DUTIR at SemEval-2026 Task 4: Narrative Story Similarity and Narrative Representation Learning

Tala Borjigin
Dalian University of Technology
DUTIR
pmbczq@foxmail.com

Liang Yang
Dalian University of Technology
DUTIR
liang@dlut.edu.cn

Abstract

This paper presents our approach for SemEval 2026 Task 4. Our method leverages a large language model fine-tuned via Low-Rank Adaptation, incorporates data cleaning, and employs a multi-prompt strategy, all trained on the official synthetic dataset. Evaluated on Track A, our system achieved an official score of 0.70, representing a reasonable performance under the given task constraints. In addition, we explore an alternative contrastive learning framework originally designed for Track B, where narrative-structure embeddings are learned and subsequently applied to Track A via similarity comparisons. Our analysis suggests that direct supervised adaptation may be more suitable for narrative reasoning tasks.[1]

1 Introduction

Recent advances in large language models (LLMs) have led to substantial improvements across a wide range of natural language processing tasks. However, despite their strong surface-level performance, LLMs often struggle with reasoning over extended narratives, where correct predictions require modeling relationships across multiple sentences rather than relying on local lexical cues. The task organizers highlight the importance of narrative-level semantic understanding and motivate further research in this direction.(Hatzel et al., 2026; Hatzel and Biemann, 2024)

Narrative structure analysis plays an important role in enhancing the reasoning capabilities of LLMs, as it encourages models to focus on intersentential relations such as temporal progression, causality, and discourse coherence. Models trained predominantly through supervised pretraining tend to exploit surface patterns—such as named entities or lexical co-occurrences—which can limit their ability to capture the underlying logical structure of narratives. (Subbiah et al., 2024) found that LLMs make faithfulness mistakes in over 50% of summaries and struggle with specificity and interpretation of difficult subtext. Experiments in (Wang and Kreminski, 2025) show that GPT-4 tier LLMs can generate causally sound stories at small scales, but planning with character intentionality and dramatic conflict remains challenging, requiring LLMs trained with reinforcement learning for complex reasoning. Investigating how LLMs represent and utilize narrative structure is important for improving abstraction and reasoning in complex language understanding scenarios.

Specifically, the task is organized into two tracks, each addressing a complementary aspect of narrative reasoning.

1.1 Track A: Narrative Similarity Classification

Track A formulates narrative similarity assessment as a binary classification problem. For each instance, in narrative space S, the system is provided with a narrative context $s_c \in S$ along with two candidate texts $s_a, s_b \in S$. By representing each input instance as an ordered triplet (s_c, s_a, s_b) drawn from the product space S^3, the objective is to determine whether s_a exhibits a higher degree of semantic similarity to the context s_c than s_b. That is, the system aims to learn a decision mapping:

$$f_a : S^3 \rightarrow \{A, B\}$$

1.2 Track B: Narrative Representation Learning

Track B focuses on learning dense vector representations for narratives. Systems are required to encode textual narratives S from narrative space into continuous embedding spaces $\mathbb{R}^d$ such that semantically related narratives are positioned closer to each other. The quality of these representations

[1] https://github.com/2B-dada/semeval2026-t4

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3010–3014

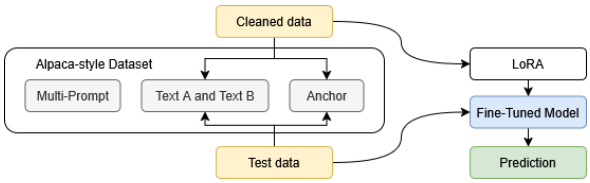

Figure 1: Overview of the proposed framework for narrative similarity classification.

is evaluated based on their ability to preserve narrative structure, semantic coherence, and relational similarity. Similarly, the system learns:

$$f_b : S \rightarrow \mathbb{R}^d$$

Our system is based on Qwen3-4B-Instruct-2507(QwenTeam, 2025) fine-tuned with LoRA(Hu et al., 2021). The best submitted model is trained on the official synthetic dataset until no further loss decrease or performance improvement. The dataset is cleaned to eliminate empty or invalid instances. In the training stage, we use the official synthetic dataset and the dev set for validation. We utilize multiple prompts so that the model will not be affected by poorly synthesized prompts or unwanted overfitting.

Our contributions are summarized as follows:

- We present a LoRA-based fine-tuning framework for narrative-level semantic similarity, combining data cleaning and a multi-prompt strategy to mitigate overfitting.
- We conduct experiments on the official dataset and report performance for task 4 obtained without using external data, achieving an official score of 0.70.
- We further explore a contrastive learning approach. Our analysis of comparisons provides empirical observations into the relative effectiveness of supervised adaptation for narrative reasoning.

