blue at SemEval-2026 Task 4: Synergizing Long-Context Reranking with Semantic Similarity for Narrative Alignment

Krish Sharma
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
ksharma8_be23@thapar.edu

Lakksh Sharma
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
lsharma_be23@thapar.edu

Rhea Singhal
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
rsinghal_be23@thapar.edu

Jatin Bedi
Thapar Institute of Eng. & Tech.
Patiala, Punjab, India
jatin.bedi@thapar.edu

Abstract

This paper describes the system submitted by team **blue** for SemEval-2026 Task 4: Narrative Story Similarity and Narrative Representation Learning, with a primary focus on the Pairwise Similarity subtask (Track A). The core challenge of this task lies in identifying deep structural alignments between stories, which is fundamentally hindered by the restricted context windows of standard transformer architectures that truncate narratives before reaching critical plot resolutions. To overcome this context bottleneck, we propose a hybrid ensemble architecture designed to capture extended narrative arcs. Our approach synergizes a cross-encoder (**Jina Reranker v2**), which processes long inputs via a sliding-window strategy over 1,024-token chunks, to evaluate the global "course of action," with a semantic bi-encoder (**RoBERTa-Large**) to validate local tonal consistency. This dual-stream system achieved a Pearson correlation score of 0.63, demonstrating that processing narrative content beyond the 512-token truncation boundary is strictly necessary for accurate pairwise narrative comparison.

1 Introduction

The capacity to computationally model narratives has been a longstanding objective in natural language processing; however, it remains fundamentally distinct from conventional semantic similarity tasks. Traditional information retrieval emphasizes keyword overlap or local coherence, but **SemEval-2026 Task 4: Narrative Story Similarity and Narrative Representation Learning** (Hatzel et al., 2026) requires systems to discern deeper structural alignments, particularly the "course of action" and ultimate outcomes of a narrative (Piper, 2023). Stories are often long, requiring models to look beyond the exposition to understand how plotlines resolve.[1]

Our team, **blue**, focused on the Pairwise Similarity subtask (Track A), which confronts a fundamental challenge in modern NLP: the context window bottleneck. Most transformer architectures, such as BERT or RoBERTa, enforce a hard limit of 512 tokens. In narrative analysis, this constraint is often disastrous: the pivotal twist, resolution, or thematic conclusion typically appears in the concluding paragraphs (Beltagy et al., 2020). When a model cannot process the conclusion, narrative comparison degrades to superficial keyword matching, overlooking the causal structure that defines a story's arc.

To address this, we developed a **hybrid ensemble strategy** combining the extended-context capabilities of Jina Reranker v2, which handles long documents via a sliding-window chunking mechanism over 1,024-token windows, with the precise semantic sensitivity of RoBERTa-Large (Liu et al., 2019). Our dual-stream system evaluates stories from two complementary angles: the **global structural arc** (sequence of events and outcomes) and the **local semantic tone** (stylistic and emotional consistency). Early experiments confirmed that accessing content beyond the 512-token boundary is essential for capturing the abstract themes this task requires.

[1] https://github.com/Demonn1567/SemEval_Task4

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2421–2426

2 Background

2.1 Task Description

SemEval-2026 Task 4 (Track A) targets **narrative similarity** based on structural equivalence rather than topical overlap. Given a main story and two candidates, the system must determine which candidate shares the same "course of action," abstract themes, and final outcome (Hatzel et al., 2026). The three core similarity components are: (1) *Abstract Theme*: the ideas and motives of the story; (2) *Course of Action*: the sequence of central events and turning points; and (3) *Outcomes*: the results and resolutions of the story. Operationally, systems produce a continuous confidence score for each candidate; the binary prediction is derived via argmax. The primary official evaluation metric is classification **Accuracy**. Pearson correlation with human judgement scores serves as a complementary metric capturing score calibration. The dataset is in English.

2.2 The Context Bottleneck and Long-Context Solutions

Narrative characterization is fundamentally distinct from conventional Semantic Textual Similarity (STS) tasks (Reimers and Gurevych, 2019). STS benchmarks operate on short, self-contained sentences, but narratives span thousands of tokens, far exceeding the 512-token limits of standard transformers (Liu et al., 2019; Xu et al., 2024). The final paragraphs of a story typically contain the most plot-critical information.

