Rank-Awareness and Angular Constraints: A New Perspective on Learning Sentence Embeddings from NLI Data

Zicheng Zhou, Min Huang, Qinghai Miao*
School of Artificial Intelligence
University of Chinese Academy of Sciences
zhoucheng23@mails.ucas.ac.cn
{huangm, miaoqh}@ucas.ac.cn

Abstract

Learning high-quality sentence embeddings from Natural Language Inference (NLI) data is often challenged by a critical signal conflict between discrete labels and the continuous spectrum of semantic similarity, as well as information loss from discarded neutral sentence pairs during training. To address this, we introduce Rank-Awareness and Angular Optimization Embeddings (RAOE), a framework that leverages the full NLI dataset (Entailment, Neutral, Contradiction) augmented with precomputed continuous similarity scores (S). RAOE employs a novel composite objective which features: (1) a Rank Margin objective that enforces rank consistency against S using an explicit margin, and (2) a Gated Angular objective that conditionally refines embedding geometry based on NLI label (L) and S score agreement. Extensive evaluations on STS tasks and the MTEB benchmark demonstrate RAOE's effectiveness. Our general-purpose RAOE-S1 model (BERT-base) significantly outperforms strong baselines, achieving an average Spearman's correlation of 85.11 (vs. SimCSE's 81.57 and AnglE's 82.43), and shows consistent improvements on MTEB. Further STS-specialized fine-tuning (RAOE-S2) establishes new state-of-the-art performance on STS (88.17 with BERT-base). These results confirm RAOE's ability to efficiently learn robust and nuanced sentence representations through the synergy of rank-awareness and conditional angular constraints. Code is available at https://github.com/Shengjingwa/RAOE.

1 Introduction

High-quality sentence embeddings are critical for advancing a wide range of Natural Language Processing (NLP) tasks (Ramesh Kashyap et al., 2024). They are fundamental for achieving strong performance in areas such as semantic textual similarity (STS) (Reimers and Gurevych, 2019; Gao et al., 2021), information retrieval (Palangi et al., 2016; Asai et al., 2023), and text clustering (Xu et al., 2023; Petukhova et al., 2025), and underpinning modern applications like Retrieval-Augmented Generation (RAG) systems for Large Language Models (LLMs) (OpenAI, 2022; Yang et al., 2024; Fan et al., 2024; Han et al., 2025).

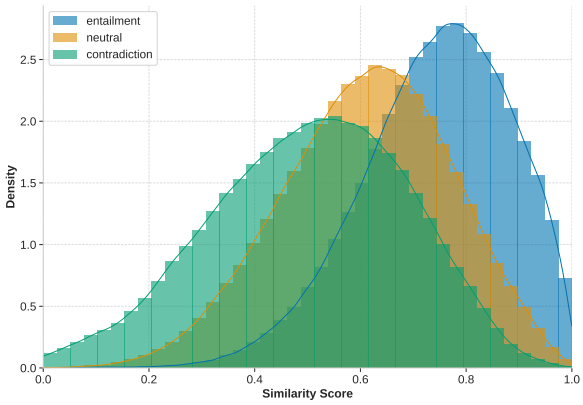

Figure 1: Distribution of Similarity Scores for Sentence Pairs by NLI Label. Note the significant overlap between distributions, indicating that discrete NLI labels do not perfectly capture the continuous spectrum of semantic similarity.

Supervised learning using Natural Language Inference (NLI) datasets, such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), has been a dominant paradigm for training effective sentence embeddings (Conneau et al., 2017; Reimers and Gurevych, 2019). Approaches like SBERT (Reimers and Gurevych, 2019) and contrastive methods like SimCSE (Gao et al., 2021) have demonstrated the power of NLI data. However, a key challenge persists: these methods primarily rely on the discrete NLI labels (Entailment(E), Neutral(N), Contradiction(C)). This reliance often struggles to capture the underlying continuous spectrum of semantic similarity, leading to suboptimal performance, particularly on

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22207–22221
November 4-9, 2025

fine-grained STS tasks. Furthermore, the common practice of discarding Neutral pairs potentially loses valuable semantic information.

The limitations of relying solely on discrete NLI labels are vividly illustrated in Figure 1, which shows the distribution of continuous similarity scores (derived from strong pre-trained embedding models, see Appendix A) for sentence pairs categorized by their NLI labels. Crucially, there is significant overlap between the distributions. For instance, a non-trivial number of Contradiction pairs exhibit higher similarity scores than some Entailment pairs. This overlap demonstrates that the discrete labels provide only a coarse and sometimes conflicting signal regarding the true semantic relatedness. Methods that treat these labels as absolute ground truth (e.g., simple classification) or make strong assumptions about their inherent ordering (e.g., assuming all Entailment pairs are strictly more similar than all Contradiction pairs) are forced to reconcile these inconsistencies, which can hinder the learning of nuanced representations. The substantial presence and overlap of Neutral pairs further underscore the information loss incurred by methods that discard them. This fundamental signal conflict between discrete labels and continuous similarity motivates the need for frameworks that can effectively integrate richer supervisory signals.

To overcome these limitations, we introduce RAOE (Rank-Awareness and Angular Optimization Embeddings), a novel framework designed to learn more robust and nuanced sentence representations from Natural Language Inference (NLI) data. RAOE achieves this by strategically integrating continuous similarity information with discrete NLI labels through a carefully designed data strategy and a synergistic learning objective. This approach directly addresses the challenge of conflicting signals and potential information loss inherent in traditional NLI-based training.

Our main contributions are:

- We address the signal conflict in NLI-based training by augmenting all labels (E, N, C) with continuous similarity scores (S) to create a more coherent supervisory signal.
- We introduce RAOE, a novel framework with a composite objective that learns from reliable rankings via a Rank Margin loss while filtering inconsistent signals using a Gated Angular loss.
- We demonstrate that RAOE is both state-of-the-art and highly efficient, outperforming strong baselines across diverse benchmarks and being substantially more computationally efficient.

