Narrative Story Similarity and Narrative Representation Learning

Apetrei Razvan Emanuel, Nechifor Alexandru

January 2026

1 Introduction

This project is based on the **Narrative Similarity Task: Understanding Stories through Similarity Judgments**, a shared task designed to evaluate whether computational models can assess narrative similarity in a way that aligns with human interpretation.

The task defines narrative similarity along three high-level dimensions: abstract theme, course of action, and outcomes. In **Track A**, which is the focus of this work, each instance consists of a reference narrative and two candidate narratives, and the goal is to determine which candidate is more similar to the reference. The task formulation is comparative rather than absolute, reflecting the inherently relative nature of narrative similarity.

In **Track B**, the goal is to produce fixed-dimensional vector representations for individual narratives such that cosine similarity between embeddings aligns with human judgments of narrative similarity. Unlike Track A, which provides explicit comparison pairs, Track B requires embeddings to be generated independently for each story at inference time. The task evaluates whether learned representations capture the underlying narrative structure well enough that simple vector similarity reflects meaningful narrative relatedness.

As the task website provides a detailed motivation and formal definition, this report emphasizes the experimental methodology and empirical findings. We explore multiple pretrained approaches, including embedding-based models, a fine-tuned cross-encoder, and prompt-based large language models, and analyze their performance through quantitative evaluation, data visualization, and qualitative error analysis.

2 Related Work

Recent state-of-the-art approaches to narrative similarity increasingly leverage large language models (LLMs). Due to their extensive pretraining on diverse textual corpora, LLMs are capable of capturing high-level semantic and discourse information and are often applied in a zero-shot or prompt-based man-

ner. In this setting, narrative similarity is framed as a reasoning or comparison task, where the model is prompted to judge which of two narratives is closer to a reference narrative. Such approaches rely on implicit representations learned during pretraining rather than on task-specific modeling.

Alongside LLM-based methods, a separate line of research emphasizes the importance of defining how narratives should be represented and compared, rather than proposing new end-to-end models. These works argue that narrative similarity arises from shared abstract structure, such as common themes, event progressions, and outcomes, and that meaningful comparison requires abstraction beyond surface-level text. This perspective motivates approaches that focus on identifying narrative dimensions or aspects that can be used as a basis for comparison, independently of the specific modeling technique employed.

Neural semantic models provide a practical middle ground between these perspectives. Embedding-based approaches such as Sentence-BERT and SimCSE encode narratives into fixed-length representations that capture semantic similarity efficiently, while cross-encoder architectures jointly encode narrative pairs to model fine-grained interactions. Cross-encoders, in particular, have been shown to be effective for ranking and comparison tasks, as they directly optimize pairwise similarity judgments [2, 3, 4].

In this work, we do not introduce a new narrative representation or similarity definition. Instead, we empirically compare representative models from these different paradigms—embedding-based models, a fine-tuned cross-encoder, and prompt-based LLM approaches—on a comparative narrative similarity task. This allows us to analyze how different modeling assumptions and architectures affect performance and error behavior.

3 Data Visualization

Before applying any modeling approaches, we performed exploratory data analysis to better understand the structure and properties of the dataset. This analysis focused on general characteristics of the narratives rather than task-specific modeling decisions.

We first examined the distribution of narrative lengths in terms of both token count and sentence count. The resulting distributions show substantial variability, with narratives ranging from very short summaries to longer, multi-sentence descriptions. This variability motivates the use of models that are robust to differences in narrative length and level of detail.

