muhammad Rafi and Mohammad Shahid Shaikh. 2013. An improved semantic similarity measure for document clustering based on topic maps. *Preprint*, arXiv:1303.4087.

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025. Musts: Multilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Tim Schopf, Emanuel Gerber, Malte Ostendorff, and Florian Matthes. 2023. AspectCSE: Sentence embeddings for aspect-based semantic textual similarity using contrastive learning and structured knowledge. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1054–1065, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony Kum Hoe Tung, and Jun Yu. 2024. A general framework for producing interpretable semantic text embeddings. *ArXiv*, abs/2410.03435.

Yiqun Sun, Qiang Huang, Anthony KH Tung, and Jun Yu. 2025. Text embeddings should capture implicit semantics, not just surface meaning. *arXiv preprint arXiv:2506.08354*.

Matthieu Tehenan, Vikram Natarajan, Jonathan Michala, Milton Lin, and Juri Opitz. 2025. Mechanistic decomposition of sentence representations. *Preprint*, arXiv:2506.04373.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Jacob Turton, Robert Elliott Smith, and David Vinson. 2021. Deriving contextualised semantic features from BERT (and other transformer model) embeddings. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 248–262, Online. Association for Computational Linguistics.

Alexandros Vasileiou and Oliver Eberle. 2024. Explaining text similarity in transformer models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7859–7873, Mexico City, Mexico. Association for Computational Linguistics.

Hongwei Wang, Hongming Zhang, and Dong Yu. 2023. On the dimensionality of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10344–10354, Singapore. Association for Computational Linguistics.

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 *Forty-second International Conference on Machine Learning*.

Shanshan Yang, Steve Yang, and Feng Mai. 2024. Financial semantic textual similarity: A new dataset and model. In *2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, pages 1–8.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903, Dublin, Ireland. Association for Computational Linguistics.

A Description of the Dataset

Table 6 presents the source domains, sample sizes, and text lengths of the datasets used in this study. Example instances sampled from each dataset are shown in Tables 7,8 and 9.

B Implementation details

Parameter	**BERT**	**Qwen3**
Base Model	mixedbread-ai/ mxbai-embed-large-v1	Qwen/ Qwen3-Embedding-8B
Torch dtype	float16	bfloat16
Epoch	10	
Learning Rate	1e-4	
Train Batch Size	64	
Random Seed	42	
Hidden size d_{emb}	1024	4096
Pooler dimension d_p	256	256
Number of layers	24	36
Pooling strategy	average pooling	last token pooling

Table 2: Hyperparameters used to train BERT and Qwen3

B.1 Hyperparameters for training poolers

[4] The configurations of the encoder models used to train the poolers, along with the corresponding hyperparameters, are presented in Table 2. To determine the optimal learning rate, we conducted preliminary experiments using the following candidate values: {1e-2, 1e-3, 1e-4, 1e-5, 1e-6}. For the dimensionality of the pooler, we explored a range of values: {8, 32, 64, 128, 256, 512, 1024}. While the optimal learning rate varied considerably depending on the specific hyperparameter setting, the pooler dimensionality had minimal impact on performance. This observation is consistent with findings from prior work (Wang et al., 2023). Furthermore, our preliminary experiments with full fine-tuning revealed that updating the encoder parameters yielded comparable accuracy to the frozen encoder approach.

B.2 Discussion on Training Design

Our framework prioritizes fairness in comparison and adherence to conventional designs, aiming to ensure that performance differences across poolers

[4] Our experimental code was developed with reference to the implementation available at https://github.com/princeton-nlp/c-sts

Sentence 1	Sentence 2	Dataset	Gold	CxC	Opus	BWS
A pole that has a couple of signs on it	A stop sign with street signs labeling an intersection.	CxC	0.80	0.26	0.62	0.61
Well, you are most welcome.	Don't mention it.	Opusparcus	0.83	0.20	0.17	0.28
Regardless of what laws are in place, there will always be young women out there who want to have an abortion.	Banning abortion risks illegal abortions.	BWS	0.77	0.56	0.69	0.50

Table 3: Case study of predicted similarities by different poolers. "Gold" denotes the ground-truth similarity score.

stem from the datasets themselves rather than from the loss functions. Consequently, optimal strategies for maximizing a pooler's performance may differ from the settings employed in this study. In our preliminary experiments, we observed that directly applying MSE without a sigmoid activation improved accuracy for PI datasets and using Supervised Contrastive Loss (Khosla et al., 2021) yielded higher classification accuracy for Triplet datasets. However, we excluded these configurations from the main study because these loss functions significantly alter the prediction tendencies of the poolers, which would obscure the dataset-driven differences we aim to analyze.

C Experiments Results

C.1 Pooler performance

Figures 7 and 8 present the performance of poolers trained using the embedding models of BERT and Qwen3, respectively. Despite the substantial differences in parameter size and model architecture between BERT and Qwen3, the results exhibit remarkably similar trends. This suggests that the types of semantic distinctions learned by the poolers are highly consistent across datasets, indicating that such behavior is not model-specific but rather a general phenomenon.

C.2 Training step analysis on Triplet task

To further investigate how the distance relationships among sentences within a triplet affect the pooler's predictions, we experimented with various parameter settings in the triplet loss defined in Equation (4) and analyzed the resulting training dynamics of the pooler. We fix $margin : m \neq 0$ beforehand and select exactly one of the following (no switching).

$$L_{\text{Triplet}} = \begin{cases} \max\Big(0,\ m + \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_n}) - \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_p})\Big), & m > 0, \\ \max\Big(0,\ |m| + \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_p}) - \cos(\mathbf{e}_{s_a}, \mathbf{e}_{s_n})\Big), & m < 0. \end{cases} \tag{8}$$

When the margin is set to positive, the $L_{Triplet}$ encourages the distance between the anchor and the positive to become smaller while pushing the anchor and the negative further apart. Conversely, when the margin is set to negative, it works in the opposite direction. Figures 11 and 12 show the F1 score for the attribute classification of anchor sentences, the Spearman's correlation with the cosine similarity calculated by the encoder model, and the TAA for each pooler at training step. When the margin is negative, both TAA and F1 scores are low, whereas a positive margin leads to higher values for both metrics. This indicates that the more strongly semantic and stylistic information are disentangled in the embedding space, the higher the accuracy of style classification becomes while the resulting predictions deviate further from those of the encoder model. The effect of the triplet loss is more pronounced in Qwen3. When $\alpha = 0$, i.e., when the model is trained solely with the cross entropy loss for anchor label classification, it tends to fall into a local optimum. These results suggest that even for style classification alone, considering the relationship between style and semantics is crucial for effective learning.

C.3 Layer-wise analysis

Figure 13, 14 show the results of training the pooler on all layers of the encoder model. While the poolers from the later layers are more advantageous for semantic and sentiment features, the middle layers are superior for the other stylistic features.

C.4 Aspect dendrogram

Figure 15, 16 show the results of the aspect dendrogram constructed using the method described in Section 4.2.3. The merged datasets exhibit interpretable characteristics, and semantic differences are integrated at an early stage of the hierarchy.

Model	STSB	SICK	CxC	STS3k	Opus	AFS	BWS	FinSTS	SemRel	APT	PARADE	Webis	AskU	PAWS	QQP
Pooler	0.90	0.84	0.87	0.84	0.68	0.74	0.67	0.92	0.81	0.79	0.73	0.37	0.83	0.73	0.90
GPT-5-mini	0.87	0.86	0.85	0.86	0.76	0.65	0.66	0.89	0.77	0.70	0.63	0.53	0.69	0.89	0.78

Table 4: Scores on STS and PI datasets using few-shot reasoning with GPT-5-mini. We sampled 100 instances from the test set for each task. Consequently, the pooler performance reported here differs slightly from the full evaluation in the main experiments.

C.5 Case Study

We conducted an error analysis on cases where the poolers failed to capture dataset-specific semantics. Table 3 presents sample instances from each dataset alongside the similarity scores predicted by each pooler. The criteria for similarity differ fundamentally across these datasets: CxC measures proximity in the context of image captions, Opusparcus focuses on colloquial equivalence, and BWS assesses the closeness of opinion or stance. Observing the predictions, it appears unlikely that the poolers differentiate between these diverse and subtle semantic criteria in the way humans do. Instead, the results suggest that the models rely primarily on general textual features rather than capturing dataset-specific semantic nuances.