2 Related Work

Learning effective sentence embeddings is fundamental to modern NLP (Kashyap et al., 2024; Reimers and Gurevych, 2019; Gao et al., 2021). Supervised approaches using large-scale Natural Language Inference (NLI) datasets (SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)) have shown significant success. Early methods like InferSent (Conneau et al., 2017), USE (Cer et al., 2018), and SBERT (Reimers and Gurevych, 2019) often used classification objectives over discrete NLI labels. However, this reliance on categories struggles to capture continuous semantic nuances and doesn't directly optimize the ranking crucial for tasks like Semantic Textual Similarity (STS).

Contrastive learning, notably SimCSE (Gao et al., 2021) using InfoNCE loss (van den Oord et al., 2018), has further advanced the field. However, these methods (Yan et al., 2021; Gao et al., 2021; Zhang et al., 2022; Chuang et al., 2022; Li and Li, 2024) often face limitations. They may treat negative samples uniformly, overlooking finer-grained semantic differences. Supervised variants often assume entailment pairs are inherently more similar than contradiction pairs—an assumption challenged by data overlap (cf. Figure 1)—and commonly discard Neutral NLI pairs (Gao et al., 2021; Li and Li, 2024), potentially losing valuable information. These observations underscore the need for strategies that effectively utilize the full NLI dataset (E, N, C) and integrate richer supervisory signals, such as continuous similarity scores (S), for finer-grained learning.

One promising avenue to leverage such richer signals and achieve finer-grained learning is by directly optimizing the rank order of sentence pairs, which is vital for representation learning. This is informed by insights from Information Retrieval's Learning to Rank (LTR) methods (e.g., (Li et al., 2007; Burges et al., 2005; Cao et al., 2007)) and margin-based losses like contrastive (Hadsell et al., 2006) and triplet losses (Schroff et al., 2015) that impose structure on embedding spaces. Applied to sentence embeddings, CoSENT (Su, 2022) optimized relative pairwise cosine similar-

ity order, while RankCSE (Liu et al., 2023) incorporated ranking consistency within contrastive learning. Our Rank Margin Objective extends CoSENT by introducing an explicit rank margin (I) derived from rank differences of pre-computed continuous similarity scores (S). This focuses optimization on pairs with substantial external similarity differences, aiming for more robust and accurate rank ordering.

Complementary to optimizing rank order, refining the angular relationships in embedding spaces offers another powerful approach. This concept, shown to enhance feature discriminability in computer vision (Liu et al., 2017; Wang et al., 2018; Deng et al., 2019) through angular margin losses, has inspired similar strategies in NLP. For sentence embeddings, AnglE (Li and Li, 2024) optimized angular distances with NLI data, and SimACE (Jeong et al., 2024) adapted angular similarity for unsupervised tasks. RAOE builds on these concepts with its Gated Angular objective, introducing a conditional mechanism activated by agreement between NLI labels (L) and continuous similarity scores (S). This offers complementary geometric constraints in real vector space, potentially mitigating conflicting supervisory signals and further enhancing the learned representations.

3 RAOE Framework

The RAOE framework enhances sentence embeddings derived from pre-trained language models (PLMs) via a two-stage process. Stage 1 trains a general-purpose model (RAOE-S1) using an enhanced Natural Language Inference (NLI) data strategy and a novel composite objective. An optional Stage 2 fine-tunes RAOE-S1 specifically for Semantic Textual Similarity (STS) tasks, yielding a specialized model (RAOE-S2). Figure 2 depicts the Stage 1 training process.

3.1 Enhanced NLI Data Strategy

To address limitations of prior NLI-based training—such as relying solely on discrete labels or discarding Neutral pairs (Gao et al., 2021; Li and Li, 2024) —RAOE employs an enhanced data strategy. It utilizes the complete NLI dataset (SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)), encompassing Entailment (E), Neutral (N), and Contradiction (C) pairs, to capture diverse semantic relationships. Crucially, since discrete labels (L) inadequately represent the continuous spectrum of similarity (as highlighted in Figure 1), each pair is augmented with a pre-computed continuous Similarity score, $S \in [0, 1]$. Specifically, S is computed as the average cosine similarity from two high-performance embedding models: `bilingual-embedding-large` (Lajavaness, 2024) and `jina-embeddings-v3` (Sturua et al., 2024), selected for their strong performance and potential complementary strengths (see Appendix A for further details on data generation). This provides a fine-grained measure of semantic relatedness.

The distributional overlap shown in Figure 1 confirms the limitations of discrete labels and motivates RAOE's approach. RAOE's composite objective is specifically designed to address this complexity by integrating information from both the discrete label L and the richer continuous score S.

3.2 The RAOE Composite Objective

RAOE utilizes a novel composite objective, L_{RAOE}, formulated to synergistically integrate signals from both the discrete NLI label L and the continuous similarity score S:

$$L_{\text{RAOE}} = w_{\text{rank}} L_{\text{rank}} + w_{\text{angle}} L_{\text{angle}} \quad (1)$$

where w_{rank} and w_{angle} are balancing hyperparameters determined via grid search on a development set.

3.2.1 Rank Margin Objective

Building upon the rank-ordering principle of CoSENT (Su, 2022), the Rank Margin objective promotes the ranking consistency between the ranking derived from pre-computed scores S and the ranking based on learned cosine similarities. These learned similarities, denoted as c_i for pair i, are calculated as $c_i = \cos(\mathbf{e}_{i1}, \mathbf{e}_{i2})$ from the learned sentence embeddings $\mathbf{e}_{i1}$ and $\mathbf{e}_{i2}$. Let s_i denote the corresponding pre-computed similarity score for the same pair i, and let $r_i = \text{rank}(s_i)$ be the rank of s_i within a batch (lower ranking positions correspond to reduced cosine similarity scores). The objective L_{rank} is defined as:

$$L_{\text{rank}} = \log \left(1 + \sum_{r_j - r_i > I} e^{\tau(c_i - c_j)} \right) \quad (2)$$

where τ is a temperature hyperparameter. The summation includes only pairs (i, j) where the