2 System Overview

Our system is implemented using the Llama-Factory(Zheng et al., 2024) framework and follows a supervised fine-tuning (SFT) pipeline with Low-Rank Adaptation (LoRA). The overall design emphasizes simplicity and reproducibility, focusing on effective adaptation of a pretrained instruction-following language model to narrative similarity classification.

2.1 Base Model and Training Framework

We adopt Qwen3-4B-Instruct as the base model. Training and fine-tuning are conducted using Llama-Factory, which provides a unified interface for parameter-efficient fine-tuning and large-scale language model training. Remote code execution is enabled to ensure compatibility with the model architecture.

2.2 Input Formatting and Prompt Design

Training instances are formatted in an Alpaca-style instruction-following format, consisting of an instruction, an input, and an output. The instruction explicitly asks the model to determine which candidate narrative (A or B) is more similar to the given main story based on narrative structure, and the output is restricted to a single label (`A` or `B`).

The input includes the full narrative context followed by two candidate stories. This design encourages the model to compare narrative-level semantics rather than relying on local lexical cues. To reduce overfitting and potentially improve stability, inspired by (Lester et al., 2021; Lu et al., 2022; Wang et al., 2021), we employ multiple prompt variants that preserve the same underlying task semantics while introducing surface-level diversity.

2.3 Exploratory Similarity-Based Baseline

In addition to our primary supervised fine-tuning approach, we implement a similarity-based method as an auxiliary baseline to better understand the effectiveness of different modeling strategies. The objective of this approach is to learn dense narrative embeddings that capture structural and semantic similarities between narratives.

We initialize the model with the same pretrained backbone as the Track A system and apply LoRA for parameter-efficient feature extraction. LoRA modules are attached to the projection layers of the attention mechanism, enabling efficient adaptation while keeping the majority of model parameters frozen.

Narrative representations are obtained by encoding each input text and computing a mean pooling over the last hidden states. The model is trained using a triplet loss objective(Schroff et al., 2015), where an anchor narrative is encouraged to be closer to a positive example than to a negative one in the embedding space. Cosine similarity is used as the distance metric, and a fixed margin is applied during training.

During training, the model processes triplets consisting of anchor, positive, and negative narratives constructed from the official dataset. After training, the learned narrative embeddings are applied to Track A by comparing similarity scores between candidate narratives. However, empirical results show that this contrastive learning approach does not outperform direct supervised fine-tuning with LoRA, and it is therefore not adopted in the final submission.

3 Experimental Setup and Results

3.1 Experimental Environment

All experiments were conducted on the AutoDL platform. The system for Track A uses the Llama-Factory training framework, while the exploratory experiments use our own implementation. Both systems were implemented in PyTorch 2.7.0 with Python 3.12 running on Ubuntu 22.04. GPU acceleration was provided by a single virtual GPU with 32GB of memory. The compute node was equipped with 80GB of system memory. CUDA version 12.8 was used throughout the experiments.

3.2 Dataset and Tracks

We participate in Track A of the shared task, which focuses on narrative similarity classification. Only the official dataset provided by the task organizers was used for training, validation, and evaluation. The synthesized data is used for training, and the dev data is used for validation. All data are preprocessed and converted into the instruction-based format described in section 2.2. The maximum sequence length is set to 8192 tokens to accommodate long narrative inputs. Data loading and preprocessing are parallelized to improve training efficiency.

Due to a format mismatch, the Track B submission was not officially evaluated. However, our offline experiments still provide useful insights into representation learning for narrative similarity.

3.3 Training Configuration

For Track A, we adopt a supervised fine-tuning strategy based on a pretrained large language model with Low-Rank Adaptation (LoRA). LoRA is applied to all eligible layers, with a rank of 16 and a scaling factor of 32. This configuration allows efficient adaptation while maintaining training stability.

Input instances are formatted as structured text pairs, including the original narrative, candidate narratives, and a binary similarity label. To mitigate overfitting, we use multiple prompt templates with identical task semantics but different surface forms, which are randomly sampled during training. Data cleaning is applied to remove malformed or duplicated samples from the official dataset.

The training process uses a cosine learning rate schedule with a warm-up ratio of 0.1. Models are trained for five epochs with a learning rate of 1×10^{-4}. Due to memory constraints, the per-device batch size is set to 1, and gradient accumulation is employed to achieve an effective batch size of 8.

The model is optimized using AdamW with a fixed learning rate. All experiments are trained on a single GPU. Unless otherwise specified, default hyperparameters provided by the training framework are used.

