NCL&HKU-NarrSim at SemEval-2026 Task 4: Aspect-Based Agents and Supervised Contrastive Embeddings for Narrative Similarity

Jianfei Xu[1], **Ting Zhu**[1], **Mingyang Chen**[2] and **Huizhi Liang**[1]
[1] School of Computing, Newcastle University, Newcastle upon Tyne, UK
[2] School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China
{j.xu65[1],t.zhu11[1], huizhi.liang[1]}@newcastle.ac.uk
u3664924[2]@connect.hku.hk

Abstract

SemEval-2026 Task 4 on Narrative Similarity requires models to assess narrative alignment between stories rather than relying on surface lexical similarity. For Track A, we introduce the Aspect-Based Narrative Similarity Agents(ABNS-Agents), a two-stage agent-based framework. It extracts three core narrative aspects aligned with the task definition under a schema constraint, and then performs aspect-aligned similarity adjudication using an LLM decision model. For Track B, Narrative Supervised Contrastive Embeddings(NSConE) is based upon supervised contrastive learning to model narrative similarity. Our experiments show that ABNS-Agents achieves 70.25% accuracy on the test set, while NSConE reaches 68.5% test accuracy, demonstrating competitive performance across both reasoning-based and representation-learning paradigms. The findings highlight the effectiveness of aspect-aligned structured modelling and task-specific supervised contrastive learning for capturing narrative similarity beyond surface semantics. The code is available at https://github.com/jianfeixu95/NCL-SemEval2026-NarrSim.

1 Introduction

Narrative similarity assessment and representation learning aim to capture structural alignment between stories rather than surface semantic relatedness (Agirre et al., 2012). Such similarity requires understanding high-level elements such as thematic coherence, event progression, and outcome consistency. SemEval-2026 Task 4 (Hatzel et al., 2026) defines the challenge through three components: *abstract theme*, *course of action*, and *outcomes*. In this work, we explicitly adopt this tripartite structure as the basis for our aspect-level modelling. The system should decide which candidate, S_A or S_B, is narratively closer to the anchor S(see Figure 1).

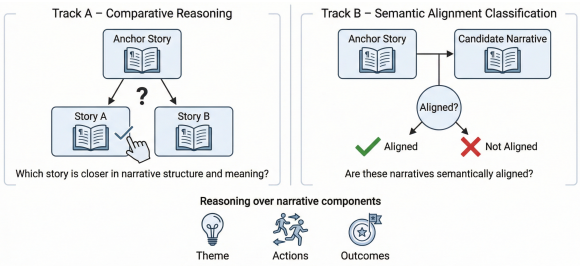

Figure 1: SemEval-2026 Task 4 Description

However, existing similarity methods primarily rely on lexical overlap and sentence-level representations, making it difficult to capture the overall narrative structure and relationships. Therefore, these models may perform well on semantic similarity benchmarks, but they may often fail to distinguish between structurally dissimilar narratives. There is a gap between superficial semantic matching and true narrative understanding. This limitation motivates structure-aware modelling strategies that can explicitly capture various aspects of the narrative and reduce reliance on lexical information.

To address this challenge, we propose two approaches aligned with the shared task tracks: (a) Aspect-Based Narrative Similarity Agents (ABNS-Agents), a structural agent-based framework that converts free-form narratives into structured high-level aspects for interpretable aspect-aware comparisons, and (b) Narrative Supervised Contrastive Embeddings (NSConE), an implicit embedding alignment. NSConE learns similarity-aware representations optimized for triple narrative ranking, providing scalable and robust narrative similarity estimation. These two approaches combine structure-aware modelling with representation learning, offering a unified framework for narrative similarity beyond surface semantic matching.

This study makes three main contributions. (a) We propose ABNS-Agents, a structured aspect-aligned LLM agent framework with schema-constrained extraction and aspect-based scoring.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1451–1461
July 3-4, 2026

(b) We introduce NSConE, a supervised contrastive embedding approach optimized for triple-wise ranking objective with auxiliary MLM training. (c) Results show that our model mitigates reliance on lexical shortcuts, achieving strong narrative-level understanding beyond surface overlap.

2 Related Work

Narrative similarity aims to measure semantic and structural alignment between stories. Early works on narrative understanding relied on symbolic representations (Banarescu et al., 2013; Ammanabrolu et al., 2020), which explicitly modelled event causality and character relations but struggled with linguistic variability. With the emergence of neural methods, representations became dominant (Peters et al., 2018; Kiros et al., 2015), enabling models to capture semantic similarity more robustly. Recent approaches to narrative similarity rely on pretrained language models (Reimers and Gurevych, 2019) and Large Language Models(LLMs). LLMs have shifted the paradigm toward prompt engineering (Liu et al., 2023) and zero-shot reasoning (Brown et al., 2020). However, without structural constraints, LLMs are prone to semantic drift, often overemphasizing superficial topical cues and failing to reliably distinguish between topically related and truly narratively similar content (Lawley et al., 2019; Yao et al., 2023).