C.6 Comparison with Reasoning-Capable LLMs

To investigate whether the similarity prediction tendencies of our pooler differ from the outputs of reasoning-capable LLMs, we conducted similarity prediction experiments on the STS and PI datasets using GPT-5-mini. We provided the model with four examples randomly sampled from the training dataset and instructed it to generate an explanation for its reasoning before predicting a score (a continuous value between 0–5 for STS, and a binary label for PI).

Table 4 presents the results for 100 samples randomly collected from each dataset, comparing the scores of GPT-5-mini with those of the pooler trained on the corresponding dataset. While GPT-5-mini achieved an average performance comparable to our pooler, we observed distinct differences in performance trends across datasets. Reasoning-capable LLMs appear to have an advantage on datasets like PAS-Wiki, where semantic differences follow systematic regularities. Conversely, encoder models (such as our pooler), which are pre-trained on massive query-document datasets, demonstrate superior performance on question equivalence tasks such as AskUbuntu and QQP.

D Orthogonal Procrustes Distance

D.1 Definition of distance

In this section, we mathematically analyze the differences across poolers by focusing on the weights and biases of their linear transformations, rather than relying on scores or correlations. Gaining an understanding of these weights and biases offers practical advantages, as manipulating the linear transformation enables controlled interventions in the pooler's predictions.

As shown in Equation (9), each pooler learns its own weight matrix and bias vector:

$$\mathbf{e}_s^{(i)} = W_i \, \mathbf{h}_s + \mathbf{b}_i, \quad W_i \in \mathbb{R}^{d_p \times d_{\text{emb}}}, \ \mathbf{b}_i \in \mathbb{R}^{d_p}, \tag{9}$$

where $W_i \in \mathbb{R}^{d_p \times d_{emb}}$ and $b_i \in \mathbb{R}^{d_p}$.

The most straightforward approach to comparing these weights is to compute the Frobenius norm of their difference. The Frobenius norm is computationally simple and widely used to quantify weight changes in transfer learning or to evaluate discrepancies in model merging. However, the Frobenius norm is sensitive to transformations irrelevant to similarity computations—specifically, it varies under orthogonal rotations and positive scalar rescalings, which leave cosine similarity invariant.

For instance, consider applying an arbitrary orthogonal matrix P and a positive scalar a to the weight matrix W and bias b, yielding transformed parameters

$$W' = aPW, \quad b' = aPb \tag{10}$$

For a hidden representation e_s, the inner product in the numerator transforms as

$$\begin{aligned} \langle e'_{s_1}, e'_{s_2} \rangle &= \langle aPe_{s_1}, aPe_{s_2} \rangle \\ &= (aPe_{s_1})^\top (aPe_{s_2}) \\ &= a^2 e_{s_1}^\top P^\top P e_{s_2} \\ &= a^2 \langle e_{s_1}, e_{s_2} \rangle, \end{aligned} \tag{11}$$

where we used $P^\top P = I$.

The denominator, i.e., the norm, transforms as

$$\begin{aligned}\|e'_s\| &= \|aPe_s\| \\ &= \sqrt{(aPe_s)^\top(aPe_s)} \\ &= \sqrt{a^2 e_s^\top P^\top P e_s} \\ &= a\|e_s\|.\end{aligned} \tag{12}$$

Thus, the cosine similarity under the transformed parameters satisfies

$$\begin{aligned}\cos_{W',b'}(e_{s_1}, e_{s_2}) &= \frac{\langle e'_{s_1}, e'_{s_2}\rangle}{\|e'_{s_1}\|\|e'_{s_2}\|} \\ &= \frac{a^2\langle e_{s_1}, e_{s_2}\rangle}{(a\|e_{s_1}\|)(a\|e_{s_2}\|)} \\ &= \cos_{W,b}(e_{s_1}, e_{s_2}).\end{aligned} \tag{13}$$

Therefore, cosine similarity remains invariant under simultaneous application of an orthogonal transformation P and a positive rescaling factor a to (W, b). Consequently, before computing Frobenius distances, we first apply Procrustes alignment to eliminate differences that arise purely from rotation and scaling, thereby ensuring that the resulting metric reflects only the differences relevant to similarity computation. To determine the optimal rotation and scaling, we solve the following optimization problem explanation:

$$\min_{a>0,\, P\in O(d_p)} \|W_i - aPW_j\|_F^2, \tag{14}$$

where $a > 0$ is a positive scaling factor, $P \in O(d_p)$ denotes an orthogonal matrix, and $\|\cdot\|_F$ represents the Frobenius norm. Expanding Equation (14), we obtain

$$\begin{aligned}&\|W_i - aPW_j\|_F^2 \\ &= \|W_i\|_F^2 + a^2\|W_j\|_F^2 - 2a\,\mathrm{tr}\Big(PW_jW_i^\top\Big).\end{aligned} \tag{15}$$

The optimal P is obtained by maximizing the third term on the right-hand side of Equation (15). Specifically, we transform this term as follows:

$$\begin{aligned}\mathrm{tr}(PW_jW_i^\top) &= \mathrm{tr}(PU\Sigma V^\top) \\ &= \mathrm{tr}(\Sigma V^\top PU),\end{aligned} \tag{16}$$

where $W_jW_i^\top$ is decomposed as $U\Sigma V^\top$ by singular value decomposition (SVD). Since $V^\top PU$ is the product of orthogonal matrices, it is itself an orthogonal matrix. Letting $X = V^\top PU$, we obtain

$$\mathrm{tr}(\Sigma V^\top PU) = \mathrm{tr}(\Sigma X) = \sum_k \sigma_{kk}x_{kk}, \tag{17}$$

where $\sigma_{kk} \geq 0$ are the singular values. It is evident that this term is maximized when X is the identity matrix. Therefore, since $V^\top PU = I$, we obtain

$$P = VU^\top. \tag{18}$$

We substitute the optimal $P = VU^\top$ into Equation (15), and

$$\begin{aligned}&\mathrm{tr}(PW_jW_i^\top) = \mathrm{tr}(VU^\top U\Sigma V^\top) \\ &= \mathrm{tr}(VV^\top\Sigma) = \mathrm{tr}(\Sigma) = \sum_k \sigma_{kk}.\end{aligned} \tag{19}$$

The objective reduces to a uni-variate quadratic in a:

$$f(a) = \|W_i\|_F^2 + a^2\|W_j\|_F^2 - 2a\sum_k \sigma_{kk}. \tag{20}$$

Taking the derivative and setting it to zero yields

$$\begin{aligned}\frac{df}{da} &= 2a\|W_j\|_F^2 - 2\sum_k \sigma_{kk} = 0 \\ \Longrightarrow \quad a &= \frac{\sum_k \sigma_{kk}}{\|W_j\|_F^2}.\end{aligned} \tag{21}$$

Since σ contains nonnegative singular values, $a \geq 0$. Because

$$\frac{d^2f}{da^2} = 2\|W_j\|_F^2 > 0, \tag{22}$$

the stationary point is a strict minimizer. Consequently, the optimal scaling a is uniquely determined; together with $P = VU^\top$, this yields the generically unique solution to the Procrustes problem. We apply the same scaling and rotation to the bias term and define a distance between poolers:

$$\begin{aligned}&d_{\mathrm{Proc}}\big((W_i, b_i), (W_j, b_j)\big) \\ &:= \|W_i - aPW_j\|_F^2 + \|b_i - aPb_j\|_2^2.\end{aligned} \tag{23}$$

This metric removes discrepancies due solely to rotation and uniform rescaling, ensuring that the measured distance reflects differences that are relevant to similarity computation.