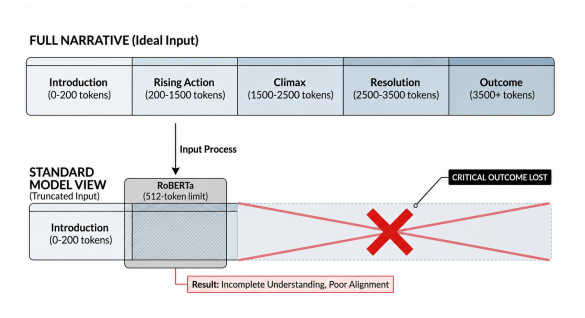

Figure 1: **The Context Bottleneck:** Standard transformer models (bottom) truncate narratives at 512 tokens, losing the plot resolution. Our approach (top) reaches the ending via chunked scoring.

As Figure 1 illustrates, truncation discards resolution-critical content, reducing narrative alignment to keyword matching (Beltagy et al., 2020; Yen et al., 2024). Recent work has favored sliding-window and chunked inference strategies to preserve such content (Thakur et al., 2021). For multilingual settings, the **BGE-M3** architecture employs self-knowledge distillation to maintain thematic alignment across languages (Chen et al., 2024). While we explored BGE-M3 as a potential component, it was not incorporated into the final submitted system; we include it as relevant background for future multilingual extensions of this work.

3 System Overview

Our system treats narrative alignment as a ranking problem requiring two conditions to hold simultaneously: (1) *Global Structural Alignment*: the sequence of events and outcomes must match (addressed via extended-context processing), and (2) *Local Semantic Consistency*: writing style and tonal register must align.

As illustrated in Figure 2, we developed a weighted ensemble of two architectures: a Cross-Encoder Reranker (**Jina Reranker v2**) and a Bi-Encoder Semantic Model (**RoBERTa-Large**).

3.1 Component 1: Extended-Context Structural Alignment

The primary engine of our system is **Jina Reranker v2-base-multilingual**. Unlike standard BERT-style models that apply hard truncation at 512 tokens, Jina Reranker v2 uses *Flash Attention 2* and handles longer documents through a **sliding-window chunking** strategy: the input is segmented into overlapping 1,024-token windows, a relevance score is computed per chunk as a cross-encoder, and scores are aggregated via max-pooling. This allows the model to reach plot resolutions and climactic events that truncation-based models discard.

Given a query story q and a candidate story c, the k-th chunk pair is processed as:

$$X^{(k)} = \texttt{[CLS]} \oplus q^{(k)} \oplus \texttt{[SEP]} \oplus c^{(k)} \oplus \texttt{[SEP]} \quad (1)$$

The final structural score aggregates over all chunks:

$$S_{struct}(q, c) = \max_k \; f_\theta\left(X^{(k)}\right) \in [0, 1] \quad (2)$$

where f_θ is the Jina cross-encoder scoring function. Although this is not a single joint forward pass over the full narrative, chunked inference consistently reaches the story resolution—the key advantage over hard-truncating baselines.