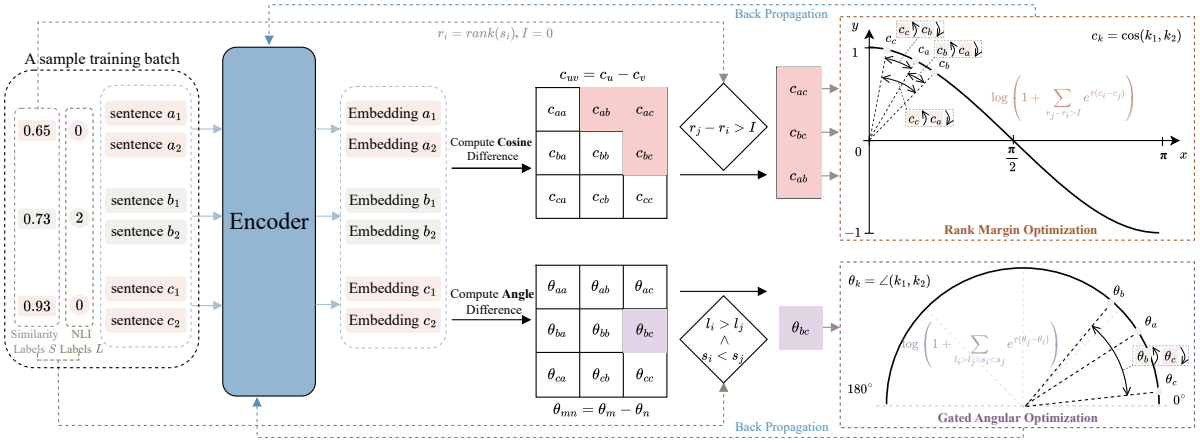

Figure 2: Overview of the RAOE Framework Training Process (Stage 1). An input batch containing sentence pairs with NLI Labels (L) and Similarity scores (S) is processed by an Encoder. Two objective components are calculated: (1) The Rank Margin objective operates on pairwise cosine similarities (c_k) and is activated when the rank margin condition $r_j - r_i > I$ holds, where $r_i = \text{rank}(s_i)$ is the rank based on similarity score S, and I is a predefined margin hyperparameter. (2) The Gated Angular objective operates on pairwise angles (θ_k) and is activated when the gating condition $l_i > l_j \wedge s_i < s_j$ holds. The combined objective is used for backpropagation to update the Encoder. The panels on the right conceptually visualize the optimization objectives within their respective cosine and angular spaces.

rank difference $r_j - r_i$ (based on S) exceeds a predefined Rank Margin I. This margin I, tuned via grid search (typically $I = 2$, see Appendix B), selects pairs where j is substantially more similar than i according to S. L_{rank} penalizes instances where this rank condition ($r_j - r_i > I$) is met, yet the learned cosine similarity order is inverted ($c_i > c_j$). Minimizing this objective thus aligns the learned similarity ranking (c) with the external score ranking (S), focusing on pairs separated by the margin I.

3.2.2 Gated Angular Objective

Inspired by the advantages of optimizing angular relationships (Li and Li, 2024), the Gated Angular objective focuses on the angle $\theta_k = \arccos(c_k)$ between embeddings. A key feature is its gating mechanism, which activates optimization only when the discrete NLI label L and the continuous score S provide consistent signals regarding the relative semantic relatedness of two pairs. We assign numerical values l_i to NLI labels such that Entailment < Neutral < Contradiction (e.g., E=0, N=1, C=2). The objective L_{angle} is then defined as:

$$L_{\text{angle}} = \log\left(1 + \sum_{l_i > l_j \wedge s_i < s_j} e^{\tau(\theta_j - \theta_i)}\right) \quad (3)$$

The gating condition, $l_i > l_j \wedge s_i < s_j$, selects pairs (i, j) for which both L and S indicate that pair i is semantically less related than pair j. If this condition is met, the objective penalizes instances where the angle for the less related pair i is erroneously smaller than the angle for the more related pair j ($\theta_i < \theta_j$). Consequently, minimizing L_{angle} promotes $\theta_i > \theta_j$ (a larger angle for the less related pair) under the gating condition. This conditional angular optimization is primarily motivated by the inconsistencies observed between discrete NLI labels and continuous similarity scores, as illustrated in Figure 1. Consequently, it offers geometric constraints complementary to L_{rank}, potentially refining the embedding space structure and mitigating cosine similarity saturation issues (Li and Li, 2024).

3.2.3 Objective Combination

The composite objective, L_{RAOE} (Eq. 1), synergistically integrates the Rank Margin objective (L_{rank}, Eq. 2) and the Gated Angular objective (L_{angle}, Eq. 3). L_{rank} enforces global ranking consistency guided by the continuous similarity scores S, while L_{angle} provides targeted, conditional angular refinement activated by the agreement between S and the discrete NLI labels L. This dual-objective strategy enables the model to learn more robust and nuanced semantic representations by concurrently leveraging similarity ranking and an-

gular geometric perspectives during Stage 1 training.

3.3 Optional STS Specialization Stage

After Stage 1 training produces the general-purpose RAOE-S1 model, an optional Stage 2 (RAOE-S2) specializes the embeddings for STS tasks. This stage involves fine-tuning the RAOE-S1 model on the STSBenchmark training dataset (Cer et al., 2017) using only the Rank Margin Objective L_{rank} (Eq. 2). L_{rank} is selected for this stage due to its direct alignment with the ranking objective inherent in STS evaluation and its independence from NLI labels, which are not present in the STSb dataset. The performance of the resulting STS-specialized model, RAOE-S2, is compared against RAOE-S1 in Section 4.

4 Experiments

In this section, we conduct a comprehensive evaluation of the RAOE framework. We begin by detailing the experimental setup in Section 4.1. Section 4.2 presents the main results, assessing RAOE's performance on STS benchmarks, its generalization on SentEval and MTEB, and its computational efficiency. Finally, Section 4.3 provides a series of ablation studies to analyze the contributions of RAOE's key components. Full implementation details are available in Appendix B.

