PEU Lab at SemEval-2026 Task 4: Pairwise Text Comparison using RoBERTa and Ranking Loss

Hangchao Ma, Jiaxu Dao, Jinli Tong, Zhuoying Li, Qingsong Zhou, Xiuzhong Tang
School of Technology
Pu'er University
Contact:{mahangchao,daojiaxu}@peu.edu.cn

Abstract

This paper describes the system developed by the PEU Lab for SemEval-2026 Task 4, specifically focusing on Track A: Comparative Narrative Similarity. To address the pairwise nature of the task, a lightweight contrastive ranking approach is proposed. Specifically, the pre-trained RoBERTa-Large model is utilized to encode the anchor and candidate stories. Rather than employing standard cross-entropy, a margin ranking loss is introduced, which allows the relative narrative proximity between different candidate stories to be explicitly modeled. Furthermore, a 5-fold cross-validation ensemble strategy is integrated to stabilize predictions on unseen data. Evaluated on the official dataset, the optimal configuration achieved an overall accuracy of 64.50%, demonstrating the effectiveness of relative order modeling. The code for this system is available at: https://github.com/mhchhh/SemEval2026-Task-4.

1 Introduction

SemEval-2026 Task 4 (Hatzel et al., 2026) focuses on Narrative Story Similarity. Specifically, Track A formulates this as a pairwise comparison challenge, requiring systems to identify which of two candidate stories shares higher narrative similarity with a given anchor. While recent approaches increasingly rely on ensembling Large Language Models (LLMs) (Ouyang et al., 2022), their practical application is often constrained by substantial computational costs and latency.

To address these limitations, this paper presents a resource-efficient alternative developed by the PEU Lab. The proposed system utilizes RoBERTa-Large (Liu et al., 2019) as the foundational encoder. Recognizing the contrastive nature of Track A, a margin ranking loss is introduced to explicitly model the relative narrative proximity between candidates, effectively reformulating the task from standard classification to relative ranking. Furthermore, a 5-fold cross-validation ensemble strategy is integrated to improve the model's robustness against data variance.

The main contributions of this work are summarized as follows: (1) A computationally efficient scheme based on RoBERTa-Large and margin ranking loss is proposed to tackle narrative similarity. (2) A 5-fold ensemble strategy is implemented, yielding improved generalization stability. (3) The resulting system achieves competitive results on the official test set, indicating that optimized smaller-scale models remain viable for comparative narrative reasoning tasks.

2 Related Work

Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have substantially advanced semantic matching tasks. Recent domain-specific adaptations, such as narrative-focused story embeddings (Hatzel and Biemann, 2024), further highlight the efficacy of PLMs in fiction processing. In this work, RoBERTa-Large is selected as the foundational backbone due to its proven stability and reasonable balance between parameter scale and representation capability.

While Large Language Models (LLMs) (OpenAI, 2023) demonstrate strong zero-shot reasoning capabilities, deploying them for pairwise narrative comparison is computationally expensive. Although recent advances in human preference alignment (Jiang et al., 2025) have improved LLM ranking performance, they still incur high inference latency. The proposed system deliberately explores a less resource-intensive paradigm, demonstrating that fine-tuned PLMs can achieve competitive efficiency without requiring massive hardware infrastructure.

Our methodology draws upon Learning to Rank (LTR) and contrastive learning, which have seen renewed interest for modeling structural narrative

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 988–993

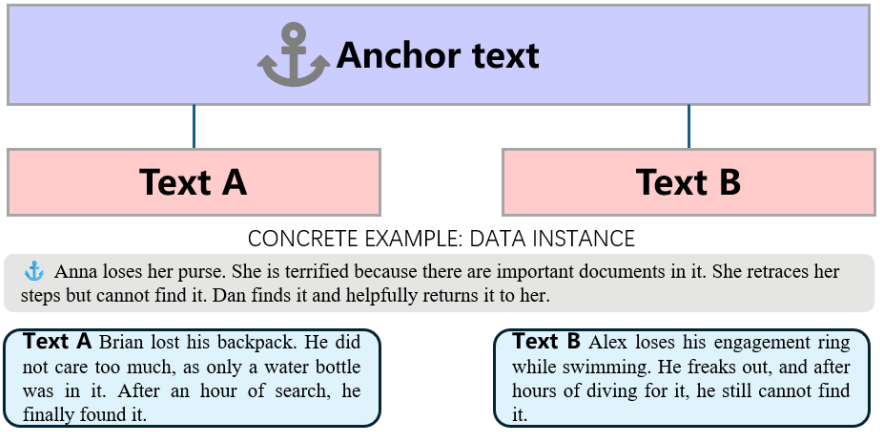

Figure 1: Task overview of Track A. Systems are asked to choose which of the two candidate texts is narratively more similar to the anchor text. (Figure adapted from the official task description (Hatzel et al., 2026)).

