Narrative Team at SemEval-2026 Task 4: Two-Stage Contrastive Learning for Narrative Similarity Assessment

Tatiana Khaidukova [1] **Ana Ciobanu** [2] **Daniela Gifu**[3,4] **Diana Trandabăț**[2]
[1]Information Technologies and Programming Faculty, ITMO University,
[2]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi
[3]Institute of Computer Science, Romanian Academy - Iasi Branch
[4]Academy of Romanian Scientists
467904@niuitmo.ru
{ciobanu.m.ana, diana.trandabat}@info.uaic.ro
daniela.gifu@iit.academiaromana-is.ro

Abstract

Narrative similarity requires capturing deeper structural and semantic relations than standard text similarity tasks. For SemEval-2026 Task 4 (Hatzel et al., 2026), we introduce a unified two-stage framework based on a RoBERTa-large encoder. Stage 1 performs contrastive pre-training on synthetic triplets to learn general narrative similarity patterns. Stage 2 fine-tunes the model with a ranking-based objective tailored to Track A. The resulting encoder supports both binary similarity classification (Track A) and narrative embedding generation (Track B) without architectural changes. Our system achieves an accuracy of 0.64 on Track A and 0.69 on Track B, outperforming single-stage baselines and demonstrating that combining synthetic contrastive supervision with task-specific ranking yields stable and reusable narrative representations.

1 Introduction

Narrative similarity assessment poses challenges that extend beyond traditional semantic similarity tasks, as narratives encode layered structures involving events, temporal progressions, character networks, and discourse-level coherence. Prior work shows that narrative understanding depends on the interaction between lexical semantics, discourse structure, and contextualized entity interpretation (Wilner et al., 2021); (Hatzel and Biemann, 2024). These observations align with studies on communication uncertainty (Vlăduțescu et al., 2014) and diachronic semantic variation (Gifu, 2015); (Gifu, 2016), which demonstrate how meaning evolves across contexts and narrative framings.

Entity-centric resources further highlight the importance of tracking characters and relations for maintaining coherence. The Quo Vadis storytelling corpus (Colhon et al., 2015), illustrates how narrative structure emerges from interactions among entities and events, while the CoRoLa construction methodology (Gifu et al., 2019) underscores the role of large, well-curated corpora in capturing semantic and stylistic variation—an aspect directly relevant to narrative similarity modelling.

SemEval-2026 Task 4 operationalizes narrative similarity through two complementary subtasks: Track A, which requires determining which of two candidate narratives is closer to an anchor, and Track B, which evaluates embedding-based narrative similarity. These tasks raise a legitimate question: *How can a model learn representations that are simultaneously sensitive to fine-grained semantic cues and robust to structural variation across narratives?* Existing approaches leverage contrastive learning, event-centric modelling, and story-level embeddings (Reimers and Gurevych, 2019); (Gao et al., 2021); (Hatzel et al., 2023), yet narratives often contain implicit relations, evolving entities, and structural transformations that require generalization beyond surface similarity.

To address these challenges, we propose a unified two-stage contrastive learning framework: Stage 1 learns generalizable narrative similarity patterns from synthetic triplets, while Stage 2 fine-tunes the encoder with a ranking-based objective tailored to Track A. This design reflects the hypothesis that narrative similarity emerges from the interaction between contrastive semantic signals and task-specific relational constraints.

Our code is released for reproducibility.[1]

[1] https://github.com/t-v-khaidukova/NarrativeTeam4/blob/main/narsim4-2026-testing.ipynb

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 146–152

2 Background

Research on narrative understanding has long emphasized that narratives differ fundamentally from other text genres due to their multi-layered structure, where meaning emerges from interactions among events, temporal progression, character dynamics, and discourse-level coherence. Foundational work on narrative schemas demonstrated that event structure and causal connectivity play a central role in how readers interpret and compare stories (Chambers and Jurafsky, 2008); (Chambers and Jurafsky, 2009). Subsequent research expanded this perspective by highlighting the importance of entity-centric modeling, showing that tracking protagonists and their evolving relations is essential for capturing deeper narrative alignment (Bamman et al., 2014). In parallel, advances in representation learning have shown that contrastive objectives can produce robust semantic embeddings capable of capturing fine-grained distinctions across texts (Reimers and Gurevych, 2019); (Gao et al., 2021). However, most contrastive frameworks are optimized for short texts or sentence-level semantics, leaving open the question of how to adapt them to long-form narratives, where similarity depends not only on lexical overlap but also on latent structural correspondences, such as parallel event progressions or analogous character roles.