BERT and RoBERTa (Devlin et al., 2019; Zhuang et al., 2021) effectively model contextual dependencies within fixed-length inputs and serve as standard backbones for sentence-level representation learning. However, for long-form narratives, representations are often computed from a limited token window, weakening the modelling of global narrative structure. Although long-context variants (Zaheer et al., 2020; Beltagy et al., 2020) mitigate the quadratic memory bottleneck of standard self-attention, extending the context window alone does not necessarily ensure improved narrative-level representations. Contrastive learning underpins modern sentence embeddings, but self-supervised objectives mainly capture surface similarity and struggle with narratives, where meaning depends on causal and temporal structure (Gao et al., 2021). To better model narrative coherence, supervised contrastive learning leverages explicit alignment labels and hard negatives (Khosla et al., 2020), encouraging embeddings to encode structural relationships rather than lexical overlap.

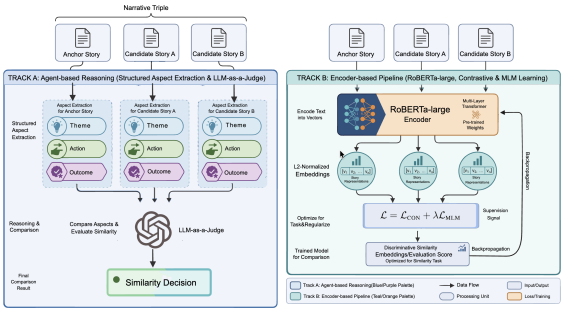

Figure 2: Architecture of our system. Track A performs aspect-based decision-making through an agent-style pipeline, while Track B learns similarity-aware narrative embeddings via supervised contrastive training with an auxiliary masked language modeling objective.

3 Methodology

This study introduces a hybrid system that integrates an LLM-based agent framework with a narrative embedding model. For Track A, we propose the Aspect-Based Narrative Similarity Agents(ABNS-Agents), a two-stage framework that performs structured narrative aspects extraction and aspect-aligned similarity scoring. For Track B, we introduce the Narrative Supervised Contrastive Embeddings(NSConE), which learns discriminative narrative representations via supervised contrastive learning based on SimCSE (Gao et al., 2021). Figure 2 illustrates the architecture of the system.

3.1 Aspect-Based Narrative Similarity Agents(Track A)

Task Definition Given an anchor narrative S and two candidate narratives S_A or S_B, the goal is to identify which candidate narrative is more structurally and semantically aligned with the anchor. The key idea is that narrative similarity cannot be reliably judged from surface-level semantics alone. Instead, narratives should be compared through structured high-level aspects. Our approach follows a two-stage pipeline:(a) aspect extraction and (b) aspect-aligned similarity decision-making.

Aspect Extraction Agent We define an index set $\mathcal{I} = \{\text{anc}, A, B\}$, where $S_{\text{anc}} = S$. Following the official task formulation (Hatzel et al., 2026), which characterizes narrative similarity through three components—*Abstract Theme*, the central topic of the narrative, *Course of Action*, the main events or plot developments, and *Outcomes*, the final resolution or consequence, we extract these

core narrative aspects for any $i \in \mathcal{I}$. For any $i \in \mathcal{I}$, we first extract core narrative aspects:(1) *Abstract Theme* is summarized as a single sentence describing the high-level meaning of the narrative; (2) *Course of Action* is represented as a list of exactly three short event descriptions; (3) *Outcomes* is summarized as a single sentence. This process maps every story into a shared aspect-aligned representation: $A(S_i) = \{a_{\text{theme}}, a_{\text{action}}, a_{\text{outcomes}}\}$

The structured prompting strategy maps narrative comparison into an aspect-aligned space, reducing sensitivity to lexical variation and enabling more robust structure-aware similarity reasoning.

Aspect-Aligned Decision Making Agent After extracting aspects from the anchor narrative S and candidate stories S_A and S_B, we use GPT-4o as an aspect-based decision-making model to compare extracted aspects and determine the closer candidate, leveraging its stronger narrative-structural reasoning to improve decision reliability. Formally, the final decision is obtained by the following:

$$\hat{y} = \arg \max_{c \in \{A,B\}} s\big(A(S_{\text{anc}}), A(S_c)\big) \quad (1)$$

where $s(\cdot, \cdot)$ denotes the LLM-based similarity function in the aspect-aligned space. The full prompt templates used for these agents are provided in Appendix E.1 and E.2.