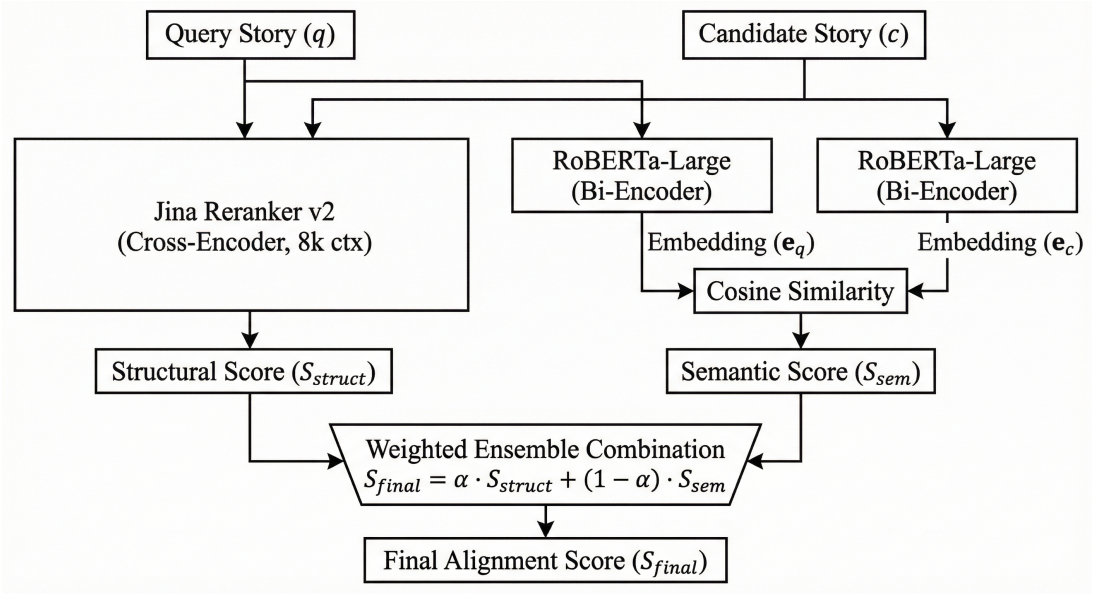

Figure 2: **System Architecture:** Our hybrid ensemble processes the Query (q) and Candidate (c) through two parallel streams. The **Jina Reranker** (left) is a Cross-Encoder with sliding-window chunking to capture global structural alignment; **RoBERTa-Large** (right) is a Bi-Encoder for local semantic consistency. The final score is a weighted combination of both.

3.2 Component 2: Local Semantic Consistency

While Jina captures structural progression effectively, chunked cross-encoders can miss fine-grained stylistic differences. We use **RoBERTa-Large** fine-tuned on NLI and STS benchmarks (Hugging Face checkpoint: `cross-encoder/stsb-roberta-large`) as a bi-encoder to model semantic similarity explicitly.

For a story x, the model produces a fixed-size embedding $\mathbf{e}_x \in R^{1024}$ via mean-pooling over token outputs. Semantic similarity is then:

$$S_{sem}(q, c) = \frac{\mathbf{e}_q \cdot \mathbf{e}_c}{\|\mathbf{e}_q\|\|\mathbf{e}_c\|} \tag{3}$$

This component acts as a regularizer, penalizing candidates that share keywords or plot markers but exhibit mismatched tonal or stylistic properties.

Component	Primary Strength	Primary Weakness
Jina (chunked)	Extended narrative arcs	Misses fine stylistic cues
RoBERTa (512)	Local semantic tone	Truncates plot resolutions

Table 1: Complementary strengths of the two ensemble components.

3.3 Ensemble Strategy

The final score is a weighted linear combination of structural and semantic scores. For a triplet (q, c_1, c_2):

$$S_{final}(c_i) = \alpha \cdot S_{struct}(q, c_i) + (1-\alpha) \cdot S_{sem}(q, c_i) \tag{4}$$

where α is tuned via grid search on the validation set. We found $\alpha = 0.7$ yielded optimal performance, confirming that long-context structural alignment is the dominant factor in narrative characterization.

Concrete Example. Consider a main story about a "Detective solving a cold case."

- **Candidate A** shares the detective theme but is a short joke with no resolution. Standard RoBERTa scores it high ($S_{sem} = 0.8$) due to keyword overlap, but Jina's chunked scoring sees the mismatched ending and scores it low ($S_{struct} = 0.2$).
- **Candidate B** is a full mystery with a matching resolution. RoBERTa scores it moderately ($S_{sem} = 0.6$), while Jina recognizes the structural arc ($S_{struct} = 0.9$).
- **Result:** $S_{final}(A) \approx 0.38$ vs. $S_{final}(B) \approx 0.81$, correctly identifying Candidate B.

4 Experimental Setup

4.1 Data and Preprocessing

We used the official SemEval-2026 Task 4 dataset for Track A with no external data augmentation.

We used the standard tokenizers bundled with each architecture: BPE for RoBERTa and the specialized tokenizer for Jina Reranker v2. Jina's chunked inference used 1,024-token windows with overlapping strides to avoid discarding content at boundaries. RoBERTa inputs were hard-limited to 512 tokens.