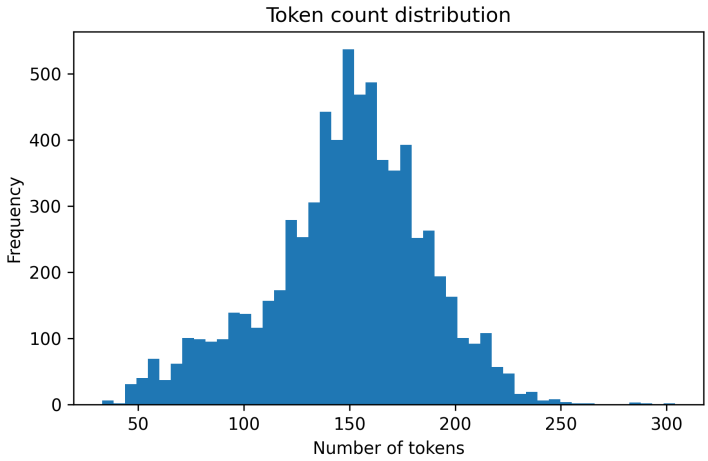

Figure 1: Distribution of token counts per narrative.

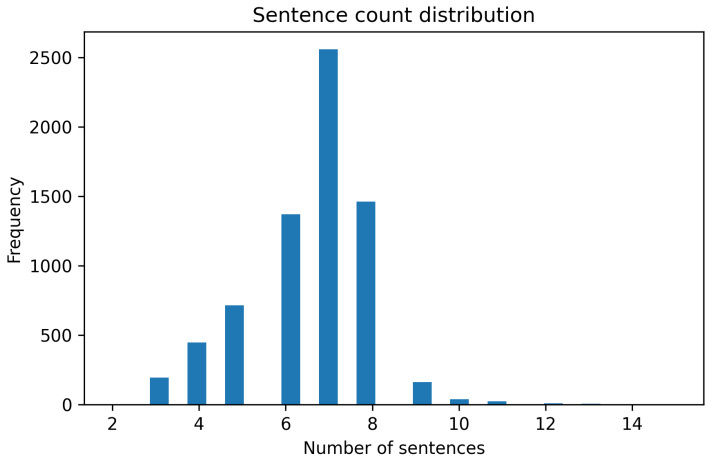

Figure 2: Distribution of sentence counts per narrative.

For Track A, we additionally analyzed the distribution of labels indicating whether candidate story A or story B was judged to be closer to the anchor narrative. The label distribution was found to be approximately balanced, suggesting that simple majority-class baselines are unlikely to perform well.

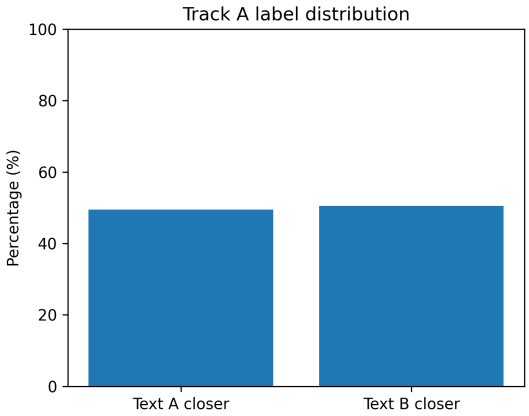

Figure 3: Distribution of labels for Track A (story A closer vs. story B closer).

To clarify the structure of the dataset, we provide simplified examples of the input format for both tracks. Track A consists of narrative triplets with a binary label, while Track B contains single narratives without similarity annotations.

```
{
  "anchor_text": "A story about finding love.",
  "text_a": "Another story about finding love.",
  "text_b": "Unrelated text.",
  "text_a_is_closer": true
}
```

```
{
  "text": "This is the story."
}
```

These observations guided our modeling choices and informed our expectations about task difficulty and model behavior.

4 Our Approaches

We evaluated three classes of pretrained models for Track A of the Narrative Similarity Task: an embedding-based baseline, a fine-tuned cross-encoder, and a prompt-based large language model. These approaches represent different modeling paradigms and allow us to compare efficiency, interpretability, and performance.

4.1 SBERT Baseline

As a baseline, we used a Sentence-BERT (SBERT) model to encode each narrative into a fixed-length vector representation. Narrative similarity was computed

using cosine similarity between embeddings. For each instance, we compared the similarity between the anchor narrative and each of the two candidate narratives, selecting the candidate with the higher similarity score.