4.1 Setup

Evaluation Benchmarks. We evaluate RAOE across several standard benchmarks. For Semantic Textual Similarity (STS), performance is measured using Spearman's rank correlation ($\rho \times 100$) on seven tasks: STS12–STS16 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (STSb) (Cer et al., 2017), and SICK-Relatedness (SICK-R) (Bentivogli et al., 2016). To assess generalization, we utilize the SentEval transfer learning suite (Conneau and Kiela, 2018) (reporting accuracy $\times 100$) and the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022). MTEB offers a comprehensive evaluation across 56 diverse tasks grouped into seven categories: Classification (12 datasets), Clustering (11), Pair Classification (3), Reranking (4), Retrieval (15), STS (10), and Summarization (1), with results typically averaged across all tasks.

Baselines. We compare RAOE against several strong baselines, including foundational supervised methods (InferSent (Conneau et al., 2017), USE (Cer et al., 2018)), recent high-performance embedding models (Jina Embeddings v3 (Sturua et al., 2024), Bilingual Embedding Large (Lajavaness, 2024)), and various NLI-based approaches. The latter category encompasses methods with distinct objectives: classification (SBERT (Reimers and Gurevych, 2019)), contrastive learning (SimCSE (Gao et al., 2021)), data augmentation (ConSERT (Yan et al., 2021)), pairwise ranking (CoSENT (Su, 2022), related to L_{rank}), angular optimization (AnglE (Li and Li, 2024), related to L_{angle}), and distillation (RankCSE (Liu et al., 2023), MSE). This selection provides a robust comparison against established, leading, and methodologically relevant models.

4.2 Main Results

STS Benchmark Performance. As shown in Table 1, the general-purpose RAOE-S1 model consistently surpasses strong baselines such as SimCSE (Gao et al., 2021) and AnglE (Li and Li, 2024) on the seven STS tasks across all evaluated backbone architectures (BERT-base/large, ModernBERT-base/large, Qwen2.5). Notably, using BERT-base, RAOE-S1 attains an average Spearman correlation of 85.11, a substantial improvement over SimCSE (81.57) and AnglE (82.43). These findings underscore the efficacy of RAOE's enhanced data strategy and novel composite objective.

The STS-specialized model, RAOE-S2, derived by fine-tuning RAOE-S1 on STSb using only L_{rank}, yields substantial further improvements, particularly on the in-domain STSb and related SICK-R tasks. RAOE-S2 establishes new state-of-the-art or highly competitive results across all STS benchmarks. For example, RAOE-S2 with BERT-large attains an average correlation of 88.90, and the Qwen2.5-7B variant reaches an impressive 89.37 average, highlighting the framework's potential for task-specific specialization.

Transfer Task Performance (SentEval). Table 2 focuses on the general-purpose RAOE-S1 model's generalization capabilities on SentEval transfer tasks. RAOE-S1 consistently achieves higher average accuracy than strong baselines like SimCSE across various classification tasks and backbones. For instance, with the BERT-base

Method	Params	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
InferSent-GloVe †	-	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
USE †	-	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
jina-embeddings-v3 ‡	572M	82.43	89.50	84.94	89.31	86.85	90.34	**86.50**	87.12
bilingual-embedding-large ‡	559M	85.52	89.37	91.61	92.02	86.29	89.15	80.56	87.79
BERT-base									
ConSERT	110M	74.07	83.93	77.05	83.66	78.76	81.36	76.77	79.37
COSENT	110M	71.35	77.52	75.05	79.68	76.05	78.99	71.19	75.69
SBERT †	110M	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SimCSE	110M	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
AnglE †	110M	75.26	85.61	80.64	86.36	82.51	85.64	80.99	82.43
RAOE-S1 (ours)	110M	79.35	87.31	84.34	89.23	83.94	87.59	84.00	85.11
RAOE-S2 (ours)	110M	83.00	90.70	91.94	92.17	84.47	88.55	86.33	88.17
BERT-large									
SimCSE ‡	340M	75.78	86.33	80.44	86.06	80.86	84.87	81.14	82.21
RAOE-S1 (ours)	340M	81.09	89.17	85.88	89.60	84.37	87.99	84.29	86.06
RAOE-S2 (ours)	340M	84.92	**92.12**	**92.93**	92.35	85.11	88.90	85.95	88.90
ModernBERT-base									
SimCSE ⋆	149M	78.96	85.60	81.16	86.49	83.74	86.05	80.28	83.18
RAOE-S1 (ours)	149M	81.78	89.12	85.43	89.53	85.72	88.66	84.23	86.35
RAOE-S2 (ours)	149M	84.82	90.24	91.38	91.72	85.78	88.36	85.84	88.31
ModernBERT-large									
SimCSE ⋆	395M	79.41	87.29	82.69	87.59	85.02	86.93	80.43	84.19
RAOE-S1 (ours)	395M	82.97	90.28	87.04	90.59	87.06	89.30	84.92	87.45
RAOE-S2 (ours)	395M	85.40	91.56	92.65	**92.62**	87.00	89.27	86.47	89.28
Qwen2.5									
SimCSE ⋆	7B	80.19	90.10	85.37	89.10	86.84	87.57	81.53	85.81
RAOE-S1 (ours)	7B	83.86	91.32	88.10	91.12	87.82	90.21	84.81	88.18
RAOE-S2 (ours)	7B	**85.57**	91.87	91.72	92.25	**88.36**	**90.49**	85.33	**89.37**

Table 1: Results on standard STS tasks, reported as Spearman correlation ($\rho \times 100$). Within each backbone model group, the best and second-best results per dataset are highlighted in blue and gray, respectively; **bold** indicates the highest score per column across all models. Results marked with † are obtained from (Li and Li, 2024). Results marked with ‡ are from evaluating official models. Results marked with ⋆ denote our own reimplementation using official code. For the remaining baselines, we refer to the corresponding original papers to obtain their results.

backbone, RAOE-S1 achieves an average accuracy of 87.85 (compared to 85.81 for SimCSE), showing significant improvements on tasks like CR, MPQA, SST2, and MRPC. This indicates that RAOE-S1 learns robust and transferable features suitable for general NLP applications. Notably, the Qwen2.5-7B variant of RAOE-S1 reaches a high average accuracy of 90.60. While RAOE-S2 excels on STS tasks (Table 1), its specialization leads to slightly lower average performance on these general transfer tasks, illustrating the expected trade-off between specialization and broad applicability (see Appendix Table 10 for RAOE-S2 SentEval results).