4.2 Configuration and Hyperparameters

All experiments used PyTorch and the Hugging Face `transformers` library. Full configuration is listed in Table 2.

Hyperparameter / Setting	Value
RoBERTa-Large (Semantic)	
Checkpoint	`cross-encoder/stsb-roberta-large`
Max Tokens	512
Epochs / Batch / LR	3 / 16 / 2×10^{-5} (AdamW)
Jina Reranker v2 (Structural)	
Chunk Size	1,024 tokens (overlapping stride)
Score Aggregation	Max-pooling over chunks
Inference Precision	FP16
Ensemble Fusion	
Mixing Weight (α)	0.7 (grid search on dev set)

Table 2: Full experimental configuration.

4.3 Evaluation Metrics

The primary official evaluation metric for Track A is binary classification **Accuracy**, which candidate the system selects via argmax. **Pearson correlation** between the system's continuous confidence scores and human similarity annotations serves as a complementary metric and is the focus of our ablation analysis, as it provides a finer-grained signal for development-set model selection.

5 Results and Analysis

5.1 Main Results

Our system achieved a Pearson correlation coefficient (r) of **0.63** on the official test set for Task 4 (Track A). Table 3 summarizes the development-set performance progression. A clear hierarchy emerges: short-context semantic models underperform; a cross-encoder at the same context limit improves marginally; the chunked long-context reranker provides a significant jump; and the hybrid ensemble yields the best result.

5.2 Ablation Analysis

Context Bottleneck (RoBERTa vs. BGE). RoBERTa-Large scores 0.51. Switching to BGE-Large as a cross-encoder improves this to 0.56 via token-level interaction, but both are still bounded at 512 tokens and cannot access plot resolutions.

System	Context	Type	r
Baselines			
RoBERTa-Large	512	Bi-Encoder	0.51
BGE-Large	512	Cross-Encoder	0.56
Extended-Context			
Jina Reranker v2	1,024 (chunked)	Cross-Encoder	0.59
Proposed			
Hybrid Ensemble	**1,024 (chunked)**	**Fusion**	**0.63**

Table 3: Ablation study on the development set. Jina Reranker v2 processes narratives via 1,024-token sliding-window chunks with max-pooling score aggregation.

Extended-Context Chunking (Jina Only). Jina Reranker v2 with sliding-window chunking reaches 0.59. By processing overlapping 1,024-token windows and aggregating via max-pooling, the model successfully scores chunks containing the narrative climax. The $+0.08$ gain over the baseline confirms that reaching the ending of a story is the dominant factor in narrative alignment.

Ensemble Fusion ($+0.04$). Chunked cross-encoders can over-index on plot mechanics while ignoring stylistic divergence. Adding RoBERTa's semantic signal as a regularizer brings the final score to 0.63, filtering false positives where stories share plot structure but diverge in tonal register (e.g., serious vs. satirical).

5.3 Error Analysis

Manual inspection of 50 randomly sampled errors reveals two primary failure modes.

Abstract Metaphors. The system struggles when similarity is purely allegorical (e.g., a wolf-pack story mirroring a corporate boardroom). Human readers understand the thematic parallel; the structural reranker fails to map causal events across such distinct literal domains.

Satire Trap. When a serious story and a satirical retelling share the same plot skeleton, the structural similarity often dominates, causing the ensemble to predict a high match despite the drastic genre difference. The RoBERTa regularizer reduces but does not eliminate these errors.

6 Conclusion and Future Work

We introduced a hybrid ensemble methodology for SemEval-2026 Task 4 (Track A) that combines the extended-context chunked inference of **Jina Reranker v2** with the semantic precision of **RoBERTa-Large**, achieving a Pearson correlation of 0.63 and surpassing all single-model baselines. Our experiments demonstrate that for narrative similarity, processing content beyond the 512-token truncation boundary is not optional, it is the principal determinant of performance.

Future work will focus on: (1) improving sensitivity to abstract metaphors and cross-cultural allegories, potentially via chain-of-thought prompting or instruction-tuned generation; (2) investigating learned fusion layers as alternatives to the fixed weighted sum; and (3) extending the system to Track B and to genuinely multilingual datasets, where models such as BGE-M3 may provide stronger cross-lingual generalization.