SBERT provides an efficient and widely used approach for semantic similarity tasks, as it allows narratives to be encoded independently and compared using a simple similarity metric. This makes it computationally inexpensive and suitable as a strong baseline.

Despite its simplicity, SBERT achieved reasonable performance on the task. However, error analysis revealed that the model often struggled with cases where narratives shared vocabulary but differed in abstract structure or outcomes.

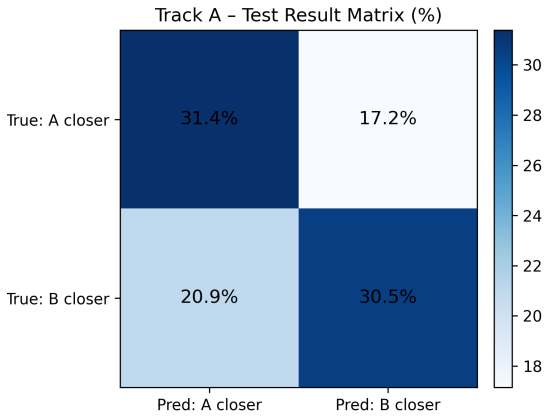

Figure 4: Confusion matrix for the SBERT baseline on Track A.

Qualitative inspection of misclassified examples showed that many errors occurred in cases where both candidate narratives were plausibly similar to the anchor, suggesting that some mistakes may also be challenging for human annotators:

```
Example:

Score difference (A - B): 0.0021
Ground truth: B closer
Prediction : A closer

ANCHOR:
 Angela is studying in the school where her father teaches. She is a
     beautiful girl and despite her boyfriend Tonino's lack of physical
     charms, she insists to be faithful to him. But when she discovers
     that she had been repeatedly betrayed, she decides to take revenge
     allowing herself to Carlo, who has been in love with her for a
     lifetime.
```

```
TEXT A:
 In a rural village, the tyrannical Jonas Lauretz intimidates his
     family, mistress and neighbours. After he disappears one night, it
     is widely believed that his eldest daughter, Silvelie, has
     murdered him. A new investigating judge arrives in the village, he
     falls in love with Silvelie. He becomes torn between his love for
     her and his duty to investigate the potential crime. Eventually it
     emerges that it was not Silvelie who murdered Jonas Lauretz but
     the village innkeeper Bndner. He is forgiven by everyone because
     they all shared his desire to murder him.

TEXT B:
 Germain, a middle-aged literature teacher, bonds with his 16-year-old
     student, Claude Garcia, while tutoring him to improve his writing
     skills. This leads the precocious and disdainful student to be
     increasingly transgressive and antisocial, demonstrating a flair
     for manipulating relationship dynamics and for finding ways to
     satisfy his needs. The student seduces his friend's mother and the
     teacher's wife. He inadvertently causes the teacher to be
     dismissed but they remain in touch due to their mutual passion in
     finding stories that excite them.
```

4.2 Cross-Encoder Model

To model narrative similarity more directly, we employed a cross-encoder architecture. Unlike SBERT, which encodes narratives independently, the cross-encoder jointly encodes a pair of narratives and directly predicts their similarity. This design enables the model to attend to fine-grained interactions between narratives, capturing subtle differences in structure, event progression, and outcomes.

We fine-tuned a pretrained cross-encoder model for six epochs using 80% of the available data, while reserving the remaining 20% for evaluation. The training data combined multiple sources, including the development set, the provided sample data, and additional synthetically generated narrative pairs. The learning objective was formulated as a pairwise comparison task, where the model is trained to assign higher similarity scores to narratives that are closer to a given anchor.

Under this evaluation setup, the cross-encoder achieved an accuracy exceeding 0.9, demonstrating a clear improvement over the embedding-based baseline. The model was particularly effective in cases where subtle narrative distinctions, such as differences in event ordering or story outcomes, determined the correct label.