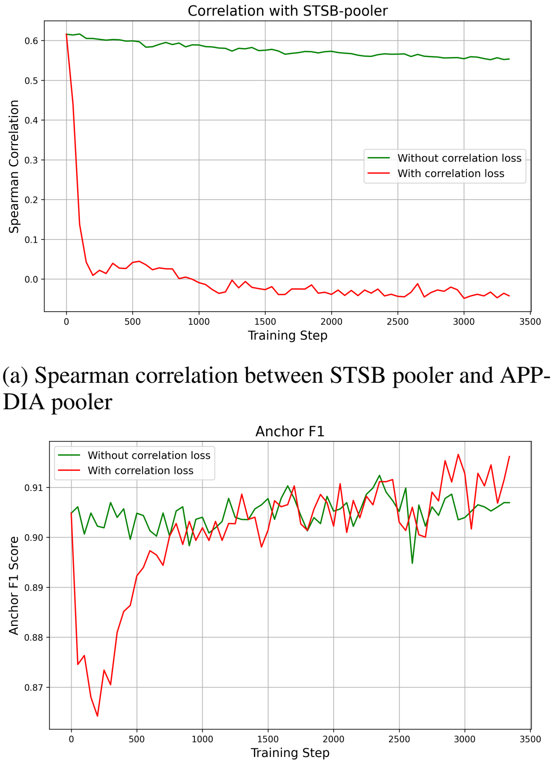
E Directions for Future Research

E.1 Learning with Consideration of Inter-Aspect Relationships

In training language models, it is common to incorporate a penalty term into the loss function that measures the distance between the model's predicted distribution and the original data distribution, in order to prevent the model from deviating excessively from the latter. This practice contributes to model stability. Conversely, when the objective is to encourage the model to diverge from the original distribution, the distance may be treated as a reward and integrated accordingly. A similar approach is applicable to our study. As shown in Section 4, semantic and style are found to be distinct concepts from the model's perspective. Taking this into account, we design a loss function that explicitly encourages low correlation between poolers.

$$L_{corr} = |\beta - \text{spearmanr}(\hat{y}^{(i)}_{batch}, \hat{y}^{(j)}_{batch})|, \quad (24)$$

where i and j denote different poolers. A penalty is imposed when the Spearman's correlation between the cosine similarities of their predictions over mini-batch pairs deviates from a user-defined target value β. Figure 6 presents the results of continued training with an additional correlation loss that enforces $\beta = 0$, applied to the STSB pooler and the APPDIA pooler. When training is guided solely by the standard loss based on agreement with the supervision labels, the correlation between poolers exceeds 0.5. However, with the correlation loss introduced, the Spearman correlation drops significantly, and the APPDIA pooler ultimately achieves improved accuracy. This suggests that explicitly encouraging separation between semantic and style representations can lead to better predictive performance. Although identifying the optimal relational constraint β is challenging due to the vast number of possible inter-pooler combinations, promoting independence or affinity between poolers can contribute to both accuracy and interpretability.



(a) Spearman correlation between STSB pooler and APPDIA pooler

(b) APPDIA pooler performance

Figure 6: Dynamics of continued training for the APPDIA pooler. The green line shows the case where training is continued normally. The red line shows the case where a penalty term based on the correlation with the STSB pooler—is added to the APPDIA pooler's loss function.

E.2 Applications for Analyzing Dataset Diversity

Models that capture diverse types of semantic differences are useful for analyzing the diversity of datasets. While it is desirable for training and evaluation datasets to cover a wide range of domains, conventional embedding models may fail to fully capture such diversity. We compute the *Diversity score* for a given dataset as follows.

$$\text{Diversity score} := 1 - \frac{1}{N}\sum_{i,j}^{N} \cos(e_{s_i}, e_{s_j}). \quad (25)$$

Here, i and j denote a randomly sampled pair of sentences, and the resulting value reflects the degree of variability in the dataset as perceived by the model. We set $N = 10,000$ and computed *Diversity score* using each pooler for the four major categories of the MMLU dataset (Hendrycks et al., 2021). The results are presented in Table 5. While the base model shows little variation in diversity across categories, the poolers trained in this study are capable of capturing diversity at a more fine-grained level. The Humanities category includes many questions related to argumentation and ethics, which are effectively captured by the AFS pooler. These observations suggest that training poolers to capture differences from more specific perspectives may be useful for dataset filtering.

Pooler	MMLU Category			
	Humanities	Social Sciences	STEM	Other
Base	0.391	0.403	0.410	**0.417**
AFS	**0.699**	0.661	0.618	0.648
FinSTS	0.579	0.610	0.589	**0.624**
AskUbuntu	0.580	0.579	**0.639**	0.621

Table 5: Diversity scores for the prompts in each category of the MMLU dataset.

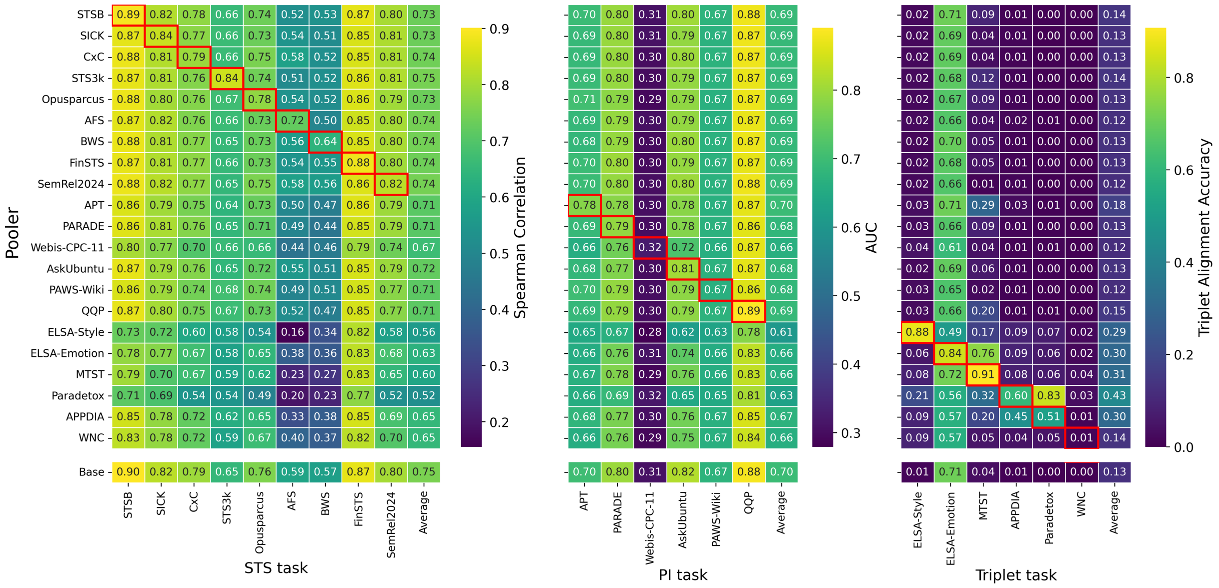


Figure 7: Performance of each BERT pooler on all evaluation tasks.