SemEval-2026 Task 4 formalizes narrative similarity through two complementary subtasks designed to probe different dimensions of narrative representation. Track A focuses on relative similarity judgments, requiring models to determine which of two candidate narratives is closer to an anchor. Track B evaluates embedding-based similarity, encouraging systems to produce continuous representations that reflect graded narrative relatedness. The task builds on insights from prior work on narrative reasoning and semantic similarity, but introduces a more challenging setting involving longer texts, richer event structures, and more subtle semantic distinctions (Wilner et al., 2021); (Hatzel and Biemann, 2024). A key contribution of the organizers is the construction of a large-scale dataset of narrative triplets, combining human-annotated examples with synthetic augmentations designed to capture fine-grained semantic contrasts (Younus and Qureshi, 2025). This dataset enables systematic evaluation of models' ability to generalize across narrative transformations such as paraphrasing, reordering, abstraction, or shifts in narrative focus. By integrating both natural and synthetic variation, the benchmark provides a controlled environment for studying how narrative similarity emerges from the interplay between lexical semantics, discourse structure, and entity-based reasoning.

Taken together, the literature and the design of SemEval-2026 Task 4 suggest that effective narrative similarity modelling requires unifying contrastive semantic signals with structural cues derived from events, entities, and discourse. This conceptual landscape, where narrative meaning is shaped by layered interactions rather than surface features, has been the central inspiration for our approach.

3 Dataset and Methods

This section presents the dataset and methodological framework used in our system. We first describe the structure, composition, and annotation principles of the Narrative Similarity Task dataset, emphasizing the characteristics that make it suitable for modeling fine-grained narrative relations. We then outline the methods implemented in our system, detailing the learning objectives, training strategy, and architectural choices that operationalize narrative similarity within the constraints of the SemEval-2026 evaluation setting.

3.1 Dataset

The SemEval-2026 Task 4 dataset provides a large-scale, carefully curated resource designed to evaluate narrative similarity across diverse story fragments. Each instance is structured as a triplet consisting of an anchor narrative and two candidate narratives whose relative similarity must be assessed. This formulation enables the dataset to probe subtle semantic distinctions arising from variations in event structure, character roles, temporal progression, and narrative focus. The hybrid construction—combining human-annotated examples with synthetically generated narrative variants—supports both semantic fidelity and controlled contrastive learning.

For Track A, each entry follows the structure `{"anchor_text": "...", "text_a": "...", "text_b": "...", "text_a_is_closer": true/false}`, while Track B uses simplified single-story entries of the form `{"text": "..." }`. The synthetic data is offered in

two formats: one mirroring the Track A triple structure (with an added `model_name` field) and another optimized for contrastive learning, `{"anchor_story": "...", "similar_story": "...", "dissimilar_story": "...", "model_name": "..." }`. The synthetic triplets used in Stage 1 are provided by organizers as part of the official dataset and are generated using large language models (LLMs).

Together, these components provide a rich and flexible benchmark for studying narrative similarity, enabling models to learn from both naturalistic variation and systematically constructed semantic contrasts. This combination makes the dataset a rigorous testbed for evaluating models' ability to capture fine-grained narrative similarity grounded in both semantic and structural cues.

3.2 Methods

Our system follows a streamlined, modular workflow designed to capture narrative similarity through a combination of contrastive learning, ranking-based optimization, and transformer-based text encoding. The overall architecture, illustrated in Figure 1, integrates data preprocessing, representation learning, and task-specific scoring into a unified pipeline that supports both evaluation tracks of the task.

