Cryptix at SemEval-2026 Task 4: Zero-Shot Bi-Encoder Modeling for Narrative Story Similarity - A Sentence Transformer Approach

Sushmitha M[1] **Sarath Kumar P**[1,2] **Thanalaxmi S**[2]
Beulah A[1]
[1]Rajalakshmi Engineering College, Chennai, Tamilnadu, India
[2]sarathkumar98236@gmail.com

Abstract

This paper describes our submission to SemEval-2026 Task 4 on Narrative Story Similarity and Narrative Representation Learning. The task evaluates systems on their ability to model narrative similarity between story summaries using a contrastive triple-based setup. We participate in both Track A (comparative narrative similarity) and Track B (narrative representation learning).Our approach employs a zero-shot bi-encoder architecture based on the pretrained sentence-transformers/all-mpnet-base-v2 model. Narrative similarity is modeled using cosine similarity between independently encoded story embeddings. For Track A, the system selects the candidate story with higher cosine similarity to the anchor; for Track B, we directly output dense embedding representations evaluated via cosine distance.This system achieves 59.50% accuracy on Track A (rank 37) and 57.50% accuracy on Track B (rank 25). Through quantitative margin analysis and qualitative error categorization, we find that general-purpose semantic encoders capture topical similarity effectively but struggle to model higher-level narrative abstractions such as causal progression and outcome alignment.

1 Introduction

Narrative understanding remains a challenging problem in natural language processing, particularly when evaluating similarity beyond surface-level lexical overlap. The SemEval-2026 Task 4 on Narrative Story Similarity requires systems to assess similarity between story summaries based on abstract theme, course of action, and outcomes (Cer et al., 2017). Unlike sentence-level semantic similarity, narrative similarity demands modeling of global structure, event progression, and conceptual alignment across longer texts.

Motivated by the practical constraints of a zero-resource setting—where no labeled training triples for this specific task were available during system development. we adopt a zero-shot bi-encoder approach. Our hypothesis is that pretrained sentence embedding models, which have been optimized for general-purpose semantic similarity, encode sufficient semantic structure to serve as a competitive baseline for narrative comparison, even without task-specific adaptation. This choice also prioritizes reproducibility and efficiency: a single pretrained model encodes each story independently, and similarity is computed via a closed-form cosine operation requiring no inference-time cross-attention.Rather than proposing a novel architecture, this work provides a systematic empirical evaluation of a strong pretrained encoder in a zero-shot setting, establishing a reproducible baseline and identifying failure modes of embedding-based approaches for narrative similarity.

Our main contribution is a reproducible, fully unsupervised baseline system for narrative similarity, paired with systematic quantitative and qualitative error analysis that illuminates the specific failure modes of embedding-based approaches on this task. We deliberately evaluate this well-understood architecture in order to establish a clear upper bound for zero-shot methods and to identify the gap that task-specific fine-tuning or structural reasoning would need to close.

Through controlled ablation experiments and detailed error analysis, this paper investigates the strengths and limitations of embedding-based approaches for modeling narrative similarity.

2 Related Work

Narrative similarity is closely related to semantic textual similarity (STS) research (Cer et al., 2017) and contextual embedding models such as BERT (Devlin et al., 2019). Transformer-based embedding approaches including Sentence-BERT (Reimers and Gurevych, 2019) and MPNet (Song et al., 2020) have demonstrated strong performance

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1282–1287

in semantic similarity and retrieval tasks. These models learn dense representations optimized for cosine similarity comparisons in embedding space.

Advances in robust pretraining strategies such as RoBERTa (Liu et al., 2019) have further improved contextual representation quality. However, modeling narrative similarity extends beyond sentence-level semantic alignment and requires capturing global structure and event progression.

Prior work on narrative understanding emphasizes event structure modeling and script induction (Chambers and Jurafsky, 2008), highlighting the importance of temporal and causal relationships between events. Additionally, long-document transformer architectures such as Longformer (Beltagy et al., 2020) have been proposed to better handle extended contexts.

Despite these advances, many practical systems rely on pretrained bi-encoder architectures that prioritize semantic proximity over structural narrative reasoning. Our work evaluates the effectiveness of such embedding-based approaches for narrative-level similarity in the SemEval-2026 shared task setting.

3 Task Overview

SemEval-2026 Task 4 focuses on Narrative Story Similarity and Narrative Representation Learning. The task operationalizes narrative similarity as a comparative judgment problem over story summaries. Rather than evaluating surface-level lexical overlap, the task emphasizes higher-level narrative characteristics such as event progression, outcomes, and abstract thematic structure.