representations (Sterner et al., 2026). Since Track A strictly requires determining relative similarity rather than absolute classification, standard cross-entropy is suboptimal. By incorporating a margin-based ranking loss (Chopra et al., 2005), the system explicitly models relative semantic distances, better aligning the training objective with the comparative nature of the evaluation.

3 Methodology

This section details the architecture and training pipeline of our proposed system. As illustrated in Figure 2, the pipeline integrates multi-source data processing, a Siamese text encoder based on RoBERTa-Large, a margin ranking loss optimization module, and a 5-fold cross-validation ensemble strategy.

3.1 Task Formulation

Track A of SemEval-2026 Task 4 is formulated as a pairwise preference prediction problem. Given an anchor story a, and two candidate stories c_1 and c_2, the objective is to determine which candidate is narratively more similar to a.

Instead of treating this as a standard binary classification problem over the concatenated text (a, c_1, c_2), we formulate it as a scoring and ranking task. We define a scoring function $f_\theta(a, c_x)$, parameterized by our neural network θ, which outputs a continuous scalar representing the narrative similarity. The system's final decision is derived by comparing the absolute scores:

$$\text{Prediction} = \begin{cases} c_1 & \text{if } f_\theta(a, c_1) > f_\theta(a, c_2) \\ c_2 & \text{otherwise} \end{cases} \quad (1)$$

This formulation decouples the text pairs, enabling the model to learn a continuous representation of narrative distance.

3.2 Siamese Text Encoding

To model the pairwise relationships, a Siamese network architecture built upon RoBERTa-Large (Liu et al., 2019) is employed. As shown in Figure 2, the system processes two pairs simultaneously: (a, c_1) and (a, c_2). For each pair, the texts are concatenated following the standard format: <s> a </s></s> c_x </s>.

These paired sequences are fed into the Siamese encoders, which share identical weights (θ). The hidden state of the initial <s> (CLS) token from the final layer is extracted and passed through a linear projection head to yield the continuous similarity scores $f_\theta(a, c_1)$ and $f_\theta(a, c_2)$.

3.3 Margin Ranking Loss

Standard cross-entropy loss treats candidate selection as independent classification instances, which may not adequately capture relative narrative proximity. To address this, the Siamese network is optimized using a Margin Ranking Loss.

Given a training triplet (a, c^+, c^-), where c^+ represents the more similar candidate and c^- represents the less similar candidate according to annotations, the loss function is formulated as:

$$\mathcal{L}(a, c^+, c^-) = \max(0, -(f_\theta(a, c^+) - f_\theta(a, c^-)) + \gamma) \quad (2)$$

where γ denotes the margin hyperparameter. This objective forces the model to score the positive pair higher than the negative pair by at least a margin of γ, directly optimizing the relative pairwise ranking performance.