We begin with minimal preprocessing to preserve narrative structure, stylistic cues, and discourse markers essential for semantic interpretation. Each narrative is encoded using a pretrained transformer model, producing dense embeddings that serve as the foundation for both subtasks. Synthetic triplets are incorporated into a contrastive learning phase in which anchor, similar, and dissimilar stories are jointly optimized to reflect their relative semantic proximity. This stage encourages the encoder to internalize patterns of narrative coherence, event progression, and entity consistency.

The training objective combines a contrastive loss with a ranking-based component aligned with the Track A evaluation protocol. Given an anchor and two candidate narratives, the model is optimized to assign a higher similarity score to the narrative labeled as closer. This objective is formalized in the following formula, which operationalizes the relative similarity constraint and ensures that the learned embedding space reflects graded narrative relatedness:

$$\text{score} = \cos(a, t_a) - \cos(a, t_b), \qquad (1)$$

where t_a and t_b are the two candidate narratives.

We use `BCEWithLogitsLoss` with labels that indicate whether t_a is narratively closer to the anchor. This stage is trained for two epochs with a learning rate of 6×10^{-6} and a batch size of 4.

We evaluate several transformer-based encoders within the proposed framework, including RoBERTa-based and sentence-transformer models. Among these, `all-roberta-large-v1` consistently yielded the strongest results and was therefore selected as our primary encoder. This model is built on RoBERTa-large and has been fine-tuned for semantic similarity tasks, making it well suited for capturing nuanced narrative relations. We apply mean pooling over non-padding tokens followed by L2 normalization, resulting in a 1024-dimensional embedding representation. The hyperparameters and training configurations used for this encoder are summarized in Table 1, including batch size, learning rate, number of epochs, and contrastive-loss temperature, each selected to balance stability and generalization. We set the maximum sequence length to 128 tokens to balance computational efficiency and model capacity. While narratives in the dataset can exceed this length, preliminary experiments showed that increasing the sequence length significantly raised memory usage.

Parameter	Value
Base Model	all-roberta-large-v1
Max Sequence Length	128
Stage 1 Learning Rate	1×10^{-5}
Stage 2 Learning Rate	6×10^{-6}
Stage 1 Batch Size	2
Stage 2 Batch Size	4
Triplet Margin	0.3
Rank Margin	0.2
Optimizer	AdamW
Weight Decay	0.01
Gradient Clipping	1.0

Table 1: Model hyperparameters

3.3 Implementation Details

All experiments were implemented in Python using the HuggingFace Transformers and Sentence-Transformers libraries. Training was conducted on a single NVIDIA GPU with mixed-precision enabled to reduce memory consumption and accelerate computation. We used the AdamW optimizer with linear warmup and weight decay, following standard fine-tuning practices for transformer-based encoders.