3.2 Narrative Supervised Contrastive Embeddings(Track B)

Task Definition Track B is a learning problem of narrative representation. Let $x \in \mathcal{X}$ denote a narrative sample, where $\mathcal{X}$ denotes the narrative space. The objective is to learn an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$, which maps a narrative x to a dense representation $\mathbf{h} = f_\theta(x)$. Narrative similarity is measured using cosine similarity:

$$s(x_i, x_j) \triangleq \cos(\mathbf{h}_i, \mathbf{h}_j) = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\| \, \|\mathbf{h}_j\|} \quad (2)$$

The submitted embeddings are evaluated by the organizers using a triple-wise similarity rating. For an annotated triple (x_i, x_i^+, x_i^-), where x_i^+ is judged more similar to the anchor x_i than x_i^-, the encoder model is preferred if $s(x_i, x_i^+) > s(x_i, x_i^-)$.

Narrative Representation with SimCSE We adopt supervised SimCSE (Gao et al., 2021), a contrastive learning framework for structured narrative embeddings. Specifically, we instantiate f_θ with the RoBERTa-large[1] Transformer encoder.

The model final-layer `[CLS]` representation is taken as the fixed-dimensional embedding. We apply ℓ_2 normalization so that cosine similarity reflects angular similarity between embeddings. Our model leverages dropout-induced positive pairs and contrastive optimization to encourage semantically similar texts to be close in the embedding space.

Auxiliary MLM Objective Following prior work, we introduce a masked language modelling(MLM) objective (Devlin et al., 2019) as an auxiliary training signal, applied in our proposed NSConE model during supervised contrastive fine-tuning. The final loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CON}} + \lambda \mathcal{L}_{\text{MLM}} \quad (3)$$

where λ is a weighting hyperparameter. This objective helps stabilize fine-tuning and mitigates catastrophic forgetting of token-level knowledge.

Specifically, given an input narrative $x = (x_1, \ldots, x_n)$, we randomly select a subset of token positions $M \subset \{1, \ldots, n\}$ and replace them with a [MASK] token, $\tilde{x}$, following the standard BERT masking strategy. The MLM loss is:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|M|} \sum_{i \in M} \log P_\theta(x_i \mid \tilde{x}) \quad (4)$$

where $P_\theta(\cdot \mid \tilde{x})$ denotes the token distribution.

Triple-wise Contrastive Fine-tuning Supervised SimCSE[2] is pretrained on sentence pairs, which does not explicitly enforce relative similarity ordering among long-form narratives. However, Track B focuses on narrative-level similarity. To bridge this gap, we fine-tune the SimCSE using the triple-wise supervision (x_i, x_i^+, x_i^-). Within each mini-batch, we optimize the encoder using an InfoNCE-style contrastive objective. Let $\mathbf{h}_i = f_\theta(x_i)$ denote the normalized embedding of narrative x_i, and define cosine similarity as $s_{ij} = \cos(\mathbf{h}_i, \mathbf{h}_j)$. The contrastive loss (Gao et al., 2021) for instance i is:

$$\mathcal{L}_{\text{CON}} = -\log \frac{\exp(s_{i,i^+}/\tau)}{\sum_j \exp(s_{i,j^+}/\tau) + \sum_j \exp(s_{i,j^-}/\tau)} \quad (5)$$

[1] https://huggingface.co/FacebookAI/roberta-large

[2] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large

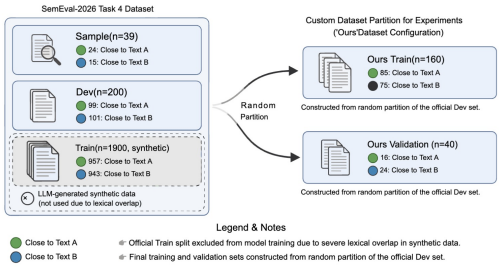

Figure 3: Our Dataset Split and Label Distribution. Our train and validation splits are constructed from the official development split.

where τ is a temperature hyperparameter. This objective encourages higher similarity between anchors and their positive stories than negatives, aligning directly with the triple-wise ranking criterion of Track B.

This task-specific adaptation refines the embedding space to better capture long-form narrative semantics and relative similarity ordering, aligning directly with the triple-wise evaluation. We retain the SimCSE architecture and loss, applying domain-adaptive supervised contrastive learning to specialize representations for narrative similarity.

4 Experiments

4.1 Datasets and Evaluation Metrics

The official dataset[3] is used. However, due to non-negligible lexical overlap in the provided synthetic training data—where models can achieve over 98% accuracy(see Table 8) by exploiting word-level overlap rather than semantic understanding(see Section 4.1 and Appendix B)—we discard the synthetic training data and instead conduct training and validation solely on the official development data and the data split is summarized in Figure 3.

This study retains the official metric, Accuracy. We additionally report results on Semantic Textual Similarity benchmarks(STS12–STS16, STS-B, SICK-R (Reimers and Gurevych, 2019)) to evaluate whether the learned embeddings preserve semantic similarity, which is a prerequisite for higher-level narrative understanding.