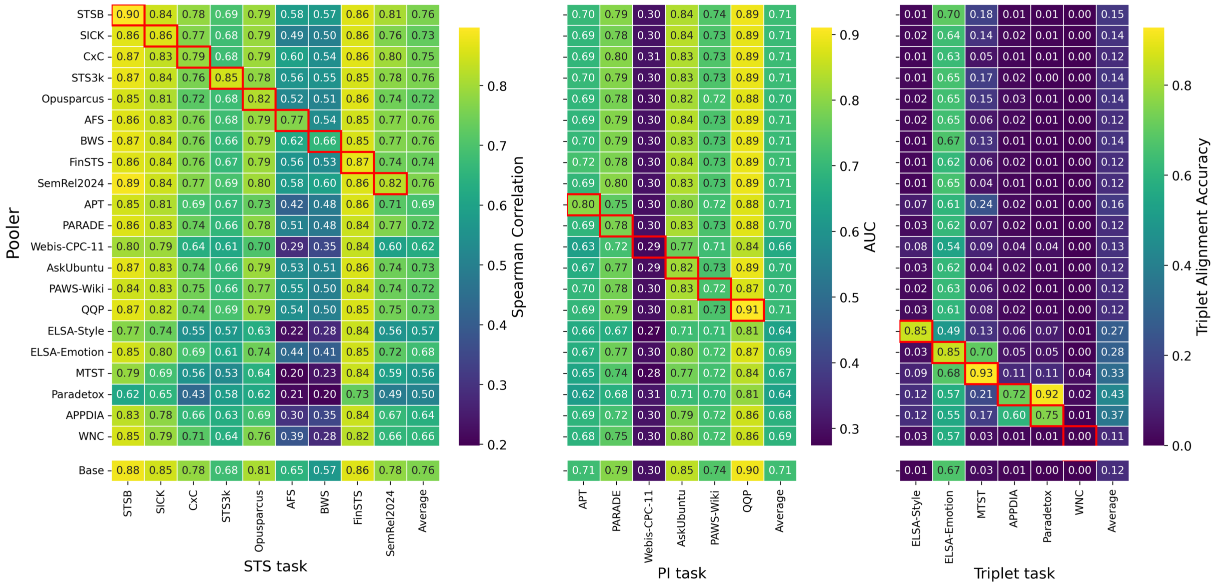
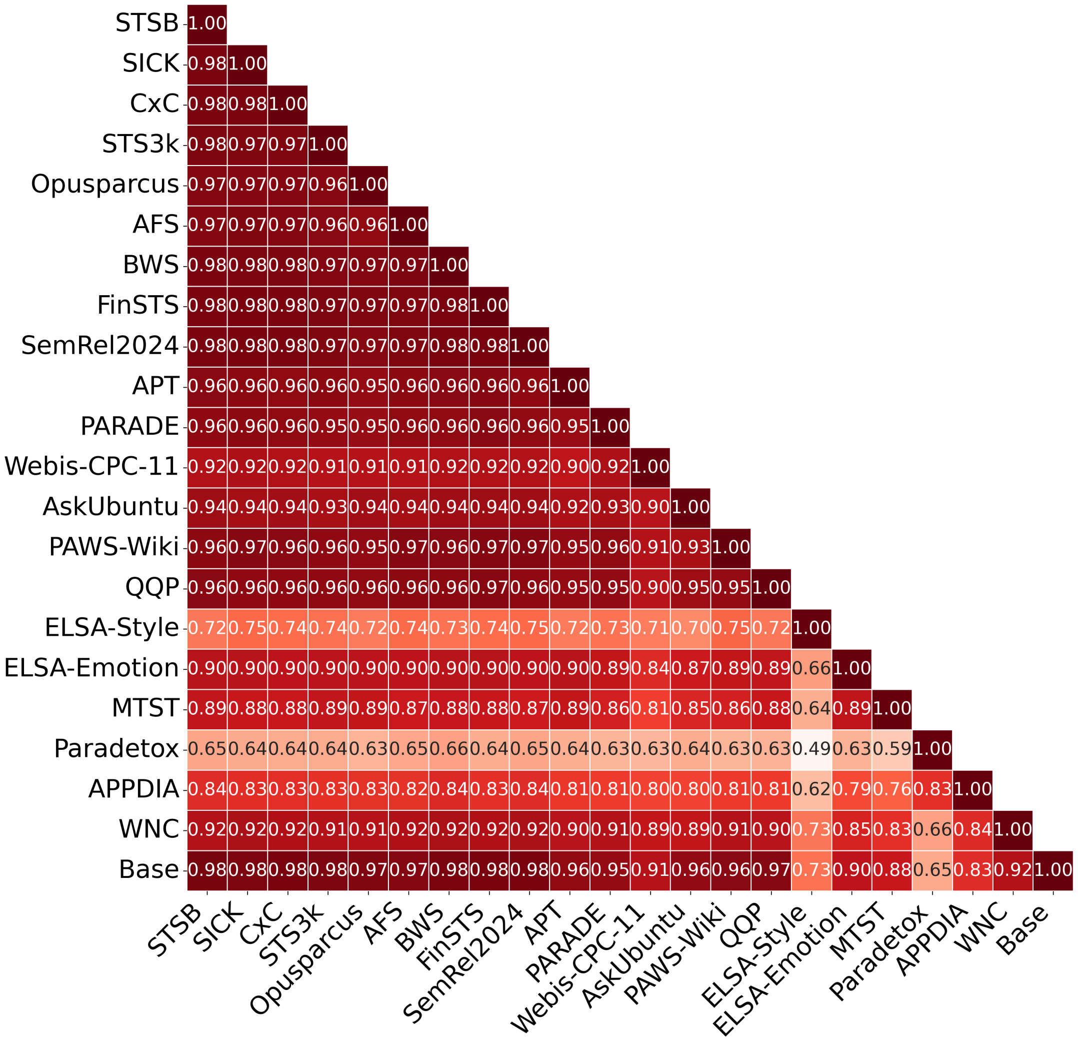


Figure 8: Performance of each Qwen3 pooler on all evaluation tasks.



	STSB	SICK	CxC	STS3k	Opusparcus	AFS	BWS	FinSTS	SemRel2024	APT	PARADE	Webis-CPC-11	AskUbuntu	PAWS-Wiki	QQP	ELSA-Style	ELSA-Emotion	MTST	Paradetox	APPDIA	WNC	Base
STSB	1.00																					
SICK	0.98	1.00																				
CxC	0.98	0.98	1.00																			
STS3k	0.98	0.97	0.97	1.00																		
Opusparcus	0.97	0.97	0.97	0.96	1.00																	
AFS	0.97	0.97	0.97	0.96	0.96	1.00																
BWS	0.98	0.98	0.98	0.97	0.97	0.97	1.00															
FinSTS	0.98	0.98	0.98	0.97	0.97	0.97	0.98	1.00														
SemRel2024	0.98	0.98	0.98	0.97	0.97	0.97	0.98	0.98	1.00													
APT	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.96	0.96	1.00												
PARADE	0.96	0.96	0.96	0.95	0.95	0.96	0.96	0.96	0.96	0.95	1.00											
Webis-CPC-11	0.92	0.92	0.92	0.91	0.91	0.91	0.92	0.92	0.92	0.90	0.92	1.00										
AskUbuntu	0.94	0.94	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.92	0.93	0.90	1.00									
PAWS-Wiki	0.96	0.97	0.96	0.96	0.95	0.97	0.96	0.97	0.97	0.95	0.96	0.91	0.93	1.00								
QQP	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.96	0.95	0.95	0.90	0.95	0.95	1.00							
ELSA-Style	0.72	0.75	0.74	0.74	0.72	0.74	0.73	0.74	0.75	0.72	0.73	0.71	0.70	0.75	0.72	1.00						
ELSA-Emotion	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.84	0.87	0.89	0.89	0.66	1.00					
MTST	0.89	0.88	0.88	0.89	0.89	0.87	0.88	0.88	0.87	0.89	0.86	0.81	0.85	0.86	0.88	0.64	0.89	1.00				
Paradetox	0.65	0.64	0.64	0.64	0.63	0.65	0.66	0.64	0.65	0.64	0.63	0.63	0.64	0.63	0.63	0.49	0.63	0.59	1.00			
APPDIA	0.84	0.83	0.83	0.83	0.83	0.82	0.84	0.83	0.84	0.81	0.81	0.80	0.80	0.81	0.81	0.62	0.79	0.76	0.83	1.00		
WNC	0.92	0.92	0.92	0.91	0.91	0.92	0.92	0.92	0.92	0.90	0.91	0.89	0.89	0.91	0.90	0.73	0.85	0.83	0.66	0.84	1.00	
Base	0.98	0.98	0.98	0.98	0.97	0.97	0.98	0.98	0.98	0.96	0.95	0.91	0.96	0.96	0.97	0.73	0.90	0.88	0.65	0.83	0.92	1.00

Figure 9: The correlation coefficients of predicted similarity scores between different BERT poolers.