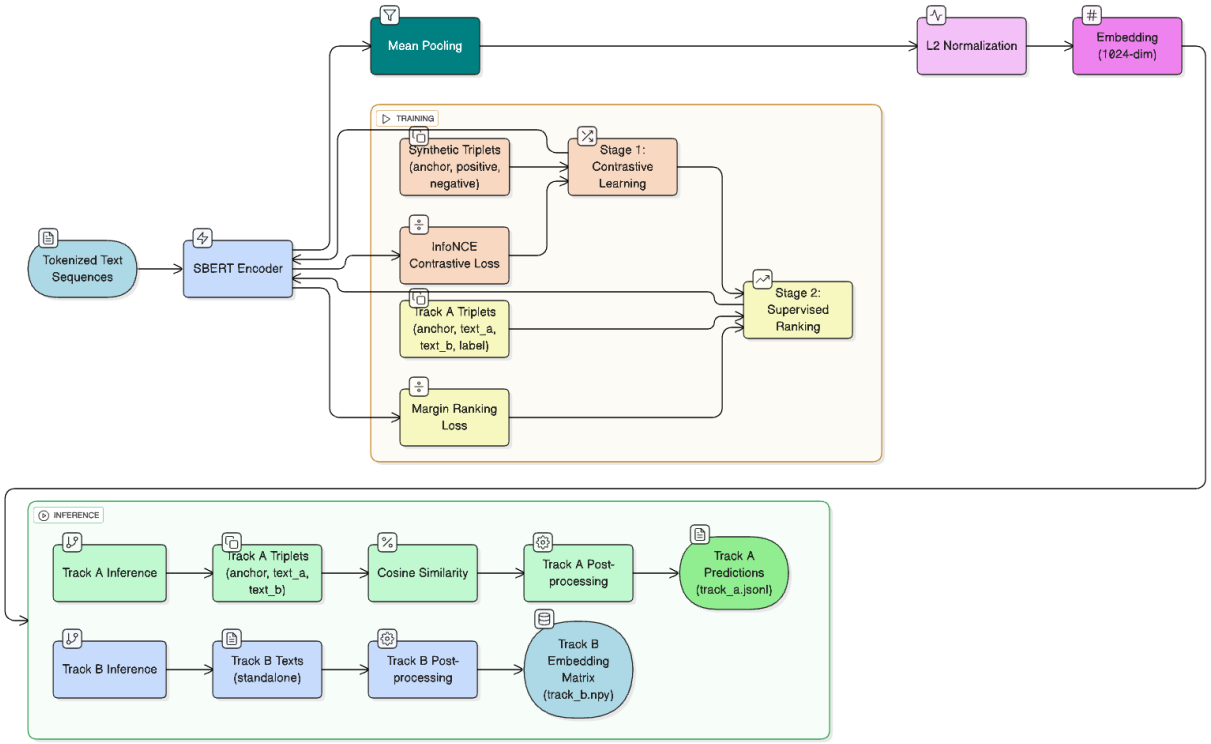

Figure 1: Overview of the proposed two-stage framework. Stage 1 learns narrative similarity via contrastive learning on synthetic triplets, while Stage 2 fine-tunes the encoder using a ranking objective.

During training, batches were dynamically padded to the longest sequence in each batch to minimize unnecessary computation. Gradient clipping was applied to stabilize updates, and early stopping was triggered based on development-set performance. The hyperparameters used in our final configuration, including the learning rate, batch size, number of epochs, and contrastive-loss temperature, are reported in Table 1, together with the settings that yielded the most stable results in both tracks.

For inference, all narratives were encoded using the same preprocessing and pooling strategy described in the previous subsection. Similarity scores were computed via cosine similarity without additional calibration or post-processing. To ensure reproducibility, random seeds were fixed across all libraries and deterministic computation was enabled when supported.

4 Results

The official results of the evaluation for both subtasks are presented in Table 2. Overall, the system achieves modest performance, with an accuracy of 0.64 on Track A and 0.69 on Track B. These scores indicate that, although the model captures certain aspects of narrative similarity, it struggles to generalize to the full semantic and structural variability present in the test set.

On Track A, the accuracy of 0.64 reflects the model's partial ability to identify which of two candidate narratives is closer to the anchor. However, many errors arise in cases where similarity depends on implicit causal relations, unstated motivations, or multi-sentence event chains. In such examples, the model tends to rely on surface lexical overlap rather than deeper narrative coherence, leading to incorrect rankings even when the semantic distinction is clear to human annotators.

On Track B, the accuracy of 0.69 suggests that the embedding space learned during training does not fully encode fine-grained narrative distinctions. The cosine-based similarity scores often compress narratives with different event structures into similar regions of the embedding space. This effect is particularly visible in longer stories or those involving multiple intertwined events, where the encoder fails to maintain consistent entity tracking or temporal alignment.