Lexical Overlap We conduct a quality analysis and identify a lexical overlap bias in the official synthetic training split: the lexical overlap between the anchor and the closer candidate is substantially higher than the less similar one. This imbalance allows models to exploit surface-level cues, leading to artificially inflated performance and shortcut learning that compromises evaluation validity. To quantify this effect, we compute multiple overlap-based metrics(see Appendix B). The synthetic training split consistently exhibits significantly higher overlap scores across all metrics(see Table 7).

[3]https://narrative-similarity-task.github.io/data/

4.2 Aspect-Based Narrative Similarity Agent

For Track A, we implement a two-stage LLM-based agent pipeline via OpenAI API(temperature = 0, max_tokens = 500, top_p = 1.0). An extraction agent first identifies three aspects(abstract theme, course of action, and outcomes), followed by a decision agent that produces the final judgment. We experiment with different LLM configurations(GPT-4o-mini, GPT-4o, and GPT-5.2). Two decision strategies are explored: direct judgment and weighted aggregation across aspects, where the weights($W_{\text{theme}} = 0.35$, $W_{\text{action}} = 0.45$, $W_{\text{outcomes}} = 0.20$) are determined via grid search.

4.3 Narrative Supervised Contrastive Embeddings

For Track B, we first evaluate embedding models from different families and training paradigms on our dataset to compare performance. The model achieving the best results across the evaluation metrics is selected as the backbone model. Then, we perform supervised contrastive fine-tuning on our training data to obtain the final model.

Backbone Model Selection To identify a suitable backbone encoder, we designed baseline experiments from two complementary perspectives: model family diversity(see Appendix D.1) and training paradigm diversity(see Appendix D.2). This strategy allows us to disentangle architectural effects from training methodology effects when modelling narrative-level similarity.

Training Setting Our final model is built on Sup-SimCSE-RoBERTa-large. We further fine-tune it for narrative similarity using triple narratives formatted as contrastive instances. The best checkpoint is selected based on validation accuracy. Due to substantial lexical overlap in the synthetic training data and the inability to effectively mitigate it, we instead train and validate on the relatively small official development set. To reduce overfitting, we monitor validation performance and apply early

Extract Aspects	Decision	Dataset	Accuracy
GPT-4o-mini	GPT-4o-mini	dev	63.50
GPT-4o	GPT-4o	dev	65.00
GPT-4o-mini	GPT-4o w*	dev	66.00
GPT-5.2	GPT-5.2	dev	67.00
GPT-4o-mini	**GPT-4o**	dev	**68.50**
GPT-4o-mini	**GPT-4o**	test	**70.25**

Table 1: Performance(%) comparison of different LLM configurations for ABNS-Agents. w* denotes the weighted decision strategy.

stopping, selecting the final model from the best-performing checkpoint. Detailed hyperparameter settings and optimization results are provided in Appendix D.3.

5 Results and Discussions

5.1 Aspect-Based Narrative Similarity Agent

LLM-based configurations consistently outperform embedding-based methods when stronger models are used for decision-making. Larger LLM variants achieve the best performance, while smaller models exhibit a noticeable decline, highlighting the importance of model capacity. The weighted decision variant does not improve results compared to direct decision-making, indicating limited benefit from explicit weighting under the current setup. Overall, structured LLM-based reasoning substantially enhances narrative–anchor similarity judgment relative to embedding-based approaches. Detailed results are reported in Table 1.

5.2 Narrative Supervised Contrastive Embeddings

To identify a robust backbone encoder for Track B, we conduct a structured comparison from two complementary perspectives: model family and training paradigm.

Multilingual BERT and RoBERTa consistently demonstrate strong and stable performance across evaluation settings(Tables 10 and 11), with competitive STS-AVG scores indicating robust general semantic representation ability. In contrast, architectures such as Longformer and Seq2Seq models show less consistent gains, suggesting that extended context modelling or architectural variation alone does not ensure improved narrative similarity performance. Across training paradigms, supervised fine-tuning clearly outperforms both pretrained-only and unsupervised variants in narrative–anchor accuracy, highlighting the importance of explicit semantic alignment signals. Among

Paradigm	Model	Acc(Valid)	Acc(Dev)	Acc(Test)	STS-AVG
Pretrain	BERT-base	62.5	54.5	–	18.63
	BERT-large	50.0	53.5	–	18.37
	ModernBERT-base	50.0	56.5	–	21.38
	ModernBERT-large	60.0	56.5	–	19.54
	DeBERTa-v3-base	62.5	51.5	–	26.48
	DeBERTa-v3-large	62.5	52.5	–	12.71
	Longformer-base	42.5	46.5	–	48.56
	Longformer-large	52.5	60.0	–	–
	RoBERTa-base	50.0	51.0	–	56.57
	RoBERTa-large	62.5	61.0	–	26.86
Unsup-finetune	Unsup-SimCSE-BERT	57.5	59.5	–	76.25
	Unsup-SimCSE-RoBERTa-base	45.0	59.5	–	76.57
	Unsup-SimCSE-RoBERTa-large	55.0	59.0	–	78.90
Sup-finetune	StoryEmb	65.0	–	–	–
	MiniLM-L6	52.5	55.0	–	73.33
	LaBSE	67.5	58.5	–	69.44
	FlanT5-base	52.5	54.0	–	16.62
	Sentence-T5-base	55.0	62.0	–	19.98
	Sentence-T5-large	60.0	63.0	–	29.26
	Sup-SimCSE-BERT	50.0	60.5		– 81.57
	Sup-SimCSE-RoBERTa-base	55.0	59.5	–	82.52
	Sup-SimCSE-RoBERTa-large	**67.5**	**62.5**	–	**83.76**
	NSConE(Ours)	**72.5**	**88.0***	**68.5**	**82.29**