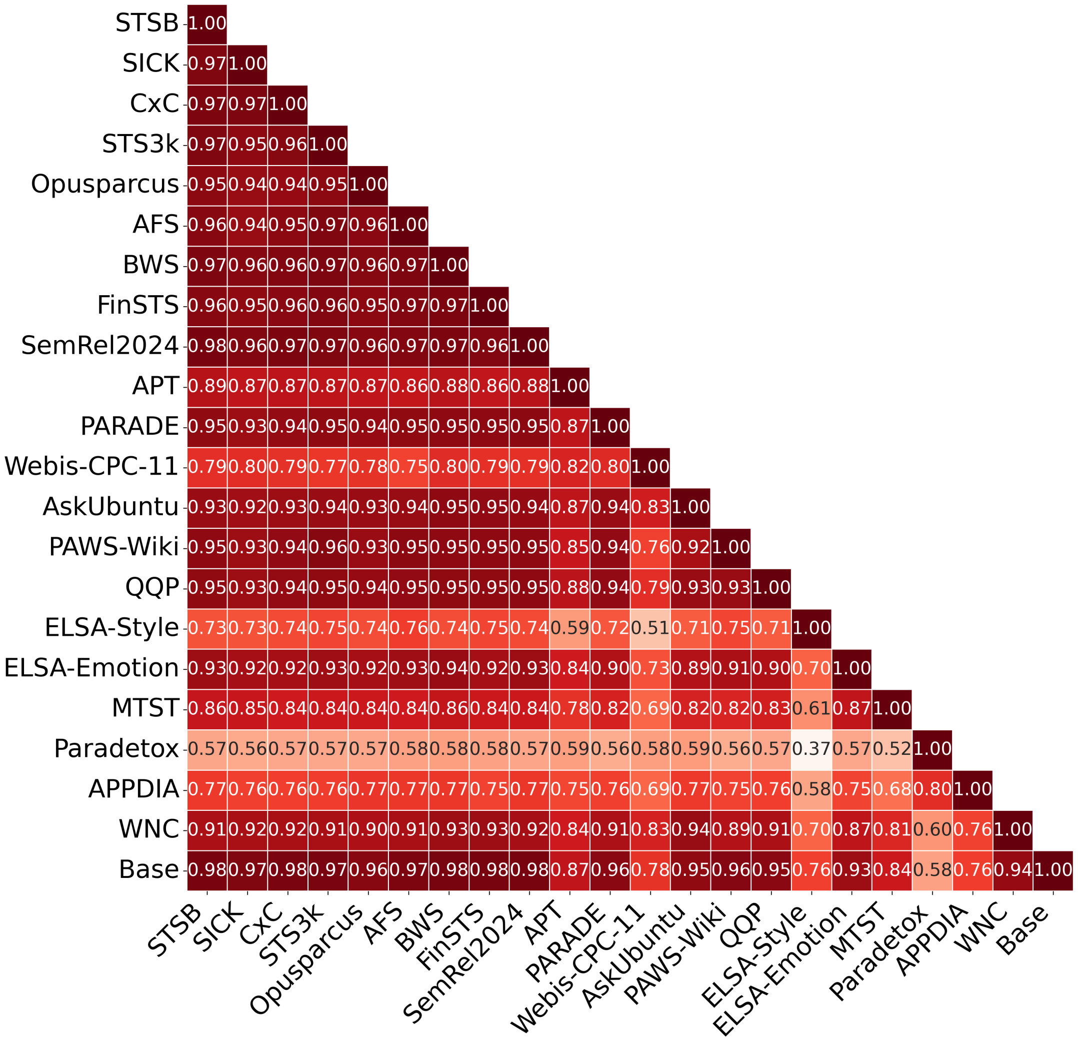


Figure 10: The correlation coefficients of predicted similarity scores between different Qwen3 poolers.

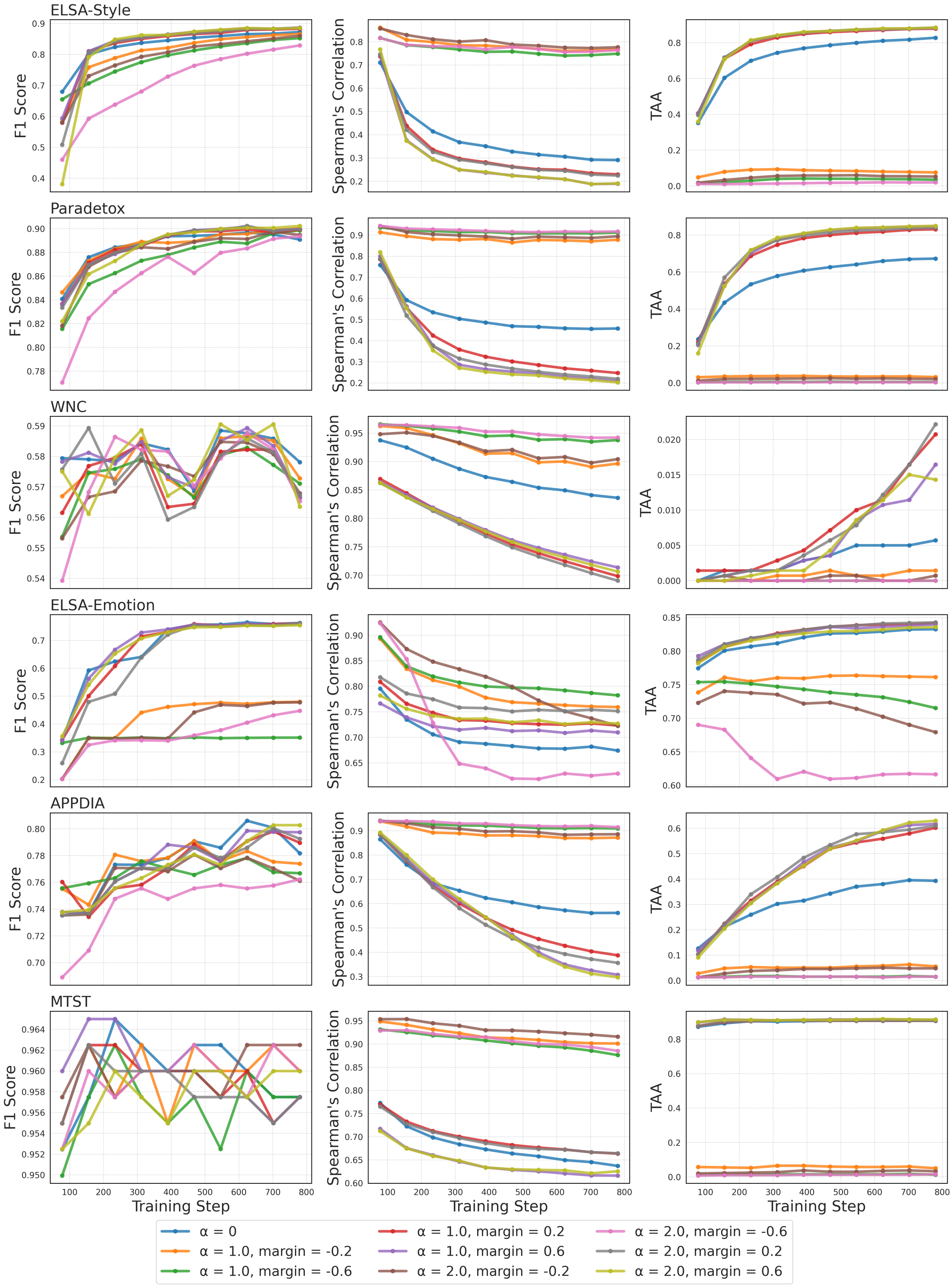


Figure 11: Training dynamics of BERT poolers with different margins and α in Triplet learning. Shown are the F1 score of anchor sentence attribute classification (left), the Spearman correlation with similarity predicted by the base model (center), and the Triplet Alignment Accuracy (right).