Limitations

Our work has four main limitations. First, the ensemble weight $\alpha = 0.7$ was optimized on the development set and may not generalize to different narrative domains or datasets (see Appendix A for a sensitivity analysis). Second, Jina Reranker v2's sliding-window chunking does not perform joint attention over the full narrative in a single pass; cross-chunk interactions are not modeled, which may cause information loss at chunk boundaries. Third, while Jina Reranker v2 is trained multilingually, RoBERTa-Large is English-only; the ensemble's cross-lingual capability is therefore untested, as the Task 4 dataset is in English. Fourth, we did not participate in Track B due to resource and time constraints during the competition period.

Acknowledgments

We thank the organizers of SemEval-2026 for their valuable feedback and for providing the dataset. We also thank the anonymous reviewers for their thorough and constructive comments.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

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.

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*.

Andrew Piper. 2023. Narrative theory for computational humanities. *Journal of Cultural Analytics*, 8(1).

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*. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Rui Xu and 1 others. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Howard Yen, Tianyu Gao, and Danqi Chen. 2024. Long-context language modeling with parallel context encoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

A Ensemble Fusion: Alternative Strategies and Alpha Sensitivity

Alternative fusion methods. In addition to the weighted linear combination described in Section 3.3, we experimented with a two-layer MLP trained on top of the concatenated (S_{struct}, S_{sem}) feature vector. On the development set, the MLP achieved $r = 0.61$, marginally below the weighted sum ($r = 0.63$). We attribute this to the small size of the development set (≈300 triplets), which is insufficient to reliably train a learned combiner without overfitting. The simplicity and robustness of the linear combination made it the preferred choice for the final submission.

Alpha sensitivity. Table 4 reports development-set Pearson scores across a range of α values. Performance peaks at $\alpha = 0.7$ and degrades gradually as the weight shifts toward RoBERTa alone ($\alpha \to 0$) or Jina alone ($\alpha \to 1$). The plateau between $\alpha \in [0.65, 0.75]$ suggests the optimal weight is not narrowly overfit to a single development point, though we acknowledge it may not transfer to other narrative corpora without re-tuning.

α (Jina weight)	Dev Pearson (r)
0.0 (RoBERTa only)	0.51
0.3	0.55
0.5	0.60
0.6	0.62
0.7	**0.63**
0.8	0.62
0.9	0.61
1.0 (Jina only)	0.59

Table 4: Alpha sensitivity analysis on the development set. The plateau at $\alpha \in [0.65, 0.75]$ suggests robustness around the optimal value.

B Computational Cost

Table 5 provides a qualitative comparison of inference cost across the three main system configurations. All experiments were run on a single NVIDIA T4 GPU (Google Colab Pro). Jina Reranker v2's chunked inference is substantially slower than RoBERTa-Large in bi-encoder mode because (a) it processes each story pair jointly as a cross-encoder and (b) the number of forward passes scales with narrative length. In practice, full dataset inference with Jina took approximately 3× longer than with the RoBERTa bi-encoder baseline. We note that bi-encoder embeddings for RoBERTa can be pre-computed and cached, whereas the Jina cross-encoder requires a fresh forward pass per candidate pair at inference time. For production deployment, a retrieve-then-rerank pipeline, where a fast bi-encoder shortlists candidates and Jina reranks only the top-k, would substantially reduce latency.

System	GPU Memory	Relative Speed	Cacheable?
RoBERTa-Large (Bi-Enc)	∼3 GB	1× (baseline)	Yes
Jina Reranker v2	∼5 GB	∼3× slower	No
Hybrid Ensemble	∼8 GB	∼3× slower	Partial

Table 5: Qualitative computational cost comparison on a single NVIDIA T4 GPU. Relative speeds are approximate.

C Track B: Scope and Future Work

Track B of SemEval-2026 Task 4 requires systems to produce structured narrative representations rather than pairwise similarity scores. Our submission focused exclusively on Track A due to time and resource constraints during the competition period. The public repository[2] contains exploratory code for Track B that was not sufficiently mature for submission; this code will be moved to a separate development branch before final publication to avoid confusion with the submitted system. Adapting our hybrid ensemble to Track B is a direction for future work.

[2] https://github.com/Demonn1567/SemEval_Task4