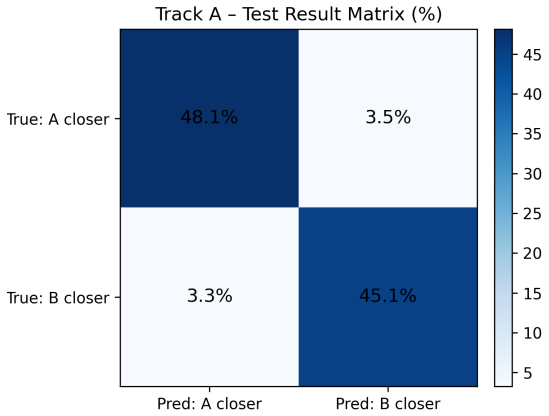

Figure 5: Confusion matrix for the fine-tuned cross-encoder on Track A.

However, when evaluated under the official submission setting, which relies solely on the development data and excludes synthetic examples, the model achieved an accuracy of 0.77. This discrepancy highlights the impact of the training data composition: while synthetic narratives can be beneficial for increasing data volume, they tend to be more homogeneous and easier to classify than real narrative summaries. As a result, performance measured on mixed or synthetic-heavy datasets may overestimate generalization ability.

This observation underscores the importance of carefully selecting and evaluating training data when fine-tuning neural models for narrative understanding tasks. While cross-encoders remain well-suited for comparative narrative similarity, their performance is strongly influenced by the realism and diversity of the data used during training.

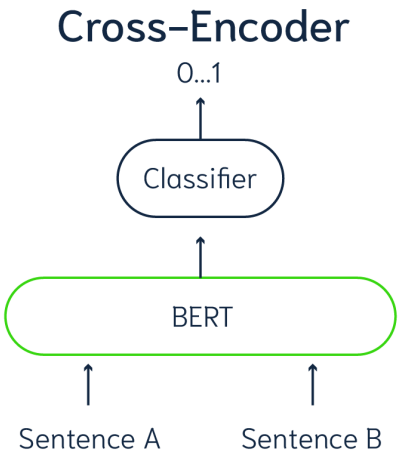

Figure 6: Cross-Encoder architecture illustrating joint encoding of narrative pairs.

4.3 Prompt-Based Phi-3 LLM Approaches

As a final approach, we explored the use of a locally hosted instruction-tuned large language model, namely *Phi-3*, through prompt engineering. The model was evaluated in a zero-shot setting, without any task-specific fine-tuning, in order to assess its ability to perform narrative similarity judgments based solely on natural language instructions.

Rather than directly comparing two candidate narratives in a single prompt, we adopted a structured scoring strategy. In this approach, each candidate narrative was evaluated independently with respect to a given anchor narrative. The model was prompted to assign similarity scores on a scale from 1 to 10 for each of the three task-defined aspects: abstract theme, course of action, and outcomes. These aspect-level scores were then combined using weighted averaging to derive a final similarity decision.

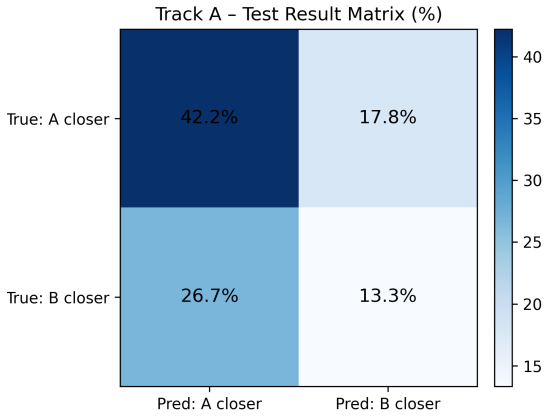

Figure 7: Confusion matrix for the aspect-based scoring prompt approach using Phi-3.