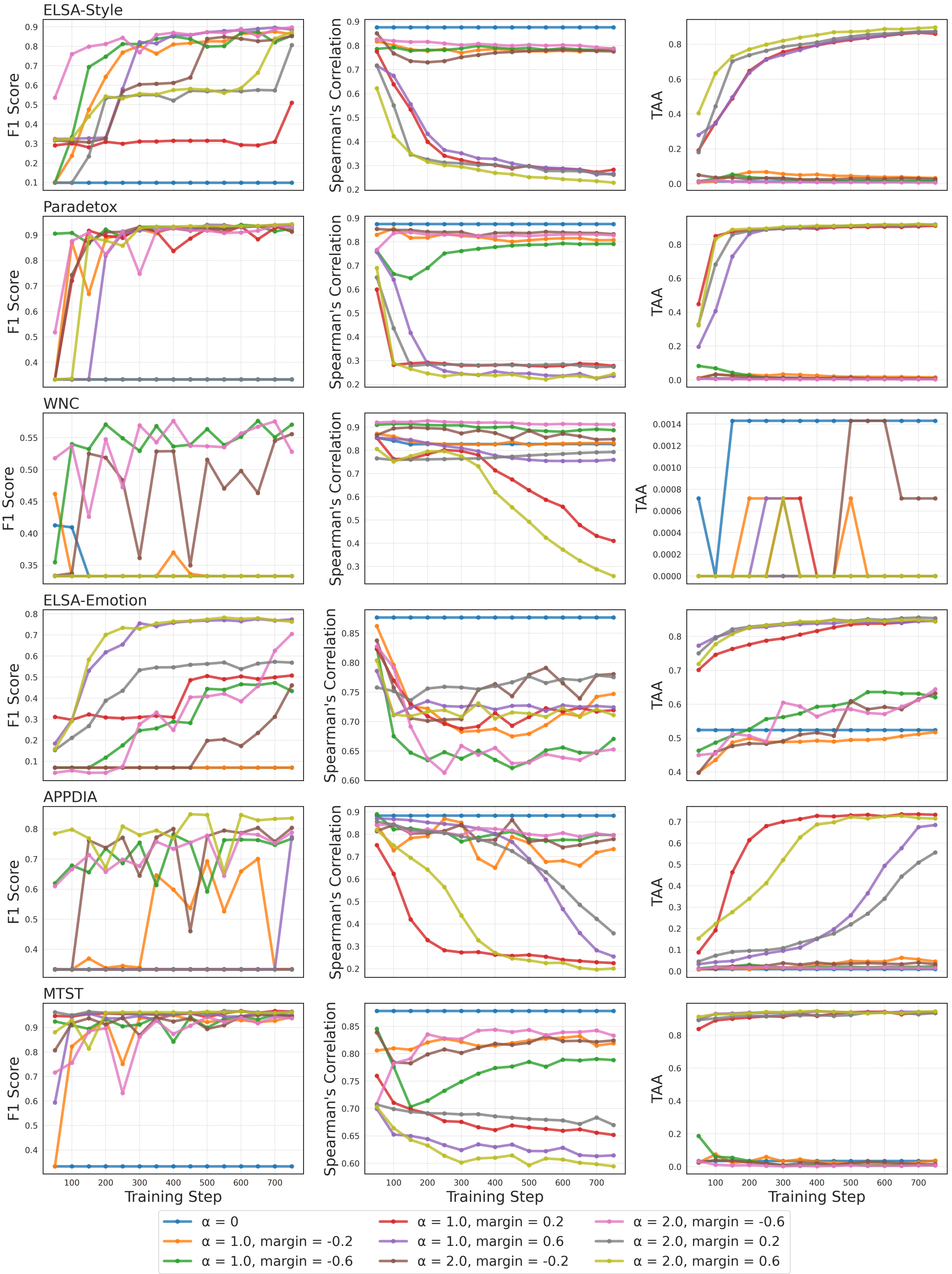


Figure 12: Training dynamics of Qwen3 poolers with different margins and α in Triplet learning. Shown are the F1 score of anchor sentence attribute classification (left), the Spearman correlation with similarity predicted by the base model (center), and the Triplet Alignment Accuracy (right).

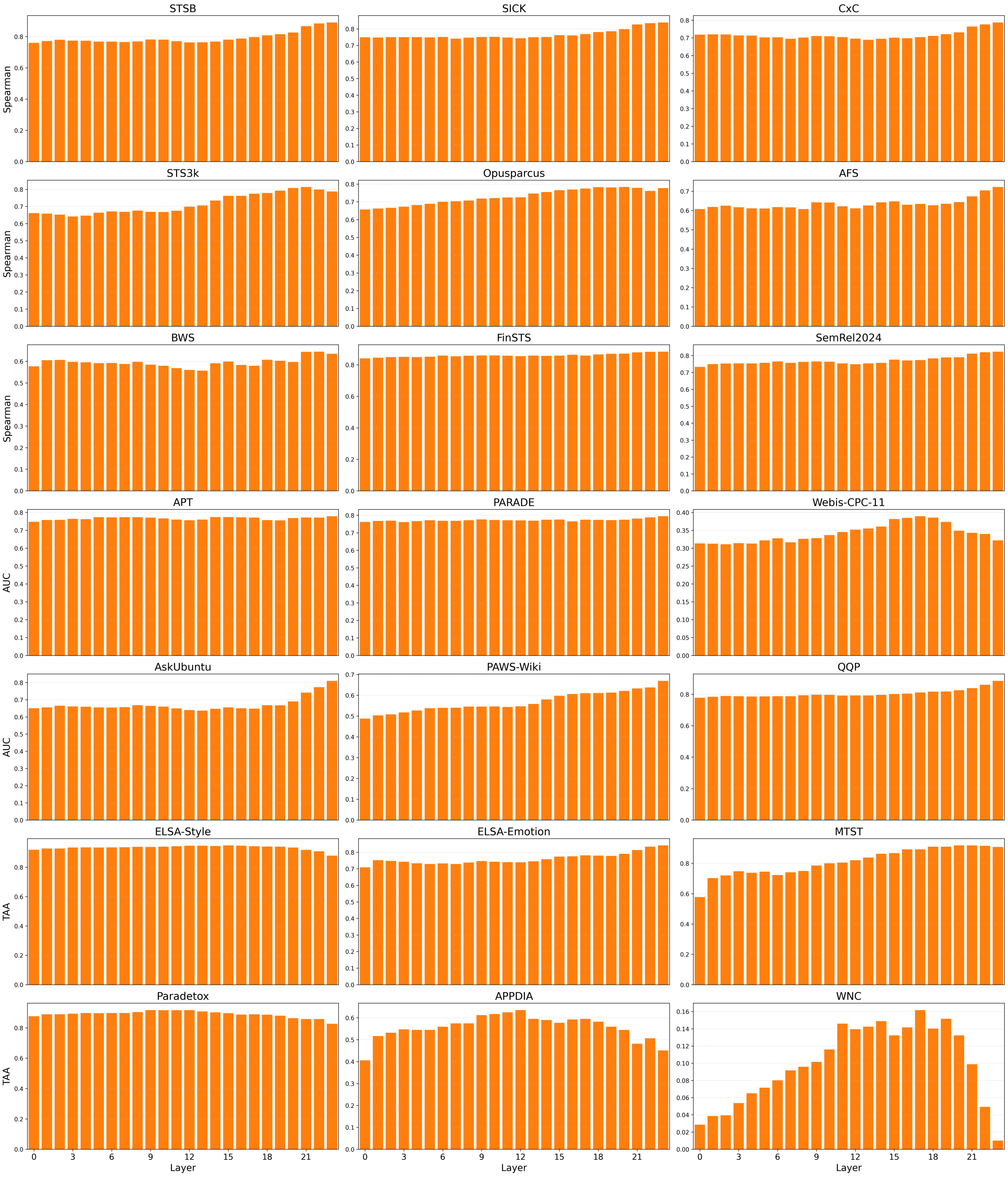


Figure 13: Results of the layer sweep experiment for the BERT pooler.