The dataset consists of story triples (a, c_1, c_2), where a is the anchor story and c_1, c_2 are candidate stories. Systems are required to determine which candidate story is narratively more similar to the anchor.

The shared task includes two evaluation tracks:

Track A: Comparative Narrative Similarity. Given a triple (a, c_1, c_2), systems must directly predict which candidate story is more narratively similar to the anchor story. The task is framed as a binary classification problem, and performance is measured using accuracy.

Track B: Narrative Representation Learning. In this track, systems generate dense embedding representations for individual stories. Similarity is evaluated indirectly by comparing cosine distances between embeddings. For each evaluation triple, the candidate story closer to the anchor in embedding space is considered the system's prediction. Performance is again measured using accuracy over the induced pairwise comparisons.

Both tracks are evaluated on manually annotated triples constructed to reflect human judgments of narrative similarity. The contrastive setup encourages systems to model abstract narrative structure rather than relying solely on surface lexical similarity.

4 Methodology

4.1 Model Architecture

The proposed system adopts a bi-encoder architecture based on the pretrained sentence-transformers/all-mpnet-base-v2 model. This model is built upon the MPNet transformer architecture (Song et al., 2020) and generates fixed-length 768-dimensional dense embeddings optimized for semantic similarity tasks.

Each story summary is encoded independently, without cross-attention between anchor and candidate texts. Input summaries are truncated to a maximum of 2000 characters prior to encoding to maintain consistency in input size while preserving most of the narrative content. Token-level representations are aggregated using the default mean pooling strategy provided by the SentenceTransformers framework (Reimers and Gurevych, 2019) to obtain a single vector representation per story.

The pretrained checkpoint is used in a zero-shot setting, and no task-specific fine-tuning is performed on the development or synthetic datasets. While the SemEval-2026 Task 4 dataset includes training triples that could be used for contrastive fine-tuning, this system intentionally operates in a zero-shot setting to isolate the representational capacity of the pretrained encoder without task-specific adaptation. Exploring supervised fine-tuning with triplet loss or Multiple Negatives Ranking (MNR) loss on the available training triples is left as future work.

4.2 System Architecture

The system uses a bi-encoder architecture built on the pretrained sentence-transformers/all-mpnet-base-v2 model to generate 768-dimensional story embeddings. Similarity between narratives is modeled using cosine similarity in embedding space.

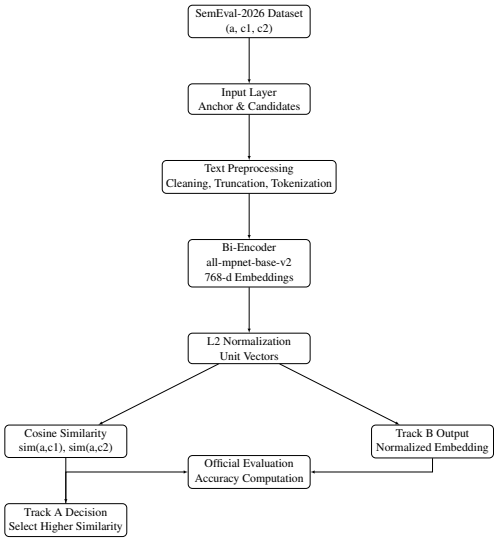

Figure 1: Architecture of the proposed system. The bi-encoder encodes each story independently using `all-mpnet-base-v2`, producing L2-normalized 768-dimensional embeddings. Track A selects the candidate with higher cosine similarity to the anchor; Track B outputs the normalized embedding directly.

4.3 Similarity Computation

Narrative similarity is modeled using cosine similarity between embedding vectors (Salton et al., 1975). Given an anchor story a and a candidate story c, similarity is computed as:

$$\text{sim}(a, c) = \frac{a \cdot c}{\|a\|\|c\|} \tag{1}$$

For Track A, given a triple (a, c_1, c_2), the system computes cosine similarity between the anchor and each candidate independently. The prediction is determined by selecting the candidate with the higher similarity score:

$$\hat{y} = \begin{cases} c_1 & \text{if } \text{sim}(a, c_1) > \text{sim}(a, c_2) \\ c_2 & \text{otherwise} \end{cases} \tag{2}$$

For Track B, the system outputs the 768-dimensional embedding representation for each story. The official evaluation script then computes cosine distances between embeddings and determines correctness based on whether the more narratively similar story is closer to the anchor in embedding space.

This formulation treats narrative similarity as a geometric proximity problem in dense embedding space.