Table 2: Performance(%) comparison across training paradigms on Track B. Acc(Valid) is performed on our 20% validation split sampled from the official dev set. Acc(Dev) and Acc(Test) refer to results on the official development and unlabelled test splits, respectively. Since the test data is unlabelled, results for other models are unavailable. STS-AVG denotes the average Spearman correlation across STS12–STS16, STS-B, and SICK-R. * denotes that the model is fine-tuned on 80% of the official development split, and thus the reported score is not directly comparable to other results. Pretrain: no fine-tuning; Sup-finetune: supervised fine-tuning with labels; Unsup-finetune: self-supervised contrastive fine-tuning.

the strongest candidates, LaBSE and Sup-SimCSE-RoBERTa-large achieve top-tier results. However, LaBSE is more sensitive to data scale, whereas Sup-SimCSE-RoBERTa-large remains stable across dev-based settings. Given that Track B training relies on dev splits due to overlap issues in the official synthetic training set, robustness is critical. Therefore, we adopt a Sup-SimCSE-RoBERTa-large model as our final baseline backbone.

Based on the results in Table 2, our model NSConE achieves the strongest overall performance across evaluation settings. Built upon Sup-SimCSE-RoBERTa-large and further fine-tuned on 80% of the official development split, our approach improves Accuracy(Validation) from 67.5% to 72.5%. In addition, it achieves 68.5% on the official test benchmark. These results indicate that domain-adaptive supervised contrastive training effectively specializes the representation space for narrative similarity, yielding consistent gains over the strong Sup-SimCSE-RoBERTa-large baseline.

6 Conclusion

We propose two narrative similarity modelling methods in this paper. For Track A, we intro-

duce ABNS-Agents, a structure-aware LLM agent framework that transforms free-form narratives into interpretable high-level aspects and performs aspect-aligned similarity inference. For Track B, we propose NSConE, a contrastive narrative representation learning method. It combines embedding optimization with a triple ranking criterion through task-specific fine-tuning of SimCSE. Experimental results demonstrate that structure-aware reasoning and ranking-consistent representation learning provide complementary strengths for narrative similarity assessment. ABNS-Agent achieves 70.25% accuracy on the official test set, which improves interpretability and narrative-level reasoning. NSConE enables scalable and robust similarity estimation in the embedding space, with 68.5% test accuracy. All of these findings highlight the importance of modelling narrative structure beyond superficial semantic overlap.

References

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.

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020. Story Realization: Expanding Plot Events into Sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7375–7382.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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.

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.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3294–3302, Cambridge, MA, USA. MIT Press.

Lane Lawley, Gene Louis Kim, and Lenhart Schubert. 2019. Towards natural language story understanding with rich logical schemas. In *Proceedings of the Sixth Workshop on Natural Language and Computer Science*, pages 11–22, Gothenburg, Sweden. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Type	Split	Num	Close to A	Close to B
Official	Sample	39	24	15
	Dev	200	99	101
	Train	1900(LLM-generated)	957	943
Ours	Train	160	85	75
	Validation	40	16	24

Table 3: Dataset statistics. The *ours* split is derived by partitioning the official dev split into training(160 instances) and validation(40 instances) sets.

Field	Split	Mean	Median	Std	P10	P90	Min	Max
anchor_text	sample	120.79	113.00	42.00	75.60	166.00	53.00	259.00
anchor_text	dev	124.59	116.00	43.11	76.00	184.10	48.00	288.00
anchor_text	train	147.56	149.00	31.86	107.00	185.00	0.00	249.00
text_a	sample	121.03	118.00	57.41	55.60	199.00	43.00	261.00
text_a	dev	120.86	119.50	49.29	59.00	184.80	36.00	304.00
text_a	train	157.82	158.00	32.80	118.00	199.10	0.00	254.00
text_b	sample	116.92	103.00	46.88	59.80	187.40	46.00	203.00
text_b	dev	121.44	116.00	48.45	59.00	193.20	36.00	283.00
text_b	train	158.14	158.00	32.67	120.00	197.10	0.00	254.00

Table 4: Word-level statistics for anchor_text, text_a, and text_b across dataset splits.