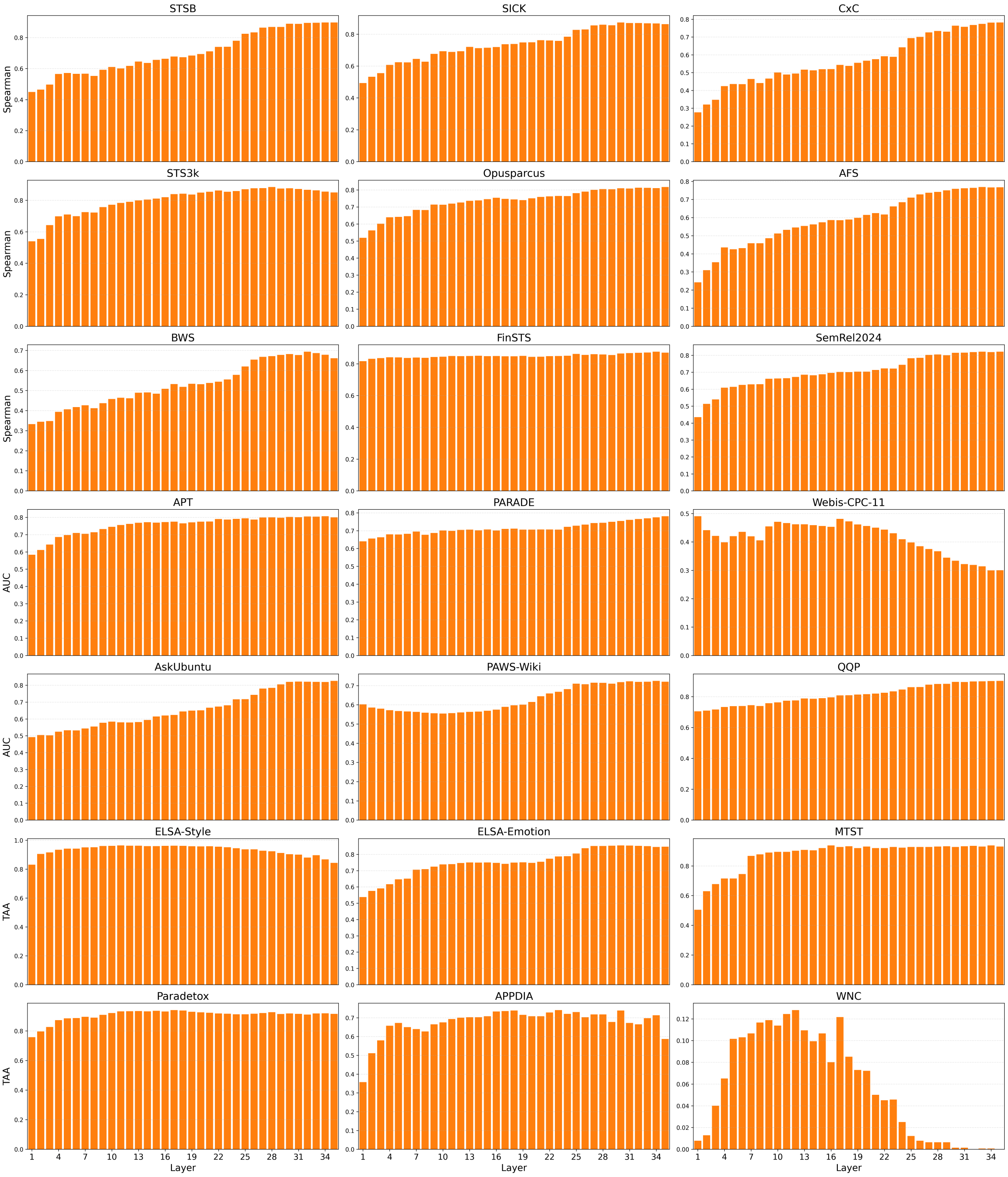


Figure 14: Results of the layer sweep experiment for the Qwen3 pooler.

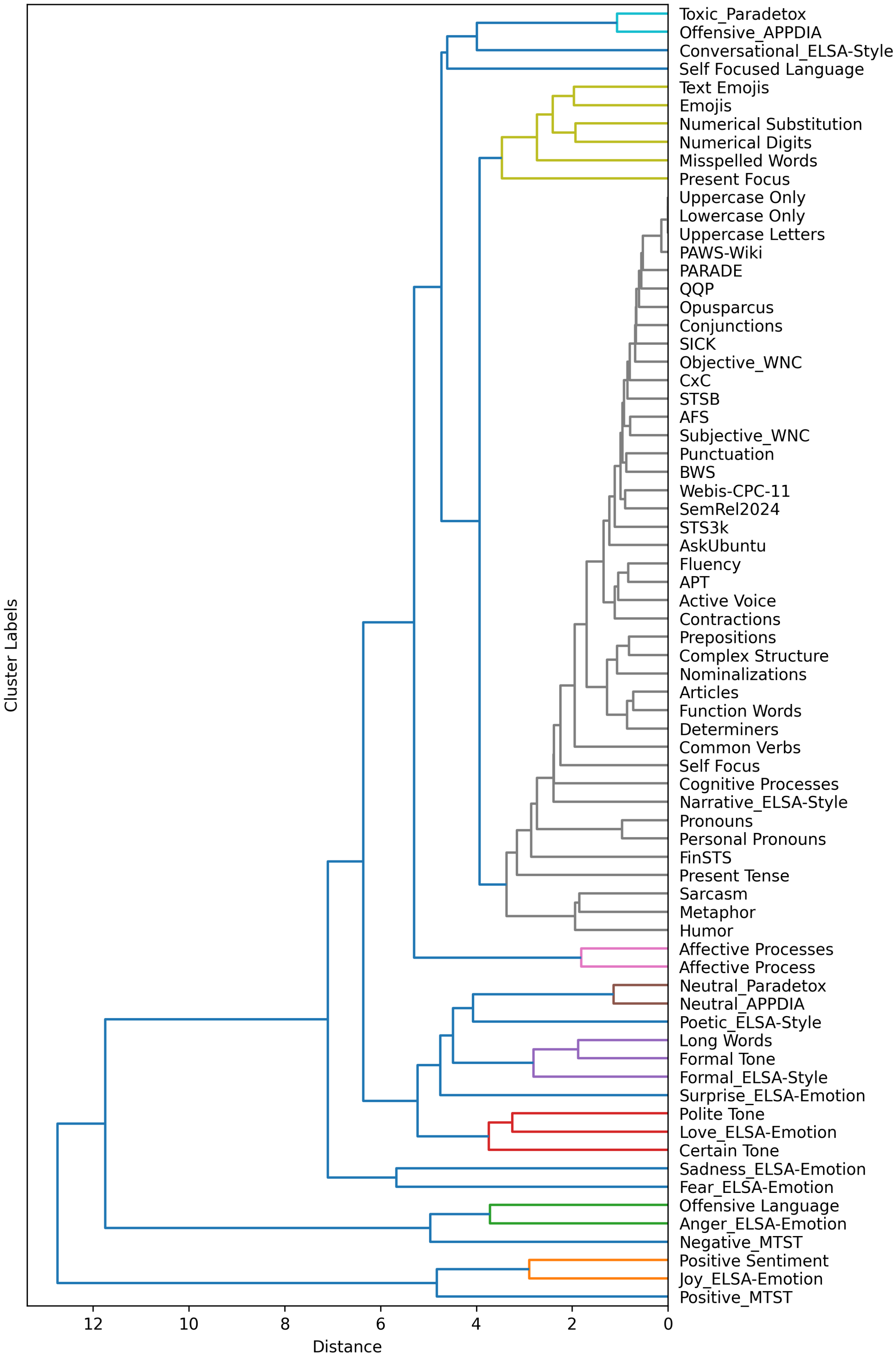


Figure 15: Aspect dendrogram based on the embedding representations of the BERT encoder. This figure is identical to Figure5, but displayed in a vertical orientation.