5 Experimental Setup

5.1 Dataset and Evaluation

Experiments were conducted on the SemEval-2026 Task 4 dataset. The task consists of story triples (a, c_1, c_2), where a is the anchor story and c_1, c_2 are candidate stories.

In Track A, systems must predict which candidate is narratively more similar. In Track B, systems generate dense embeddings evaluated using cosine distance.

Performance is measured using classification accuracy.

5.2 Model Configuration

All experiments use the pretrained sentence-transformers/all-mpnet-base-v2 model. The model produces 768-dimensional embeddings optimized for semantic similarity. Story summaries are truncated to 2000 characters prior to encoding.

6 Official Results

Table 1 presents the official test set performance of Team Cryptix in SemEval-2026 Task 4. We report classification accuracy and final leaderboard rank for both Track A (comparative similarity) and Track B (representation learning).

Model	**Tr.A (%)**	**Tr.B (%)**	**Rk.A**	**Rk.B**
GPT-4o-mini (Baseline)	73.00	58.00	7	7
MiniLM-L6-v2 (Baseline)	62.00	56.00	40	40
Cryptix (ours)	59.50	57.50	37	25

Table 1: Official results. Tr.A = Track A accuracy, Tr.B = Track B accuracy, Rk = Rank.

The results indicate that zero-shot bi-encoder methods can exceed chance-level performance (50%) on this task, confirming that pretrained semantic representations carry useful narrative similarity signal. Track B ranks higher than Track A (rank 25 vs. rank 37), suggesting that the raw embedding representations generalize somewhat better across evaluation triples than the hard comparative decisions derived from them.

7 Ablation Study

We conduct ablation experiments on the development set to assess the contribution of individual design choices.

Removing L2 normalization does not affect development accuracy, which is expected because

Model Variant	Dev Accuracy (%)
MPNet (Baseline)	61.5
No Normalization	61.5
Sentence-Level Averaging	56.5
Truncate at 1000 Characters	62.0
RoBERTa-Large Encoder	57.5

Table 2: Ablation study on the development set.

cosine similarity is scale-invariant. Sentence-level averaging reduces accuracy by 5 points, suggesting that full-text mean pooling captures narrative coherence more effectively than averaging individual sentence embeddings. It should be noted that differences within 1–2% between ablation variants (e.g., 61.5% vs 62.0%) should be interpreted with caution, as these margins may fall within natural variance and may not be statistically significant without confidence interval analysis or significance testing such as McNemar's test. These results are therefore treated as indicative trends rather than conclusive performance differences.Truncating at 1000 characters yields a marginal 0.5-point improvement, possibly because shorter inputs reduce noise from less informative narrative tail content. The RoBERTa-Large encoder underperforms the MPNet baseline by 4 points, indicating that optimization objective matters more than model size for this task.

8 Error Analysis

To better understand system behavior, we conduct a detailed analysis on the development set, examining prediction confidence characteristics and recurring qualitative error patterns.

8.1 Margin-Based Analysis

We define the prediction margin as the absolute difference between cosine similarity scores of the two candidate stories:

$$\text{Margin} = |\text{sim}(a, c_1) - \text{sim}(a, c_2)| \quad (3)$$

Table 3 reports the average margin for correct and incorrect predictions.

Correct predictions exhibit nearly double the average margin of incorrect ones. This indicates that model errors predominantly occur in low-confidence scenarios where similarity differences between candidates are subtle. These findings suggest that narrative similarity distinctions often lie near the decision boundary in embedding space, making them difficult to separate using general-purpose semantic encoders.

Prediction Type	**Avg. Margin**	**Count**
Correct Predictions	0.1101	298
Incorrect Predictions	0.0606	102

Table 3: Average cosine similarity margin for correct and incorrect predictions on the development set.

8.2 Length-Based Analysis

To assess whether input length influences prediction accuracy, we compare the average summary lengths of correctly and incorrectly classified samples.

Prediction Type	**Avg. Length (chars)**
Correct Predictions	706.71
Incorrect Predictions	705.06

Table 4: Average summary length for correct and incorrect predictions.

As shown in Table 4, there is no substantial difference between the lengths of correctly and incorrectly classified stories. This suggests that prediction errors are not driven by input size or truncation effects, but rather by deeper semantic or structural challenges inherent to narrative similarity.

8.3 Qualitative Categorization of Errors

To further understand the nature of model failures, we manually inspected a subset of incorrectly predicted development samples and categorized them based on narrative characteristics. The resulting categories are shown in Table 5.