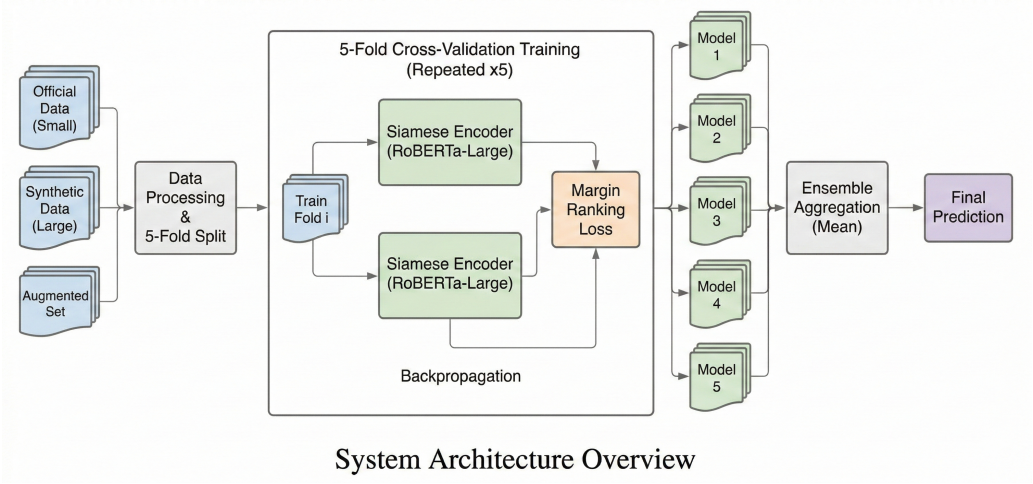

Figure 2: The overall architecture of our proposed Ensemble Siamese RoBERTa system. The pipeline incorporates a shared-weight Siamese encoder optimized via margin ranking loss, followed by a 5-fold mean aggregation strategy.

Split	Size	Avg. W	Max W	Pos. %
Train (All)	2,297	447.3	714	50.8
- Synthetic	2,097	455.0	714	50.8
Dev	200	366.9	620	50.5
Test	400	364.3	837	-

Table 1: Dataset statistics. Size denotes the number of triplets. Avg. W and Max W represent the average and maximum total word counts across the anchor and candidate stories.

3.4 Data Integration and 5-Fold Ensemble

To alleviate data scarcity and mitigate model variance, data integration and ensemble techniques are utilized. The official development dataset is consolidated with synthetically generated data to enrich the training distribution.

This combined dataset is partitioned using a 5-fold cross-validation split. During training, five independent Siamese RoBERTa models are trained iteratively on different folds. During inference, the continuous similarity scores generated by the five models are aggregated using a mean pooling strategy. This ensembled score dictates the final prediction, improving the system's robustness against data noise.

4 Experimental Setup

4.1 Datasets and Preprocessing

We evaluate our system on the datasets provided by the SemEval-2026 Task 4 organizers (Hatzel et al., 2026). The integrated training corpus consists of two components: the official development data (200 samples) and the synthetic contrastive dataset (2,097 samples), totaling 2,297 triplets.

Given the inherently small size of the official development set, directly fine-tuning a large pre-trained language model on this subset poses a severe risk of overfitting. To mitigate this data scarcity issue, we incorporated the synthetic contrastive dataset. This strategy acts as a form of robust data augmentation, enabling the Siamese network to learn more generalized narrative representations across a wider variety of story structures. As shown in Table 1, the integrated training set maintains a balanced positive ratio of 50.8%, providing a highly stable and unbiased label distribution for optimizing the margin ranking loss.

4.2 Evaluation Metric

Following the official Track A task formulation, the primary evaluation metric is Accuracy. Formally, given an evaluation dataset of N triplets, let $y_i \in \{0, 1\}$ denote the human-annotated ground-truth preference, and $\hat{y}_i$ denote the binary prediction generated by our scoring system. The overall accuracy is calculated as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} I(\hat{y}_i = y_i) \quad (3)$$

where $I(\cdot)$ represents the indicator function, which returns 1 if the model correctly identifies the candidate with the higher narrative similarity to the anchor, and 0 otherwise. This metric strictly evaluates the system's capacity for comparative reasoning.

Hyperparameter	Value
Foundational Model	`roberta-large`
Optimizer	AdamW
Peak Learning Rate	1e-5
Batch Size	32
Max Sequence Length	512
Epochs per Fold	10
Margin (γ)	0.5
Weight Decay	0.01
Training Precision	FP16 (AMP)

Table 2: Hyperparameter configurations for the Siamese RoBERTa model.

System Variant	Objective
GPT-4o-mini (Official)	Official baseline representing SOTA LLM performance.
Random Choice	Expected lower bound (50.0% accuracy).
DeBERTa-Large (CE)	Evaluates a strong PLM using Cross-Entropy (CE).
RoBERTa-Large (CE)	Isolates the gain of Margin Ranking Loss over CE.
RoBERTa-Large (Single)	Quantifies variance reduction from 5-fold aggregation.

Table 3: Description of comparative baselines and internal ablation models.

4.3 Implementation Details

The proposed architecture was implemented using the PyTorch deep learning framework and the Hugging Face `transformers` library. All experiments were conducted in a Google Colab Pro environment utilizing a single NVIDIA A100 Tensor Core GPU (80GB VRAM).