Several factors contribute to these limitations. First, the training data is dominated by synthetic triplets, which, while useful for contrastive learn-

Model		Stage 1	Accuracy	
Track A	**Track B**	**Use**	**Track A**	**Track B**
roberta-base	roberta-base	No	0.52	0.64
roberta-base	roberta-base	Yes	0.58	0.57
roberta-large	all-mpnet-base-v2	Yes	0.56	0.64
all-roberta-large-v1	all-roberta-large-v1	No	0.61	0.62
all-roberta-large-v1	**all-roberta-large-v1**	**Yes**	**0.64**	**0.69**

Table 2: Model configurations and CodaBench accuracy for Track A and Track B.

ing, exhibit more regular structure and narrower semantic variation than human-written narratives. This mismatch reduces the model's ability to generalize to the stylistic and thematic diversity of the test set. Second, the relatively small development set limits the reliability of hyperparameter tuning, making it difficult to identify configurations that transfer well across narrative types. Finally, transformer-based encoders, even large ones such as `all-roberta-large-v1`, remain sensitive to long-range dependencies and implicit narrative cues, which are central to this task.

Taken together, the results in Table 2 highlight both the strengths and the current limitations of our approach. While the system provides a stable baseline, substantial improvements will require richer training data, more explicit modelling of narrative structure, and mechanisms capable of capturing deeper semantic relations beyond lexical similarity.

4.1 Error Analysis

We grouped the most common errors into four qualitative categories. First, lexical-overlap errors occur when the model selects a candidate that shares many words with the anchor but differs in the underlying event sequence or outcome. Second, implicit-causality errors arise when two narratives describe similar surface actions, but only one preserves the motivation or consequence that makes it closer to the anchor. Third, entity-tracking errors appear in stories with several characters or shifting roles, where the encoder fails to distinguish who performs the central action. Finally, truncation-related errors occur when important information appears after the 128-token limit, especially in narratives where the final sentence changes the interpretation of the preceding events.

A typical lexical-overlap failure involves an anchor and a candidate that share the same setting and several content words, while the truly closer candidate uses different wording but preserves the same event progression. A typical implicit-causality failure involves stories where a character's decision is explained in one candidate but only implied or contradicted in another. These patterns show that the model often captures broad topical similarity but is less reliable when the correct decision depends on causal chains, late narrative developments, or stable character-role assignment.

5 Limitations

Although our system provides a coherent and reproducible baseline for narrative similarity, several limitations constrain its overall performance. A first limitation concerns the quality and distribution of the training data. The synthetic triplets used during contrastive learning exhibit narrower semantic variability and more regular structure than human-written narratives. This mismatch reduces the model's ability to generalize to the stylistic diversity, implicit reasoning, and multi-event complexity present in the official evaluation set.

A second limitation arises from the encoder's difficulty in modelling long-range dependencies. Even with a large RoBERTa-based architecture, the system struggles to maintain consistent entity tracking, temporal alignment, and causal coherence across multi-sentence narratives. These weaknesses are reflected in the relatively low accuracy obtained in Track B and the inconsistent ranking decisions observed in Track A.

The system is also sensitive to implicit narrative cues, such as unstated motivations, background assumptions, or subtle shifts in perspective. Because the model relies primarily on surface-level lexical and semantic similarity, it often fails to capture deeper narrative structure, resulting in embedding compression and reduced discriminative power.

Finally, the limited size of the development set restricts the reliability of hyperparameter tuning. Small variations in learning rate, batch size, or contrastive-loss temperature can lead to substantial fluctuations in performance, making it difficult to identify configurations that generalize robustly across narrative types.

These limitations suggest that future work should explore richer training data, architectures capable of modelling discourse-level structure, and mechanisms that explicitly encode event relations and narrative progression.

6 Conclusion

This study set out to evaluate whether a transformer-based architecture, enhanced through contrastive learning on synthetic triplets, can reliably capture narrative similarity across diverse story types. The results demonstrate that, while the system provides a coherent and reproducible baseline, its performance remains limited, with an accuracy of 0.64 on Track A and 0.69 on Track B. These findings raise a legitimate question: what does this approach actually achieve, and why is it still valuable despite its modest scores?