Error Category	**Proportion (%)**
Theme Overlap but Structural Difference	34
Structural Similarity with Lexical Variation	29
Outcome Divergence	21
Surface Lexical Bias	16

Table 5: Qualitative categorization of development set errors.

We observe several recurring patterns:

Theme Overlap but Structural Difference. In many cases, the model favors stories sharing similar motifs (e.g., revenge, betrayal, romance) even when the sequence of events or causal progression differs substantially. This suggests that the model prioritizes topical similarity over narrative structure.

Structural Similarity with Lexical Variation. Some errors occur when two stories share a similar event progression but use different vocabulary and settings. In these cases, the model fails to recognize deeper structural parallels due to limited sensitivity to abstract event patterns.

Outcome Divergence. In several triples, stories share initial developments but differ in final resolution. The model occasionally overemphasizes early narrative similarity and fails to account for outcome alignment.

Surface Lexical Bias. A subset of errors appears driven by higher lexical or semantic overlap at the surface level, even when narrative structure does not align closely with the anchor story.

Overall, these findings indicate that while pretrained semantic encoders capture topical similarity effectively, they struggle to model higher-level narrative abstractions such as causal structure and outcome alignment.

9 Discussion

The results highlight both the strengths and limitations of using a pretrained bi-encoder for narrative similarity. While the sentence-transformers/all-mpnet-base-v2 model effectively captures topical and semantic similarity, it struggles with deeper narrative abstractions such as causal progression and outcome alignment.

Development–test accuracy gap The development accuracy of 62.0% compared to test performance of 59.50% on Track A and 57.50% on Track B suggests that the hidden evaluation triples involve more subtle or structurally complex narrative distinctions.

Track B analysis Track B ranks higher than Track A (rank 25 vs. rank 37), despite reporting lower absolute accuracy (57.50% vs. 59.50%). This apparent discrepancy is explained by the difference in the two tracks' evaluation populations: Track B measures how well the raw embedding geometry reflects human narrative similarity judgments, whereas Track A measures hard binary decisions. The better relative standing in Track B suggests that while the embedding space does encode some narrative similarity structure, the hard comparative decision in Track A amplifies errors near the decision boundary—consistent with our margin analysis showing that most errors occur in low-margin cases. Improving Track A accuracy would therefore require either better-calibrated similarity scores or a re-ranking mechanism that accounts for uncertainty.

Baseline comparison Compared to the official baselines, our zero-shot bi-encoder system falls notably below GPT-4o-mini on Track A (59.50% vs 73.00%), suggesting that large generative models with instruction-following capabilities hold a significant advantage in comparative narrative judgment tasks. However, our system performs comparably to the all-MiniLM-L6-v2 baseline on both tracks, with a marginal improvement on Track B (57.50% vs 56.00%), indicating that the larger all-mpnet-base-v2 encoder provides slight representational benefit for embedding-based evaluation despite operating under identical zero-shot conditions.

Implications for future work While pretrained embedding models provide a competitive zero-shot baseline, future work should explore several concrete directions: contrastive fine-tuning using the available training triples with triplet loss or Multiple Negatives Ranking (MNR) loss to adapt the encoder to narrative-specific similarity, cross-encoder architectures that allow joint attention between anchor and candidate stories, enabling richer interaction modeling, structured event graph representations that explicitly capture causal and temporal relationships between narrative events, and ensemble approaches combining embedding-based similarity with generative model judgments, motivated by the strong Track A performance observed in GPT-4o-mini.

10 Conclusion

We presented a zero-shot bi-encoder system for SemEval-2026 Task 4 on Narrative Story Similarity. The system encodes story summaries independently using the pretrained sentence-transformers/all-mpnet-base-v2 model and ranks candidate stories by cosine similarity to the anchor. Our approach achieved 59.50% accuracy on Track A (rank 37) and 57.50% on Track B (rank 25).The primary finding is that pretrained

semantic encoders provide a reasonable zero-shot baseline substantially above chance but exhibit a systematic gap between development and test performance. Margin analysis reveals that errors are concentrated in low-confidence cases near the similarity decision boundary, and qualitative inspection identifies theme-overlap-with-structural-difference as the dominant failure mode. Together, these results demonstrate that narrative similarity is not reducible to topical or lexical similarity: it requires modeling event sequencing, causal relationships, and outcome alignment, none of which are explicitly encoded by current general-purpose sentence embeddings.These findings establish a clear picture of where embedding-based methods succeed and where they fail on narrative comparison tasks, providing a useful reference point for future systems that incorporate structural reasoning or task-specific fine-tuning.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In *Advances in Neural Information Processing Systems*, volume 33, pages 16156–16170.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo 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.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.