MTEB Performance. To assess its broad generalization capabilities, we evaluated the general-purpose RAOE-S1 model on the comprehensive MTEB benchmark (Muennighoff et al., 2022). As shown in Table 3, RAOE-S1 demonstrates superior performance over the strong SimCSE baseline across all evaluated backbones. The trend is consistent, with RAOE-S1 achieving higher average scores and outperforming SimCSE in nearly every task category. The improvements are particularly notable on the ModernBERT-large backbone, where RAOE-S1 achieves a final average score of 56.50 versus SimCSE's 53.52. These results reinforce RAOE-S1's standing as a robust, general-purpose sentence embedding model. As detailed in Appendix Table 11, while the STS-specialized RAOE-S2 model expectedly improves upon RAOE-S1 within the MTEB's STS category,

Model	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	Avg.
				BERT-base				
SBERT †	80.10	86.25	94.61	88.78	84.90	89.00	73.25	85.27
AnglE †	83.00	89.38	94.72	89.87	87.20	89.00	75.54	86.96
SimCSE	81.18	86.46	94.45	88.88	85.50	89.80	74.43	85.81
RAOE-S1 (ours)	83.78	89.96	94.62	90.52	89.02	89.80	77.28	87.85
				BERT-large				
SimCSE ‡	85.53	90.97	95.47	90.65	90.72	90.00	77.22	88.65
RAOE-S1 (ours)	85.05	91.21	95.05	90.76	90.66	92.20	**77.74**	88.95
				Qwen2.5-7B				
SimCSE ⋆	**87.79**	89.86	**96.55**	89.08	91.65	94.60	71.54	88.72
RAOE-S1 (ours)	87.50	**92.98**	95.56	**90.95**	**94.07**	**96.80**	76.35	**90.60**

Table 2: Evaluation results of the general-purpose RAOE-S1 model on SentEval transfer tasks (accuracy $\times 100$). RAOE-S1 is compared against baselines across different backbones. **Bold** indicates the best result per column across all models shown. Results marked with † are obtained from (Li and Li, 2024). Results marked with ‡ are from evaluating official models. Results marked with ⋆ denote our own reimplementation using official code. For the remaining baselines, we refer to the corresponding original papers to obtain their results.

Model	Classification	Clustering	PairClassification	Reranking	Retrieval	STS	Summarization	Avg.
				BERT-base				
SimCSE	67.32	33.43	73.68	47.54	21.82	79.12	**31.25**	48.72
RAOE-S1 (ours)	69.34	34.31	82.12	49.76	24.09	82.24	**31.25**	51.25
				BERT-large				
SimCSE	68.92	35.17	76.33	47.65	21.66	79.66	30.89	49.75
RAOE-S1 (ours)	69.73	34.89	82.25	50.07	23.95	83.49	29.63	51.63
				ModernBERT-base				
SimCSE	68.65	36.72	78.32	49.56	24.10	80.30	29.64	50.98
RAOE-S1 (ours)	70.43	36.83	83.40	52.30	30.67	83.05	29.64	54.10
				ModernBERT-large				
SimCSE	71.13	37.22	81.37	50.44	29.46	81.61	30.22	53.52
RAOE-S1 (ours)	**72.22**	**38.32**	**85.09**	**53.73**	**33.84**	**84.63**	29.02	**56.50**

Table 3: Average MTEB benchmark scores (total 56 datasets) comparing the general-purpose RAOE-S1 against SimCSE. **Bold** indicates the better result per row. Baseline results sourced from the official leaderboard or our evaluations.

its overall average score is lower, further illustrating the trade-off between broad applicability and task-specific fine-tuning.

Efficiency Comparison. RAOE delivers state-of-the-art accuracy while remaining computationally efficient. Table 4 yields two key observations. First, RAOE achieves the highest average STS score (85.11), surpassing InfoNCE (81.50), scores embedding distillation (MSE2, 81.06), rank distillation (RankCSE, 84.38), and direct embedding distillation (MSE, 84.76). This supports the premise that optimizing relative order and geometric structure, as RAOE does, is more effective than fitting absolute teacher scores. Second, RAOE matches the efficiency of InfoNCE—21.17 GB memory and 18.04 min per 1k steps—while RankCSE and MSE are substantially more resource-intensive (higher memory and over 4× slower). Hence, RAOE combines effectiveness with practicality, making it well suited for real-world deployment.

4.3 Ablation Studies

We conducted ablation studies to analyze the contributions of RAOE's key components: pooling strategy, objective functions (L_{rank}, L_{angle}), and NLI data utilization.

Method	Score	Memory	Time
InfoNCE	81.50	21.17 GB	17.93 min
MSE2	81.06	21.15 GB	18.40 min
RankCSE	84.38	26.71 GB	77.15 min
MSE	84.76	26.15 GB	72.28 min
RAOE	**85.11**	**21.17 GB**	**18.04 min**

Table 4: Performance–efficiency comparison of RAOE versus distillation and rank-based baselines on BERT-base. Score is the average Spearman correlation ($\rho \times 100$) across 7 STS benchmarks; Time is training time per 1,000 steps. `MSE` distills similarities from teacher-generated embeddings (on-the-fly), whereas `MSE2` distills the static, pre-computed similarity scores (S).