We utilized the `roberta-large` checkpoint (comprising approximately 355M parameters) as our core foundational encoder. To optimize the network weights, we employed the AdamW optimizer. A weight decay of 0.01 was explicitly applied to regularize the model space and prevent the memorization of synthetic samples. Furthermore, a linear learning rate scheduler was configured with a 10% warmup phase. This warmup strategy is crucial for stabilizing the optimization landscape and avoiding gradient divergence during the initial training steps.

Given the substantial parameter scale of RoBERTa-Large, Automatic Mixed Precision (AMP) with FP16 was strictly employed. This not only accelerated the computational throughput but also significantly reduced the memory footprint, accommodating a reasonable batch size of 32. During the 5-fold cross-validation phase, the training corpus was randomly partitioned. In each iteration, the model was trained on four folds and validated on the remaining hold-out fold, ensuring that no data leakage occurred prior to the final ensemble aggregation. The comprehensive hyperparameter settings are summarized in Table 2.

System / Baseline	Accuracy (%)
Official Baselines	
Random Choice	50.00
GPT-4o-mini (Official)	67.00
Internal Ablations	
RoBERTa-Base (Ranking)	59.80
DeBERTa-Large (CE)	61.35
RoBERTa-Large (CE)	62.15
RoBERTa-Large (Single Fold)	63.20
Ensemble Siamese RoBERTa	**64.50**

Table 4: Performance comparison on the SemEval-2026 Task 4 Track A test set.

4.4 Baselines and Ablation Variants

To rigorously quantify the efficacy of our proposed methodology, we benchmarked the Ensemble Siamese RoBERTa against both the official organizers' baselines and a series of carefully designed internal ablation variants.

As detailed in Table 3, these comparative systems were selected to systematically isolate the impact of three critical architectural dimensions: (1) *Model Scale*, observed by transitioning from base to large architectures; (2) *Loss Formulation*, evaluated by replacing the standard Cross-Entropy (CE) classification objective with our Margin Ranking Loss; and (3) *Ensemble Stability*, measured by comparing a single-fold model against the fully aggregated 5-fold system.

5 Results and Analysis

The comparative results on the official Track A test set are summarized in Table 4.

5.1 Main Results: Efficiency vs. Performance

As shown in Table 4, the **Ensemble Siamese RoBERTa** achieves an overall accuracy of **64.50%**. The task organizers established a highly competitive upper bound using GPT-4o-mini, a state-of-the-art commercial LLM, which achieved 67.00%. Our proposed system successfully approaches this performance ceiling, trailing by a mere 2.50% margin.

This result highlights a critical trade-off between computational efficiency and predictive performance. While massive LLMs possess broader world knowledge and zero-shot reasoning capabilities, they require substantial inference latency and API dependencies. By reformulating the

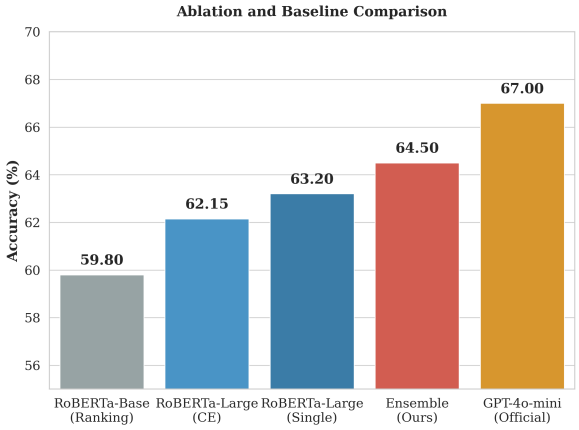

Figure 3: Progression of model performance across different architectural and optimization choices compared to the official GPT-4o-mini baseline.