Field	Split	Mean	Median	Std	P10	P90	Min	Max
anchor_text	sample	5.46	5.00	1.74	3.80	8.00	3.00	10.00
anchor_text	dev	5.70	5.00	1.59	4.00	8.00	3.00	12.00
anchor_text	train	6.75	7.00	1.30	5.00	8.00	0.00	15.00
text_a	sample	5.67	6.00	1.68	3.00	8.00	3.00	8.00
text_a	dev	5.72	6.00	1.65	3.00	8.00	3.00	11.00
text_a	train	6.81	7.00	1.26	5.00	8.00	0.00	15.00
text_b	sample	5.56	6.00	1.61	3.80	8.00	3.00	8.00
text_b	dev	5.66	6.00	1.73	3.00	8.00	3.00	11.00
text_b	train	6.85	7.00	1.29	5.00	8.00	0.00	14.00

Table 5: Sentence-level statistics for anchor_text, text_a, and text_b across dataset splits.

Field	Split	Mean	Median	Std	P10	P90	Min	Max
anchor_text	sample	155.69	153.00	54.99	89.00	219.80	67.00	313.00
anchor_text	dev	159.85	149.00	57.65	98.80	235.50	57.00	384.00
anchor_text	train	177.48	179.00	36.95	131.00	221.10	0.00	334.00
text_a	sample	153.59	148.00	73.19	68.40	265.00	50.00	330.00
text_a	dev	153.74	149.00	63.38	81.90	238.00	39.00	389.00
text_a	train	193.48	195.00	38.59	147.00	241.00	0.00	293.00
text_b	sample	150.82	130.00	63.00	80.40	240.80	50.00	272.00
text_b	dev	154.31	150.50	63.68	70.00	249.10	39.00	338.00
text_b	train	194.29	195.00	39.39	147.00	241.10	0.00	370.00

Table 6: RoBERTa token statistics for anchor_text, text_a, and text_b across dataset splits.

A Dataset Description

The SemEval-2026 Task 4 dataset consists of four splits: sample(39 instances), development(200 instances), training(1,900 instances, including 1,898 LLM-generated narratives), and test(400 instances). Label distributions are well-balanced across splits, with nearly equal numbers of instances closer to $text\_a$ and $text\_b$, minimizing potential class bias(see Table 3). Narratives are moderately long and consistent in structure: anchor texts and candidate narratives typically contain 5–7 sentences(see Table 5) and 120–160 words on average(see Table 4), with slightly longer narratives in the training split. Token-level analysis using the RoBERTa-large tokenizer[4] shows that most narratives fall well below common transformer input limits, indicating that truncation is rarely required(see Table 6).

B Lexical Overlap

We analyze lexical overlap between anchor narratives and candidate narratives using multiple metrics(see Table 7), including unigram Jaccard overlap, asymmetric anchor–candidate coverage ratios, n-gram overlap(2-gram and 3-gram), ROUGE-1/ROUGE-2, and POS-filtered overlap. All statistics are computed separately for positive and negative pairs across dataset splits.

A clear contrast emerges between the training split and the evaluation-oriented splits(sample and development). In the sample and development sets, positive and negative narratives exhibit nearly identical overlap distributions across all metrics, with mean and median differences close to zero. This indicates that surface-level lexical similarity provides little discriminative signal for evaluation.

By contrast, the training split shows consistent and substantial differences between positive and negative pairs. Positive narratives demonstrate higher unigram overlap, greater anchor and candidate coverage, and increased n-gram and ROUGE overlap, reflecting stronger alignment in local event expressions and lexical choices. Similar trends are observed for POS-filtered overlap, suggesting closer alignment in content-bearing entities and actions.

Percentile statistics(p10 and p90) further support these findings: while the overlap distributions diverge noticeably for positive and negative pairs in the training data, such differences largely disappear in the evaluation splits. Overall, the dataset provides informative lexical signals during training while effectively minimizing lexical shortcuts during evaluation, encouraging models to rely on deeper narrative-level understanding.

The accuracy of unfine-tuned baseline embedding models across different dataset splits for Track B is summarized in Table 8. Notably, all models achieve exceptionally high performance on the synthetic training split, exceeding 98% accuracy, while performance drops substantially on the sample and dev splits. This discrepancy suggests that models can exploit lexical overlap patterns in the synthetic training data without genuine semantic modelling. In contrast, the more moderate results on the human-curated splits better reflect the true