Model	Avg. STS Score
$\text{BERT}_{\text{base}}$ + CLS	84.84
$\text{BERT}_{\text{base}}$ + Mean	**85.11**
$\text{BERT}_{\text{base}}$ + Max	84.48
$\text{ModernBERT}_{\text{base}}$ + CLS	**86.35**
$\text{ModernBERT}_{\text{base}}$ + Mean	86.08
$\text{ModernBERT}_{\text{base}}$ + Max	85.05

Table 5: Ablation study on pooling methods using RAOE-S1. Average Spearman correlation ($\rho \times 100$) across 7 STS benchmarks is reported. Best result for each backbone model is in **bold**.

Impact of Pooling Strategy. Table 5 shows the results of evaluating Mean, CLS, and Max pooling for RAOE-S1. Mean Pooling yielded the best average STS score for BERT-base (85.11), while CLS Pooling was optimal for ModernBERT-base (86.35), consistent with prior findings for this architecture (Warner et al., 2024). These results validate our default pooling choices.

Effect of Objective Components. Table 6 compares the performance of the full RAOE objective against its individual components. The results indicate that both components contribute positively; using only L_{rank} or only L_{angle} reduces the average STS score by 0.80 and 1.09 points, respectively, compared to the full objective. The composite loss yields the best performance, confirming the synergistic effect of combining rank-ordering and angular constraints.

Effect of Rank Margin (I). The inclusion of a rank margin I in L_{rank} proved consistently beneficial, with its impact becoming more pronounced

Model	Avg. STS Score
RAOE	**85.11**
Only L_{rank}	84.31
Only L_{angle}	84.02

Table 6: Ablation study on RAOE objective components (L_{rank}, L_{angle}) using BERT-base backbone. Average Spearman correlation ($\rho \times 100$) across 7 STS benchmarks is reported. Best overall result in **bold**.

Model	With Margin	Without Margin	Improvement
BERT-base	**85.11**	85.06	+0.05
BERT-large	**86.06**	85.90	+0.16
Qwen2.5-7B	**88.18**	87.73	**+0.45**

Table 7: Ablation of the Rank Margin I in L_{rank}. Scores are average STS Spearman ($\rho \times 100$)

on larger models. As detailed in Table 7, the performance gain was modest on BERT-base (+0.05) but grew substantially with model capacity, reaching +0.45 on Qwen2.5-7B. This suggests that for stronger encoders, focusing optimization on pairs with significant external rank differences is an increasingly effective strategy.

Angular Optimization Strategy	Avg. STS Score
RAOE (Full Gating: L + S)	**85.11**
Gating on S only	84.30
Gating on L only	76.79

Table 8: Ablation of the gating mechanism in L_{angle} on BERT-base. Results are average Spearman correlation ($\rho \times 100$) over seven STS benchmarks. The full gating activates the angular loss only when the NLI label (L) and the continuous score (S) agree.

Gating Mechanism in L_{angle}. To prevent the model from learning from conflicting supervisory signals, we introduced a gating mechanism in L_{angle}. Its importance is underscored in Table 8. The full gating strategy, which activates the loss only upon agreement between the NLI label (L) and the similarity score (S), achieves the best performance (85.11). Relying solely on the label (L) for gating causes a sharp performance drop (–8.32), confirming that our conditional approach effectively filters out misleading signals inherent in the NLI data.

Utilizing Neutral Pairs. Table 9 examines the impact of including Neutral pairs from the NLI

Model	$E+N+C$	$E+C$	Gain from N Pairs
RAOE (ours)	**85.11**	84.64	**+0.47**
SimCSE	81.61	81.57	+0.04
AnglE	81.48	82.43	-0.95

Table 9: Ability to utilize Neutral NLI pairs. $E+N+C$ uses all NLI labels (Entailment, Neutral, Contradiction); $E+C$ uses only Entailment and Contradiction. Gain is $(E+N+C)-(E+C)$. Scores are average STS Spearman ($\rho \times 100$).

dataset during training. RAOE clearly benefits from explicitly modeling Neutral pairs, showing a +0.47 point gain. In contrast, SimCSE gains only marginally (+0.04), while AnglE's performance degrades (–0.95). This demonstrates that RAOE's composite objective, guided by the continuous score S, can successfully extract useful semantic signals from Neutral pairs rather than treating them as noise.

Collectively, these ablation studies validate the key design choices of RAOE. The results confirm the benefits of the synergistic combination of L_{rank} and L_{angle}, the effectiveness of the rank margin and gating mechanisms, and the framework's unique ability to leverage the full NLI dataset (including Neutral pairs) when augmented with continuous similarity scores.

5 Conclusion

We presented RAOE, a novel framework that addresses the critical signal conflict between discrete NLI labels and continuous similarity scores in sentence embedding learning. By leveraging the full NLI dataset augmented with continuous scores (S), RAOE implements a synergistic objective that combines a Rank Margin component to enforce consistency with reliable rankings and a Gated Angular component to filter conflicting signals and selectively refine embedding geometry. Our extensive evaluations demonstrate that RAOE not only achieves state-of-the-art performance on STS tasks and strong results across diverse benchmarks, but does so with remarkable computational efficiency. RAOE effectively integrates rank-awareness and conditional angular constraints, yielding robust and nuanced sentence representations.

Ethics Statement

Our work utilizes publicly available datasets (SNLI, MNLI, STS benchmarks). While standard for research, these datasets may contain inherent societal biases present in the source text, which could be reflected in the learned embeddings. Additionally, our method relies on similarity scores from existing large embedding models, potentially inheriting their limitations and biases. We release our code to promote transparency and reproducibility. Users should be mindful of potential biases when deploying models trained with our method in downstream applications.

Limitations

RAOE's performance relies on the quality of the pre-computed similarity scores (S), potentially inheriting limitations from the source models. Additionally, the precise geometric effects of the composite objective in high-dimensional space warrant further investigation. While the STS-specialized RAOE-S2 model excels on similarity tasks, it exhibits reduced generalization compared to the RAOE-S1 model, highlighting a trade-off for specific applications. Future work could explore dynamic similarity score generation, deeper geometric analysis, and extending RAOE to diverse languages or domains.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2025ZD0122005) and the National Natural Science Foundation of China (No. 62271485).