While this structured prompting strategy produced more interpretable outputs than direct comparison prompts, the resulting accuracy remained substantially lower than that of the embedding-based and cross-encoder models. The model frequently generated inconsistent or ambiguous scores, particularly in cases where the narrative differences between candidate stories were subtle. This variability made it difficult to enforce stable and reliable comparative judgments through prompting alone.

These results highlight an important limitation of locally hosted large language models for fine-grained narrative similarity tasks. Although Phi-3 offers strong instruction-following capabilities relative to its size, its limited parameter count constrains its ability to robustly reason over long and complex narrative inputs. Achieving competitive performance with prompt-based approaches would likely require access to significantly larger models, extensive in-context examples, or task-specific fine-tuning.

In practice, such requirements are more naturally met through API-based access to large, well-established language models, which offer substantially greater capacity and more stable reasoning behavior. However, deploying and evaluating these models lies beyond the scope of local experimentation on standard personal computing hardware. Consequently, while prompt-based approaches provide useful qualitative insights, they remain less effective than specialized neural architectures for narrative similarity when constrained to local LLM deployments.

5 Track B: Narrative Representation Learning

While Track A evaluates comparative similarity judgments, Track B focuses on learning fixed-dimensional vector representations for individual narratives. The

goal is to produce embeddings such that cosine similarity between vectors aligns with human judgments of narrative similarity. Crucially, embeddings must be generated independently for each story at inference time, without access to comparison pairs.

5.1 Approach

We fine-tuned a pretrained embedding model using the comparative judgments from Track A as supervision. The key insight is that Track A triplets implicitly define a similarity ordering that can be used to shape the embedding space.

5.1.1 Base Model

We used **BGE-base-en-v1.5** (BAAI General Embedding) as our base model. BGE is a 110M parameter transformer trained with RetroMAE pre-training and sophisticated contrastive learning objectives, achieving a good performance on semantic similarity benchmarks. The model produces 768-dimensional embeddings.

5.1.2 Training Data Transformation

Track A triplets of the form (anchor, text_a, text_b, text_a_is_closer) were transformed into training samples:

- If `text_a_is_closer = true`: create triplet (anchor, positive=text_a, hard_negative=text_b)
- Otherwise: create triplet (anchor, positive=text_b, hard_negative=text_a)

This transformation yields approximately 2,136 training triplets, including synthetic data augmentation.

5.1.3 Loss Function: Multiple Negatives Ranking Loss

We employed **Multiple Negatives Ranking Loss (MNRL)**, which is particularly effective for contrastive representation learning. Given a batch of N triplets, MNRL constructs a similarity matrix comparing each anchor to all positives and hard negatives in the batch.

For anchor a_i with positive p_i, the loss is:

$$\mathcal{L}_i = -\log \frac{\exp(sim(a_i, p_i)/\tau)}{\sum_{j=1}^{N} \exp(sim(a_i, p_j)/\tau) + \sum_{j=1}^{N} \exp(sim(a_i, h_j)/\tau)} \tag{1}$$

where $sim(a, b) = \cos(f(a), f(b))$ is cosine similarity, $f(\cdot)$ is the encoder, h_j denotes hard negatives, and τ is a temperature parameter.

MNRL provides two sources of negative examples:

- **In-batch negatives**: All positives and hard negatives from other samples in the batch serve as additional negatives, providing $2(N-1)$ extra negative examples per sample.
- **Hard negatives**: The explicit negative from Track A forces the model to learn fine-grained distinctions between narratively similar but distinct stories.

5.1.4 Training Configuration

Parameter	Value
Base model	BAAI/bge-base-en-v1.5
Embedding dimension	768
Batch size	6
Epochs	10
Learning rate	2×10^{-5}
Warmup steps	10% of total steps
Precision	Mixed (FP16)
Optimizer	AdamW

Table 1: Training hyperparameters for Track B embedding model.