[4]https://huggingface.co/FacebookAI/roberta-large

Metric		Train			Sample Diff	Dev Diff
		Anchor–Closer Candidate	Anchor–Another Candidate	Diff		
Jaccard Overlap	Mean	31.89	15.03	**16.86**	0.35	0.23
	Std	17.92	6.73	**11.19**	0.37	0.00
Anchor Ratio	Mean	47.12	27.15	**19.97**	0.66	0.29
	Std	17.65	9.93	**7.72**	0.96	0.31
Candidate Ratio	Mean	45.03	24.37	**20.66**	1.02	1.17
	Std	17.71	8.98	**8.73**	0.15	0.19
2-gram Overlap	Mean	17.83	4.91	**12.92**	0.10	0.00
	Std	18.04	5.26	**12.78**	0.10	0.06
3-gram Overlap	Mean	10.80	1.86	**8.94**	0.02	0.00
	Std	16.59	3.66	**12.93**	0.02	0.01
ROUGE-1	Mean	54.29	35.49	**18.79**	0.19	0.08
	Std	14.76	8.85	**5.92**	0.95	0.19
ROUGE-2	Mean	27.30	8.96	**18.34**	0.18	0.04
	Std	20.88	8.40	**12.49**	0.95	0.08
POS-filtered	Mean	25.32	10.11	**15.21**	0.85	0.15
	Std	18.74	6.68	**12.06**	0.46	0.17

Table 7: Lexical overlap statistics across dataset splits. The synthetic training split shows substantially larger Anchor–Closer Candidate vs. Anchor–Less Similar Candidate differences compared to the sample and dev splits.

Embedding Model	Train	Sample	Dev
all-MiniLM-L6-v2[5]	**98.89**	58.97	55.00
Unsup-SimCSE-RoBERTa-base[6]	**99.74**	58.97	59.50
Sup-SimCSE-RoBERTa-base[7]	**99.95**	69.23	59.50

Table 8: Accuracy(%) of unfine-tuned baseline embedding models across different official dataset splits for Track B. The results show that all models achieve substantially higher accuracy on the synthetic training data than on the other splits.

difficulty of narrative similarity judgment.

C Evaluation Metrics

To comprehensively evaluate model performance on narrative–anchor similarity modelling, we adopt two complementary evaluation settings: a task-specific classification metric and a standardized semantic textual similarity benchmark evaluation.

Narrative–Anchor Similarity Accuracy The primary evaluation metric for Track B is Accuracy, which measures the proportion of correctly predicted similarity labels between a narrative text and its corresponding anchor story. For each narrative–anchor pair, the model produces a binary similarity decision, and accuracy is computed as the ratio of correct predictions to the total number of instances. This metric directly reflects the model's ability to perform discrete similarity judgment in alignment with the task formulation. Accuracy is reported under both the 20% dev split setting(held-out evaluation) and the full dev setting for comparative backbone analysis.

Standard STS Evaluation via SentEval In addition to task-specific accuracy, we evaluate all baseline encoders using the SentEval toolkit[8] to assess their general semantic textual similarity capability. Sentence embeddings are extracted from each model and evaluated on the official STS benchmark datasets provided in the SentEval repository, reporting the averaged STS score(STS-AVG). This auxiliary evaluation provides a standardized measure of semantic representation quality and allows us to examine the relationship between general STS performance and narrative-level similarity modelling.

[8]https://github.com/facebookresearch/SentEval

D Experiment Details

To identify a robust backbone encoder for Track B, we conduct a structured comparison across three complementary perspectives: model family, training paradigm, and overall ranking(see Table 9).

D.1 Model family diversity

We selected representative models from multiple architectural families to ensure broad coverage of encoder design choices and contextual modelling capabilities. The evaluated model families include:

- Multilingual BERT family(e.g., LaBSE)
- BERT family
- RoBERTa family
- DeBERTa family
- MiniLM family
- Longformer family
- Seq2Seq architectures(e.g., T5-based models)
- Domain-Specific pretrained models

These families differ in pretraining objectives, architectural innovations(e.g., disentangled attention in DeBERTa), parameter scale, and context modelling mechanisms(e.g., sparse attention in Longformer). Including long-context models(Longformer) is particularly relevant for narrative similarity, as narrative texts often exceed standard sentence length. Seq2Seq-based encoders were included to assess whether text-to-text pretraining objectives provide advantages in modelling higher-level semantic alignment.

By covering both standard sentence encoders and long-document architectures, we aim to evaluate how different structural inductive biases affect narrative-to-anchor similarity modelling.

Type	Paradigm	Model
Multilingual BERT	Sup-finetune	LaBSE
RoBERTa	Sup-finetune	Sup-SimCSE-RoBERTa-large
Domain-Specific	Sup-finetune	StoryEmb
BERT	Pretrain	BERT-base-uncased
DeBERTa	Pretrain	DeBERTa-v3-base
DeBERTa	Pretrain	DeBERTa-v3-large
RoBERTa	Pretrain	RoBERTa-large
BERT	Pretrain	ModernBERT-large
Seq2Seq	Sup-finetune	Sentence-T5-large
BERT	Unsup-finetune	Unsup-SimCSE-Bert-base-uncased
Seq2Seq	Sup-finetune	Sentence-T5-base
RoBERTa	Sup-finetune	Sup-SimCSE-RoBERTa-base
RoBERTa	Unsup-finetune	Unsup-SimCSE-RoBERTa-large
MiniLM	Sup-finetune	all-MiniLM-L6-v2
Seq2Seq	Sup-finetune	FlanT5-base
Longformer	Pretrain	longformer-large-4096
BERT	Pretrain	BERT-large-uncased
BERT	Pretrain	ModernBERT-base
RoBERTa	Pretrain	RoBERTa-base
BERT	Sup-finetune	Sup-SimCSE-Bert-base-uncased
RoBERTa	Unsup-finetune	Unsup-SimCSE-RoBERTa-base
Longformer	Pretrain	longformer-base-4096

Table 9: Model types, training paradigms, and evaluated models.