Here, we sincerely thank the reviewers and ACs for their valuable input, which has greatly improved our work.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei

Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 497–511, San Diego, California. 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.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics.

L. Bentivogli, R. Bernardi, M. Marelli, S. Menini, M. Baroni, and R. Zamparelli. 2016. Sick through the semeval glasses: lessons learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50:95–124.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank. In *Proceedings of the 24th international conference on Machine learning*.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen tau Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. *Preprint*, arXiv:2204.10298.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive learning of sentence

embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. 2025. Retrieval-augmented generation with graphs (graphrag). *Preprint*, arXiv:2501.00309.

Yoo Hyun Jeong, Myeongsoo Han, and Dong-Kyu Chae. 2024. A simple angle-based approach for contrastive learning of unsupervised sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5553–5572, Miami, Florida, USA. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2024. A comprehensive survey of sentence representations: From the BERT epoch to the CHATGPT era and beyond. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1738–1751, St. Julian's, Malta. Association for Computational Linguistics.

Lajavaness. 2024. bilingual-embedding-large. https://huggingface.co/Lajavaness/bilingual-embedding-large.

Ping Li, Qiang Wu, and Christopher Burges. 2007. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Xianming Li and Jing Li. 2024. Angle-optimized text embeddings. *Preprint*, arXiv:2309.12871.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. RankCSE: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13785–13802, Toronto, Canada. Association for Computational Linguistics.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 212–220.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

OpenAI. 2022. Introducing chatgpt. https://openai.com/index/chatgpt/.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.

Alina Petukhova, João P. Matos-Carvalho, and Nuno Fachada. 2025. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6:100–108.

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2024. A comprehensive survey of sentence representations: From the BERT epoch to the CHATGPT era and beyond. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1738–1751, St. Julian's, Malta. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Jianlin Su. 2022. Cosent (1): A more effective sentence vector scheme than sentence bert.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.

Lingling Xu, Haoran Xie, Zongxi Li, Fu Lee Wang, Weiming Wang, and Qing Li. 2023. Contrastive learning models for sentence representations. *ACM Transactions on Intelligent Systems and Technology*, page 1–34.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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 Generation of Enhanced NLI Data with Continuous Similarity Scores

To augment the standard Natural Language Inference (NLI) datasets with richer supervisory signals, we generated a continuous Similarity score (S) for each sentence pair in the combined training sets of SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). This process involved leveraging two distinct, high-performance sentence embedding models:

1. **Initial Encoding (S1):** All sentence pairs were first encoded using the `bilingual-embedding-large` model (Lajavaness, 2024). The cosine similarity between the resulting embeddings for each pair yielded an initial set of scores, denoted as S_1.

2. **Second Encoding (S2):** Independently, the same sentence pairs were encoded using the `jina-embeddings-v3` model (Sturua et al., 2024). Cosine similarity calculation on these embeddings produced a second set of scores, S_2.

3. **Final Score Computation (S):** The final continuous similarity score S used in our enhanced data strategy was computed as the average of the scores obtained from the two models:

$$S = \frac{S_1 + S_2}{2} \tag{4}$$

This averaging approach yields a robust estimate of semantic similarity by integrating signals from two distinct, high-performance embedding models, leveraging their complementary strengths and mitigating potential biases of any single model. This process resulted in a dataset of 941,581 entries, each comprising a sentence pair ('text1', 'text2'), its NLI label ('label'), and the computed similarity score ('similarity'). Figure 3 illustrates the resulting enhanced data format, which includes the original discrete NLI label (L) and the computed continuous similarity score (S) for each sentence pair. The complete Enhanced NLI Dataset will be released publicly upon acceptance.

• Traditional NLI Data:

```
{"text1": "sent A", "text2": "sent B", "label": 1} // Neutral
{"text1": "sent C", "text2": "sent D", "label": 2} // Contradiction
{"text1": "sent E", "text2": "sent F", "label": 0} // Entailment
```

• Enhanced NLI Data:

```
{"text1": "sent A", "text2": "sent B", "label": 1, "similarity": 0.44}
{"text1": "sent C", "text2": "sent D", "label": 2, "similarity": 0.10}
{"text1": "sent E", "text2": "sent F", "label": 0, "similarity": 0.65}
```

Figure 3: Illustration of Enhanced NLI Data Format with Labels and Similarity Scores.

B Implementation Details

Models and Training Setup. We employed several Pre-trained Language Models (PLMs): BERT (Devlin et al., 2018) with Mean Pooling, ModernBERT (Warner et al., 2024) with CLS Pooling, and Qwen2.5 (Yang et al., 2024). For the 7B-parameter Qwen2.5 model, we utilized QLoRA (Dettmers et al., 2023) for fine-tuning with an initial learning rate of 1×10^{-4}. Sentence embeddings for Qwen2.5 were obtained via the prompt "Summarize sentence {text} in one word:", using the last token's embedding, following AnglE (Li and Li, 2024).

Stage 1 training used the combined and enhanced SNLI and MNLI datasets for 5 epochs. The optional Stage 2 fine-tuning used the STS-Benchmark training set for 10 epochs. For BERT and ModernBERT models, we used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-5} and a batch size of 512, consistent with SimCSE (Gao et al., 2021). Experiments involving Qwen2.5 were conducted on 4 NVIDIA A100 GPUs, while all other experiments ran on a single NVIDIA 4090 GPU. For reference, Stage 1 training of RAOE with a BERT-base backbone takes approximately 95 minutes on one NVIDIA 4090 GPU.

Hyperparameters. Across all experiments, we used a fixed temperature $\tau = 20$ and set the objective balancing weights to $w_{\text{rank}} = 1$ and $w_{\text{angle}} = 1$. The Rank Margin I (Eq. 2) was determined via grid search on the development set, resulting in an optimal value of 16 for the Qwen2.5 backbone and 2 for all other models. Following SimCSE (Gao et al., 2021), we used a fixed random seed of 42 for all main experiments to ensure reproducibility.