5.2 Inference

At inference time, each Track B narrative is encoded independently:

```
embedding = model.encode(story_text) # 768-dim vector
```

Narrative similarity between any two stories is computed as the cosine similarity between their embeddings.

5.3 Results

Our fine-tuned model achieved a score of **0.64** on the official Track B evaluation, compared to 0.50 for random chance. This demonstrates that comparative similarity judgments from Track A transfer effectively to the embedding space.

5.4 Embedding Visualization

To qualitatively assess the learned representations, we applied UMAP dimensionality reduction to project the 768-dimensional embeddings into two dimensions. The resulting visualization shows a continuous distribution rather than discrete clusters, suggesting the model captures a spectrum of narrative similarity rather than categorical distinctions.

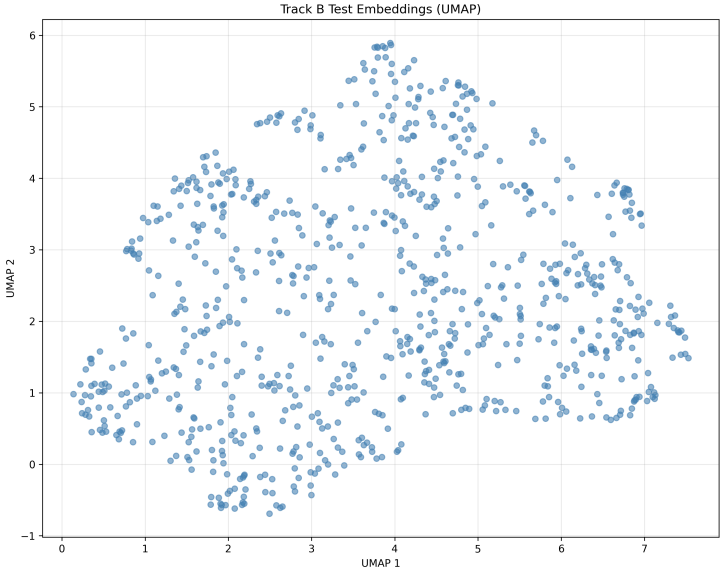

Figure 8: UMAP projection of Track B test embeddings. Points represent individual narratives; proximity indicates learned similarity.

5.5 Discussion

The Track B approach demonstrates that pairwise comparative judgments can effectively supervise representation learning. The use of hard negatives from Track A was particularly important, as random in-batch negatives alone achieved only 0.56 accuracy. By explicitly training on difficult contrasts, the model learns to distinguish narratives that share surface-level features but differ in abstract structure or outcomes.

A limitation of this approach is the relatively small training set. With only ∼2,000 triplets, the model may not fully capture the diversity of narrative similarity patterns. Future work could explore data augmentation strategies or leverage larger pretrained models with stronger zero-shot transfer capabilities.

References

[1] Loizos Michael. *Similarity of Narratives.* In Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2017. Available at: https://www.researchgate.net/publication/316280937_Similarity_of_Narratives

[2] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* In Proceedings of the Conference on Em-

pirical Methods in Natural Language Processing (EMNLP), 2019. Available at: https://arxiv.org/abs/1908.10084

[3] Tianyu Gao, Xingcheng Yao, and Danqi Chen. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021. Available at: https://arxiv.org/abs/2104.08821

[4] Rodrigo Nogueira and Kyunghyun Cho. *Passage Re-ranking with BERT*. arXiv preprint arXiv:1901.04085, 2019. Available at: https://arxiv.org/abs/1901.04085

[5] Jason Wei et al. *Large Language Models Are Zero-Shot Reasoners*. arXiv preprint arXiv:2205.11916, 2022. Available at: https://arxiv.org/abs/2205.11916

[6] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. *C-Pack: Packaged Resources To Advance General Chinese Embedding*. arXiv preprint arXiv:2309.07597, 2023. Available at: https://arxiv.org/abs/2309.07597