Model Family	Accuracy(20% Dev)	Accuracy(Dev)	STS-AVG
BERT	55.00	56.83	39.29
DeBERTa	62.50	52.00	19.60
Domain-Specific	65.00	–	–
Longformer	47.50	53.25	48.56
MiniLM	52.50	55.00	73.33
Multilingual BERT	**67.50**	**58.50**	**69.44**
RoBERTa	**55.83**	**58.75**	**67.53**
Seq2Seq	55.83	59.67	21.95

Table 10: Performance comparison across different model families on Track B. Multilingual BERT and RoBERTa demonstrate similar performance levels and consistently outperform other model families.

D.2 Training Paradigm Diversity

Beyond architectural differences, we further compare models under different training paradigms:

- Pretrain: Directly using pretrained encoders without task-specific contrastive supervision.
- Supervised Fine-tuning(Sup-finetune): Models trained with supervised similarity or NLI-style objectives.
- Unsupervised Fine-tuning(Unsup-finetune): Models trained using self-supervised contrastive learning objectives(e.g., dropout-based contrastive learning as in SimCSE).

This design enables us to investigate whether supervised semantic alignment objectives transfer better to narrative similarity compared to unsupervised representation learning, and whether pretrained-only encoders already provide competitive semantic representations.

By orthogonally varying model family and training paradigm, the baseline experiments provide a structured comparison space for selecting a robust backbone encoder for subsequent optimization in Track B.

Training Method	Acc(20% Dev)	Acc(Dev)	STS-AVG
Pretrain	55.50	54.35	27.68
Sup-finetune	**58.33**	**59.38**	**57.06**
Unsup-finetune	52.50	59.33	77.24

Table 11: Performance comparison across different training paradigms on Track B. The supervised fine-tuning paradigm consistently outperforms alternative training approaches.

D.3 Narrative Supervised Contrastive Embeddings Model Training Setting

To better adapt the supervised SimCSE backbone to narrative-level similarity modelling, we systematically tuned key hyperparameters, including the learning rate, effective batch size(via gradient accumulation), maximum sequence length, MLM loss weight, and hard negative weight. These adjustments were designed to balance contrastive alignment strength with contextual representation capacity under limited training data.

Training runs for 10 epochs with a learning rate of 5e-6, a per-device batch size of 2, and gradient accumulation of 8(effective batch size 16), and a maximum sequence length of 128 with CLS pooling. We adopt supervised contrastive learning(temperature=0.05) with hard negatives(weight=0.2) and an auxiliary MLM objective(weight=0.1).

E Prompts

E.1 Aspect Extraction Prompt

Figure 4 is the template for aspect extraction of ABNS-Agents.

E.2 Aspect-Aligned Decision Prompt

Figure 5 is the template for decision-making of ABNS-Agents.

1. System Instruction

```
You are an expert annotator for
narrative similarity.  Extract ONLY
the three aspects:  1) Abstract
Theme:  high-level theme, ignore
names/time/place and writing style.
2) Course of Action:  key events
and their order.  3) Outcomes:
final outcome/end state only.
Return STRICT JSON that matches
the provided schema.
```

2. User Query

```
Narrative: {story_text}
Output STRICT JSON with schema:
{"theme":"...", "action":["..."],
"outcome":"..."}
Rules:  theme=1 sentence;
action=EXACTLY 3 short events;
outcome=1 sentence.
```

Figure 4: Aspect extraction prompt template used in ABNS-Agents with GPT-4o-mini.

1. System Instruction

```
You are an expert annotator for
narrative similarity.  You must
decide whether Story A or Story
B is narratively closer to the
Anchor.  You MUST base your
decision ONLY on these aspects:
Abstract Theme, Course of Action,
Outcomes.  Ignore names/time/place
and writing style.  Return STRICT
JSON only.
```

2. User Query

```
Anchor Aspects: {anchor_aspects}
Story A Aspects: {storyA_aspects}
Story B Aspects: {storyB_aspects}
Output STRICT JSON with schema:
{"closer":"A|B", "aspects":
{"abstract_theme":true/false,
"course_of_action":true/false,
"outcomes":true/false},"why":"short"}
```

Figure 5: Decision-making prompt template(GPT-4o) used in ABNS-Agents for Track A.