Baseline Implementation for Ablations. For the efficiency comparison (Table 4), we implemented several baselines on an 80GB NVIDIA A100 GPU. These included the standard contrastive loss InfoNCE, two MSE variants (MSE and MSE2), and RankCSE. Note that RankCSE is an unsupervised method; for comparability, we trained it on 10^6 English Wikipedia sentences released with SimCSE, whereas the other baselines in this comparison were trained on our Enhanced NLI Data.

For the neutral pair utilization study (Table 9), we adapted SimCSE and AnglE to incorporate neutral pairs. For SimCSE, both `Neutral` and `Contradiction` pairs were treated as hard negatives. For AnglE, we similarly treated `Neutral` and `Contradiction` pairs as hard negatives in its contrastive loss component. For its angular loss component, we replaced the original binary assumption with a three-way ranking objective: `Angle(Entailment) < Angle(Neutral) < Angle(Contradiction)`.

Datasets and Evaluation. The models were fine-tuned using our Enhanced NLI Data as the training set. The STS Benchmark (STSb) development set served for hyperparameter tuning, and the test sets from the seven standard STS tasks were used for final evaluation, following the protocol of SimCSE (Gao et al., 2021).

C Full SentEval Results

Table 10 provides a detailed breakdown of the performance across all SentEval transfer learning tasks for the different models and backbones evaluated in this study.

D Full MTEB Results

Table 11 provides a detailed breakdown of performance across all MTEB task categories for the different models and backbones.

Model	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	Avg.
				BERT-base				
Avg. BERT †	78.66	86.25	94.37	88.66	84.40	92.80	69.54	84.94
BERT-CLS †	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS-BERT	81.09	87.18	94.96	88.75	85.96	88.64	74.24	85.83
DiffCSE	82.69	87.23	95.23	89.28	86.60	90.40	76.58	86.86
SBERT †	80.10	86.25	94.61	88.78	84.90	89.00	73.25	85.27
AnglE †	83.00	89.38	94.72	89.87	87.20	89.00	75.54	86.96
SimCSE	81.18	86.46	94.45	88.88	85.50	89.80	74.43	85.81
RAOE-S1 (ours)	83.78	89.96	94.62	90.52	89.02	89.80	77.28	87.85
RAOE-S2 (ours)	83.88	89.70	94.11	90.27	89.07	88.20	77.45	87.53
				BERT-large				
SimCSE ‡	85.53	90.97	95.47	90.65	90.72	90.00	77.22	88.65
RAOE-S1 (ours)	85.05	91.21	95.05	90.76	90.66	92.20	**77.74**	88.95
RAOE-S2 (ours)	85.42	90.86	94.75	90.50	89.95	91.60	76.41	88.50
				ModernBERT-base				
SimCSE ⋆	85.33	89.94	93.51	87.90	91.21	90.40	71.01	87.04
RAOE-S1 (ours)	85.65	91.52	94.01	89.28	92.20	92.80	70.49	87.99
RAOE-S2 (ours)	84.68	90.01	92.55	88.32	90.55	89.40	74.67	87.17
				ModernBERT-large				
SimCSE ⋆	87.11	92.18	94.45	88.93	92.26	92.40	72.00	88.48
RAOE-S1 (ours)	87.14	92.53	94.82	89.19	93.41	93.20	71.30	88.80
RAOE-S2 (ours)	86.99	90.30	93.63	88.68	91.98	88.20	72.46	87.46
				Qwen2.5-7B				
SimCSE ⋆	**87.79**	89.86	**96.55**	89.08	91.65	94.60	71.54	88.72
RAOE-S1 (ours)	87.50	92.98	95.56	**90.95**	**94.07**	**96.80**	76.35	**90.60**
RAOE-S2 (ours)	87.24	**93.09**	94.82	90.25	93.30	93.60	74.49	89.54

Table 10: Evaluation results on SentEval transfer tasks (accuracy $\times 100$). Best (blue) and second-best (gray) results per dataset are shown within each backbone group. **Bold** indicates the overall best score per column. Baseline sources: † (Li and Li, 2024); ‡ official models; ⋆ our reimplementation; unmarked from original papers.

Model	Classification	Clustering	PairClassification	Reranking	Retrieval	STS	Summarization	Avg.
			BERT-base					
SimCSE	67.32	33.43	73.68	47.54	21.82	79.12	31.25	48.72
RAOE-S1 (ours)	69.34	34.31	82.12	49.76	24.09	82.24	31.25	51.25
RAOE-S2 (ours)	67.37	27.73	80.84	47.92	16.49	83.91	**32.38**	47.62
			BERT-large					
SimCSE	68.92	35.17	76.33	47.65	21.66	79.66	30.89	49.75
RAOE-S1 (ours)	69.73	34.89	82.25	50.07	23.95	83.49	29.63	51.63
RAOE-S2 (ours)	68.58	31.32	82.06	49.72	20.65	85.61	30.17	50.16
			ModernBERT-base					
SimCSE	68.65	36.72	78.32	49.56	24.10	80.30	29.64	50.98
RAOE-S1 (ours)	70.43	36.83	83.40	52.30	30.67	83.05	29.64	54.10
RAOE-S2 (ours)	68.55	32.05	82.19	51.16	23.46	84.31	31.92	50.95
			ModernBERT-large					
SimCSE	71.13	37.22	81.37	50.44	29.46	81.61	30.22	53.52
RAOE-S1 (ours)	**72.22**	**38.32**	**85.09**	**53.73**	**33.84**	84.63	29.02	**56.50**
RAOE-S2 (ours)	70.81	33.72	84.29	52.73	28.65	**85.80**	30.19	53.61

Table 11: Performance on the Massive Text Embedding Benchmark (MTEB). Results are average scores per evaluation category. Within each backbone model group, the best and second-best results per task category are highlighted in blue and gray, respectively. **Bold** indicates the highest score per column across all models. Baseline results are from the official MTEB leaderboard or our evaluations.