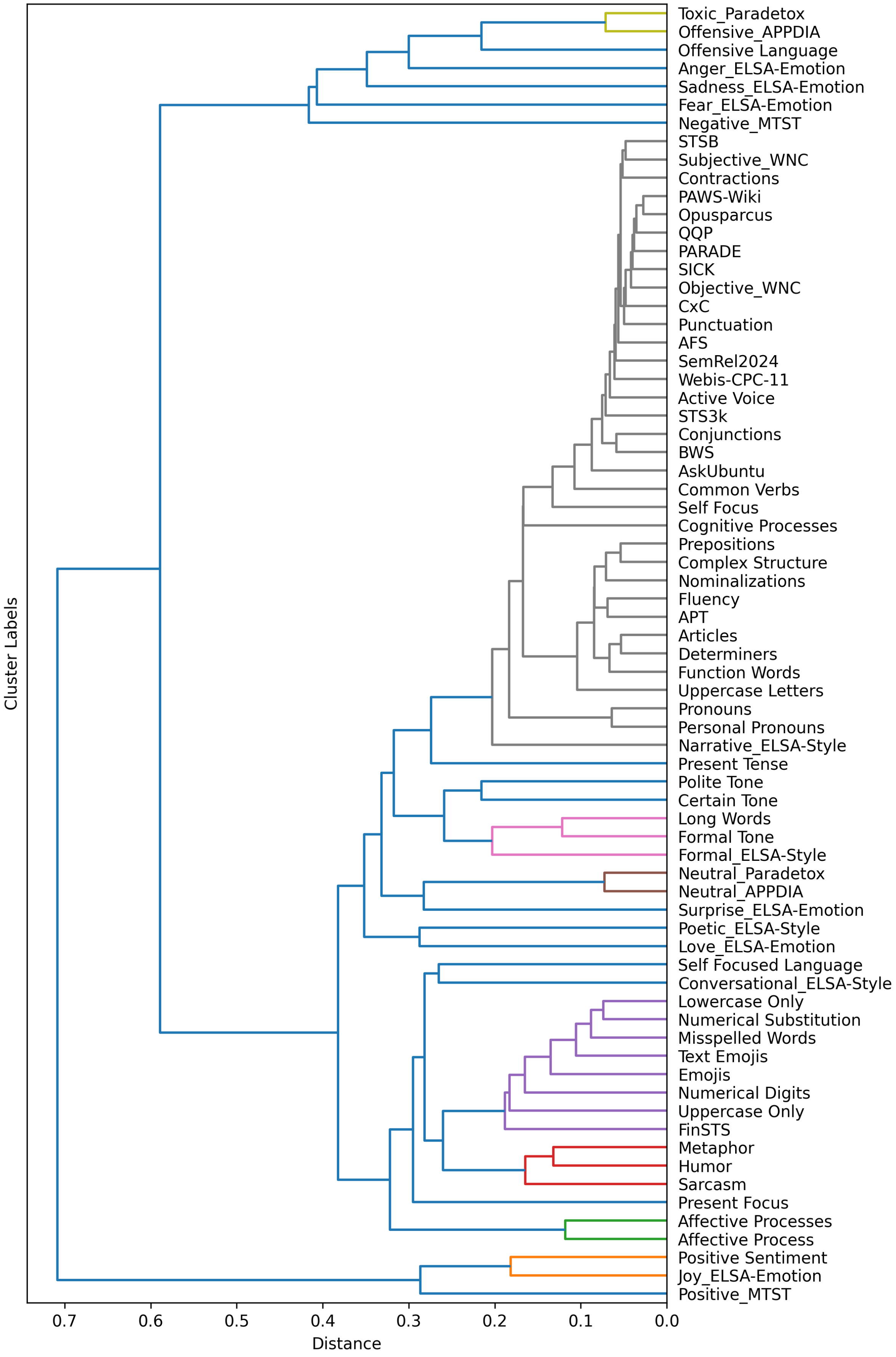


Figure 16: Aspect dendrogram based on the embedding representations of the Qwen3 encoder.

Dataset	**Category**	**Source domain**	**Samples**	**Lengths**
STSB (Cer et al., 2017)	*STS*	news headlines, video/image captions, NLI	8,577	10.2
SICK (Marelli et al., 2014)		video/image captions	9,840	9.6
CxC (Parekh et al., 2021)		video/image captions	88,052	10.4
STS3k (Fodor et al., 2025)		hand crafted by author	2,800	8.1
Opusparcus (Creutz, 2018)		OpenSubtitles database	2,900	4.7
AFS (Misra et al., 2016)		Internet Argument Corpus	6,000	22.2
BWS (Thakur et al., 2021)		UKP's Argument Facets, IBM Debater data	3,400	26.7
FinSTS (Yang et al., 2024)		Corporate Annual Report	3,988	26.9
SemRel2024 (Ousidhoum et al., 2024)		news, Wikipedia, SNS	8,350	12.2
APT (Nighojkar and Licato, 2021)	*PI*	MSRP, PPNMT	4,449	11.6
PARADE (He et al., 2020)		Computer Science web platforms	10,182	17.0
Webis-CPC-11 (Burrows et al., 2013)		Project Gutenberg	4,243	119.8
AskUbuntu (Lei et al., 2016)		Stack Exchange technical forum	6,557	92.0
PAWS-Wiki (Zhang et al., 2019)		Wikipedia	65,345	21.4
QQP (Quora, 2017)		Quora	404,308	11.1
ELSA (Gandhi and Gandhi, 2025)	*Triplet*	GoEmotions	10,434	20.5
MTST (Mukherjee et al., 2023)		Yelp	2,000	8.9
Paradetox (Logacheva et al., 2022)		Reddit, other online forums	11,927	10.3
APPDIA (Atwell et al., 2022)		Reddit	1,981	11.9
WNC (Pryzant et al., 2020)		Wikipedia	6,990	21.3
StyleDistance (Patel et al., 2025)		generated by GPT-4	4,000	14.7

Table 6: Meta information of the datasets. “Length” indicates the average number of words after preprocessing.

Dataset	**Sentence 1**	**Sentence 2**	**Score**
STSB	He will replace ron dittemore, who announced his resignation April 23.	Dittemore announced his plans to resign on April 23.	0.52
SICK	A panda that is cute is lying down	The panda bear is lying on the logs	0.83
CxC	A gray stripped cat standing on a blue rug.	A cat with a look of annoyance standing on a toilet lid.	0.14
STS3k	The organizers changed the start time of the conference.	The conference started a major change.	0.18
Opusparcus	Come on guys, have a seat.	Please sit down.	0.83
AFS	People have been found out to be innocent after the death penalty ha already been carried out....	There has been so much cases of people who have been executed, only to be found innocent after their death.	1.00
BWS	And dangerous felons are trying to get guns ; 39 % of state firearms applications are denied because of a felony conviction.	We had an armed intruder who was a career felon , mentally ill , 6 ' 4 " tall , 247 pound , and armed.	0.53
FinSTS	Generation considers capacity factor useful measure to analyze the nuclear fleet performance between periods.	Animal feed products occasionally contain contaminants due to inherent defects in those products or improper storage or handling.	0.20
SemRel2024	In 2007 , mit formed the kerberos consortium to foster continued development.	Mit developed kerberos to protect network services provided by project athena.	0.62

Table 7: STS examples.

Dataset	**Sentence 1**	**Sentence 2**	**Label**
APT	Typhoon maemi later moved out over the sea of japan, where it weakened considerably, the meteorology department said.	Maemi rapidly weakened to tropical storm status while moving over land, and was undergoing extratropical transition by the time it entered the sea of japan.	0
PARADE	- variables declared inside a function - local scope - they can only be used inside the function where they were declared	Memory for the formal parameters and variables declared in the body of a function	1
Webis-CPC-11	Peace is much desirable than war. the pains of a camp and war-zone may seem great, but the sufferings of a constitution that is being violated, and separatist rampant on the country, will be more devastating in nature.	Dear as war may be, a dishonorable peace will prove much dearer. great as may be the sufferings of the camp and the battle-field, yet the prolonged tortures of a murdered union, a violated constitution, and secession rampant over the country, will be found to be greater.	1
AskUbuntu	Connect my galaxy s duos phone as mass storage or drivers I want the drivers of my phone to be installed in ubuntu 12.04 lts please help me on this	I am unable to connect my samsung gt-s7562 with data cable I have installed ubuntu 12.04 and samsung galaxy s duos. I am unable to connect it with my system via data cable. please tell me steps so that I can transfer my files to my system.	1
PAWS-Wiki	It was chosen as the 19th best movie at the 7th Yokohama Film Festival .	It was chosen as the 7th best film at the 19th Yokohama Film Festival .	0
QQP	How do I fill in Address Line 1 and Address Line 2?	How do I register desired web address?	0

Table 8: Paraphrase identification examples.

Dataset	**Anchor**	**Positive**	**Negative**	Label
ELSA - Style	A heart heavy with sorrow, for your thoughts are tangled in the web of distorted tales.	In his presence, laughter dances like sunlight on a playful breeze.	I can't help but feel a deep sadness for you, knowing that your beliefs are molded by such one-sided sources.	{conversational, formal, **poetic**, narrative}
ELSA - Emotion	It's like I'm staring at an empty space without that photo.	I have adjusted my perspective on food moderation, yet there remains a lingering sense of loss for the indulgences I once enjoyed.	As I reflected on their incredible work, I felt a wave of gratitude wash over me, encouraging me to embrace my own blogging path, no matter how small.	{joy, anger, **sadness**, love, fear, surprise}
MTST	The beer was bad and warm!	Its not cheap and its not worth it.	The beer was nice and cold !	{positive, **negative**}
Paradetox	So sad nobody boycotts this shit.	Taking it easy fuck that ! ! winning	So sad no body boycott this	{**toxic**, neutral}
APPDIA	Okay, you really are not smart	That he did drugs	Okay, you are that stupid.	{offensive, **neutral**}
WNC	The next morning, she claimed it was an "honest mistake."	The front national is a political party in france , viewed as extreme right-wing by some.	The next morning, she admitted her statement was false.	{subjective, **objective**}

Table 9: Triplet Examples. The underlined and bold label represents the stylistic attribute of the anchor sentence.