3.4 Results

Our primary submission for Track A is built upon the supervised LoRA fine-tuning framework detailed in the previous sections. On the official evaluation set, our system achieved an accuracy of 0.70, representing an improvement over the base model (Qwen3-4B-Instruct), which scored 0.57 in local evaluations without task-specific adaptation. This performance gap underscores the efficacy of our multi-prompt strategy and parameter-efficient tuning in capturing narrative-level semantic similarity. However, the current results also suggest that relying exclusively on synthetic data may constrain the model's robustness and generalization across diverse real-world narrative patterns. To further narrow this gap, future research could explore adaptive prompting or self-refinement strategies during inference, enabling the model to dynamically adjust its reasoning logic based on specific input characteristics. Additionally, (Gunasekar et al., 2023; Wang et al., 2022) shows that potentially improvement about extending the training pipeline to incorporate large-scale or weakly supervised narrative data remains a promising direction for enhancing model performance while maintaining task fairness. (Mostafazadeh et al., 2016; Reagan et al., 2016; Valls-Vargas et al., 2014) also provides some useful resources to further development.

Regarding the contrastive learning approach explored, while it facilitates the learning of dense narrative embeddings, its zero-shot transfer performance to Track A (0.56) was markedly inferior

Model	Accuracy
Random	0.51
Contrastive learning approach	0.56
w/o fine tune	0.57
w/o multi-prompt	0.67
Ours	0.70

Table 1: Accuracy comparison between a random baseline, exploratory approach, ablations, and our LoRA fine-tuned system on Track A.

to direct supervised adaptation. This disparity indicates that a simple mean pooling strategy over token representations may be insufficient for encoding complex narrative structures, such as the course of action or long-range discourse relations. To improve the representation of the three core similarity components—abstract theme, course of action, and outcomes—future iterations should consider incorporating structure-aware pooling mechanisms or discourse-level modeling. Furthermore, investigating hybrid approaches that combine embedding-based similarity estimation with lightweight, task-specific classifiers could better align representation learning with downstream decision objectives, providing a more robust framework for narrative reasoning tasks.

4 Conclusion

In this paper, we present a system description for our submission to the shared task, focusing on the design choices, implementation details, and empirical observations across multiple tracks. Our primary submission for Track A is based on a large language model fine-tuned with LoRA, combined with data cleaning and a multi-prompt inference strategy. This approach relies solely on the officially provided dataset and achieves a performance of 0.70 on the Track A evaluation, suggesting its practical applicability under the current setting.

In addition to the main submission, we also investigate a similarity-based approach as an auxiliary method and use it to analyze the differences between embedding-based and classification-based strategies. The similarity-based model was trained to learn narrative-level representations using a triplet loss objective. The learned embeddings were further applied to infer similarity judgments for Track A by comparing narrative structure representations. Although this two-stage approach did not outperform the direct LoRA fine-tuning strategy and could not be officially evaluated in Track B due to submission format misalignment, it provides valuable insights into the strengths and limitations of representation-based methods for narrative similarity tasks.

Overall, our system emphasizes practical and reproducible design choices, discusses the trade-offs between discriminative fine-tuning and embedding-based modeling under realistic resource constraints.

5 Limitations

Despite the effectiveness of the proposed system, several limitations remain. First, our Track A approach relies on prompt-based binary classification, which, while robust in practice, is sensitive to prompt design and may not fully capture fine-grained narrative relations beyond the given label space. Although a multi-prompt strategy was adopted to mitigate overfitting, prompt engineering still introduces an additional degree of heuristic design.

Second, the contrastive learning approach explored in Track B adopts a relatively simple embedding construction strategy based on mean pooling over token representations. Such a design is likely insufficient to capture higher-level narrative structures, discourse relations, or long-range dependencies that are crucial for narrative similarity reasoning. Moreover, the embedding-based inference used in the Track A transfer experiment lacks an explicit decision boundary aligned with the task-specific labels, which may partly explain its inferior performance compared to direct fine-tuning.

Finally, all experiments were conducted using only the official dataset and limited computational resources. While this ensures fairness and reproducibility, it may restrict the model's ability to generalize to more diverse narrative patterns.

References

Suriya Gunasekar and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2025 task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA.

Hans Ole Hatzel and Chris Biemann. 2024. Story embeddings — narrative-focused representations of fic-

tional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*.

Yao Lu and 1 others. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Proceedings of ACL*.

Nasrin Mostafazadeh, Nathanael Chambers, He He, Devi Parikh, Dhruv Batra, James Vanderwende, Margaret Mitchell, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

QwenTeam. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Andrew J. Reagan, Lewis Mitchell, Danforth Kile, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310.

Josep Valls-Vargas, Santiago Ontañón, and Jicheng Zhu. 2014. Narrative plot graphs: A structurally-based method for analyzing stories. In *Proceedings of the 7th International Conference on Interactive Digital Storytelling (ICIDS)*.

Bin Wang and 1 others. 2021. Structure-aware pooling for narrative representation learning. In *Proceedings of the Workshop on Narrative Understanding*.

Liang Wang and 1 others. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Yi Wang and Max Kreminski. 2025. Can llms generate good stories? insights and challenges from a narrative planning perspective. In *2025 IEEE Conference on Games (CoG)*, pages 1–8.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Ye Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llama-factory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.09440*.