Our analysis shows that the method succeeds in learning stable, general-purpose narrative embeddings that capture broad semantic relatedness. However, it struggles with deeper aspects of narrative understanding, such as implicit causality, multi-event coherence, and entity-centric reasoning. These limitations highlight a fundamental gap between current transformer-based similarity models and the richer discourse-level representations required for narrative comprehension.

At the same time, the results provide a clear diagnostic signal for future research. The contrastive-learning framework proves effective for structuring the embedding space, but synthetic triplets alone are insufficient for modelling the complexity of human-written narratives. The task therefore calls for richer training data, architectures capable of modelling long-range dependencies, and mechanisms that explicitly encode event structure and narrative progression.

In this sense, the contribution of our system is not its performance level, but its role as a transparent, reproducible, and analytically useful baseline. It clarifies which aspects of narrative similarity current models can capture and, more importantly, which aspects they cannot. By exposing these gaps, the system provides a foundation for developing more sophisticated approaches that move beyond surface similarity toward genuine narrative understanding.

Acknowledgments

This work was carried out partially within the project “Tools for Processing Online Texts Specific to Cultural and Scientific Diplomacy”, funded by the Academy of Romanian Scientists.

References

Mousumi Akter and Shubhra Kanti Karmaker. 2023. FaNS: A facet-based narrative similarity metric. In *Proceedings of a Conference on Narrative or Text Similarity*.

David Bamman, Brendan O’Connor, and Noah A. Smith. 2014. Learning latent personas of film characters. In *Proceedings of ACL 2014*, pages 352–361. 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. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL-IJCNLP 2009*, pages 602–610. Association for Computational Linguistics.

Mihaela Colhon, Daniela Gifu, and Dan Cristea. 2015. The "quo vadis" storytelling. In *Proceedings of the 11th International Conference “Linguistic Resources and Tools for Processing of the Romanian Language”*, pages 3–108. "Alexandru Ioan Cuza" University Publishing House, Iaşi.

Dan Cristea, Daniela Gifu, Mihaela Colhon, Paul Diac, Anca Bibiri, Cătălina Mărănduc, and Liviu-Andrei Scutelnicu. 2016. Quo vadis: A corpus of entities and relations. *Language Resources and Language Engineering*, 408:505–543.

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.

Daniela Gifu. 2015. Contrastive diachronic study on romanian language. In *Proceedings FOI-2015*, pages 296–310. Institute of Mathematics and Computer Science, Academy of Sciences of Moldova.

Daniela Gifu. 2016. *Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language*. Ph.D. thesis, Alexandru Ioan Cuza University of Iași, Iași, Romania.

Daniela Gifu, Alex Moruz, Cecilia Bolea, Anca Bibiri, and Maria Mitrofan. 2019. The methodology of building corola. *Revue Roumaine de Linguistique (Romanian Review of Linguistics)*, pages 241–253.

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

Hans Ole Hatzel and Chris Biemann. 2024. Story embeddings — narrative-focused representations of fictional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

Hans Ole Hatzel, Fynn Petersen-Frey, Tim Fischer, and Chris Biemann. 2023. Dimensions of similarity: Towards interpretable dimension-based text similarity. In *Proceedings of the 26th European Conference on Artificial Intelligence*. IOS Press.

Juri Opitz, Lucas Möller, Andrianos Michail, Sebastian Padó, and Simon Clematide. 2025. Interpretable text embeddings and text similarity explanation: A survey. *Preprint*, arXiv:2502.14862.

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.

Ștefan Vlăduțescu, Florin Smarandache, Daniela Gifu, and Alina Țenescu. 2014. *Topical Communication Uncertainties*. Sitech, Craiova and & Zip Publishing, Ohio/SUA.

Sean Wilner, Daniel Woolridge, and Madeleine Glick. 2021. Narrative embedding: Re-Contextualization through attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arjumand Younus and Muhammad Atif Qureshi. 2025. nlptuducd at semeval-2025 task 10: Narrative classification as a retrieval task through story embeddings. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1742–1746. Association for Computational Linguistics.