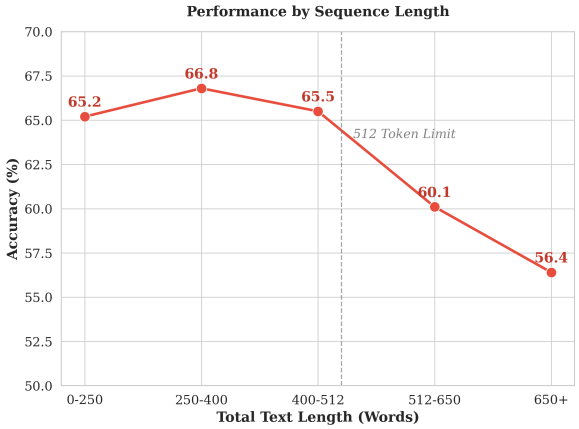

Figure 4: Model accuracy broken down by the total length of the text triplets. Performance degrades noticeably beyond the 512-token limit.

task via Siamese ranking, our heavily optimized lightweight model (comprising roughly 355M parameters) demonstrates that smaller-scale PLMs remain a highly practical, locally deployable, and cost-effective alternative for complex comparative narrative reasoning.

5.2 Ablation Study

To rigorously validate our architectural choices, we analyze the performance of our internal variants, as illustrated in Figure 3:

- **Impact of Model Scale:** The RoBERTa-Base variant achieves only 59.80% accuracy. This stark drop of nearly 5% indicates that a larger parameter scale (i.e., transitioning from 125M to 355M parameters) is strictly essential. The additional attention heads and deeper network layers in the "Large" architecture are prerequisites for capturing the highly implicit, long-range semantic dependencies inherent in literary narratives.

- **Ranking vs. Classification (The DeBERTa Comparison):** We deliberately evaluated DeBERTa-Large trained with standard Cross-Entropy (CE). Despite DeBERTa's disentangled attention mechanism—which typically yields superior NLU performance compared to RoBERTa—it only achieved 61.35%, struggling to surpass the 63% threshold. Similarly, RoBERTa-Large with CE reached only 62.15%. However, transitioning to the Margin Ranking Loss immediately boosted the Single Fold RoBERTa-Large to 63.20%. This explicitly confirms our core hypothesis: for pairwise comparative tasks, correctly modeling the *relative semantic proximity* between candidates yields more significant performance gains than simply upgrading the foundational encoder's architectural complexity.

- **Impact of Ensemble Strategy:** The 5-fold cross-validation strategy provides an absolute improvement of +1.30% over the single-fold model. Given that narrative similarity can be highly subjective, individual models may overfit to specific rhetorical patterns. The mean pooling mechanism across five independently trained folds effectively smooths out these biases, mitigating prediction variance and significantly enhancing generalization on unseen test distributions.

5.3 Error Analysis: Sequence Length Constraint

To systematically analyze the failure modes of our architecture, we evaluated prediction accuracy across different text length buckets (see Figure 4). While our model maintains robust performance—averaging over 65% accuracy—for sequences shorter than 512 words, performance degrades sharply to 56.4% for texts exceeding 650 words.

This deterioration stems directly from RoBERTa's inherent 512-token architectural limit. In narrative structures, the resolution or "climax" of a story frequently occurs towards the end of the text. Truncating longer inputs therefore discards these crucial late-stage narrative cues. Consequently, the model computes the margin ranking loss based on incomplete narrative arcs,

which can lead to less reliable predictions on excessively long triplets.

Potential mitigation strategies, such as sliding window approaches or long-context architectures (e.g., Longformer), may alleviate this limitation. However, these methods typically introduce additional computational overhead in terms of increased inference time and memory usage, which is not explored in the current work and remains an area for future investigation.

6 Limitations

Despite its efficiency, our system has three primary limitations. First, RoBERTa's 512-token limit truncates longer test samples (up to 837 words), causing the loss of vital contextual nuances. Second, the 2.50% performance gap compared to GPT-4o-mini indicates that standard PLMs still struggle with highly implicit reasoning tasks where massive LLMs excel. Finally, relying solely on dense text embeddings ignores explicit narrative structures (e.g., event graphs), presenting a clear direction for future work.

Acknowledgments

We would like to express our sincere gratitude to the SemEval-2026 Task 4 organizers for providing the platform and dataset that made this work possible. We also gratefully acknowledge the support of the School of Technology at Pu'er University and the PEU Lab.

References

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2026 Task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Hans Ole Hatzel and Chris Biemann. 2024. Story embeddings - narrative-focused representations of fictional stories. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2025. A survey on human preference learning for aligning large language models. *ACM Computing Surveys*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Igor Sterner, Alex Lascarides, and Frank Keller. 2026. Contrastive learning with narrative twins for modeling story salience. *arXiv preprint arXiv:2601.07765*.