UTD-HLTRI at SemEval 2026 Task 4: Reasoning like an Expert for Inferring Narrative Similarity

Rakshitha Rao Ailneni, Maitry Bhavsar, and Sanda M. Harabagiu
Human Language Technology Research Institute
The University of Texas at Dallas
{rxa220074, maitry.bhavsar, sanda}@utdallas.edu

Abstract

Narrative similarity is a challenging problem that requires reasoning over three *aspects* of narratives, including (1) the abstract theme; (2) the course of action and (3) the outcomes of narratives. We present **UTD.HLTRI-SIM.NARRATIVES**, our method developed for SemEval 2026 Task 4 (Narrative Story Similarity), which combines contrastive reasoning prompting with careful selection of few-shot examples to guide a Large Language Model (LLM) toward decisions of narrative comparative similarity. A curriculum learning framework orders examples of narrative triplets presented to the LLM by using a score that quantifies the impact of common narratives aspects with information discerned from several *distractors* of narrative similarity between pairs of narratives [1].

1 Introduction

The computational modeling of narrative similarity can be valuable for tasks such as Question Answering, and information retrieval, enabling tools capable to analyze collections of real or fictional narratives, as pointed out in (Chaturvedi et al., 2018). For example, given a news story, a digital archives analyst might identify similar stories from the past. Additionally, similarities between movie scripts can also be detected by this technology, informing creative writing on new scenarios. Several methods for measuring the similarity between texts are available, and they have been also applied to narrative similarity. For example, Story-Emb (Hatzel and Biemann, 2024a) generates narrative embeddings from story summaries extracted from Wikipedia, to be used in measuring similarity. In contrast, (Chaturvedi et al., 2018) represents narratives in terms of plot events, and resemblances between characters and their social relationships to create a *story kernel* that measures similarity between pairs of stories. In addition (Nguyen et al., 2014) reported on the various dimensions considered by experts and non-experts when judging narrative similarity.

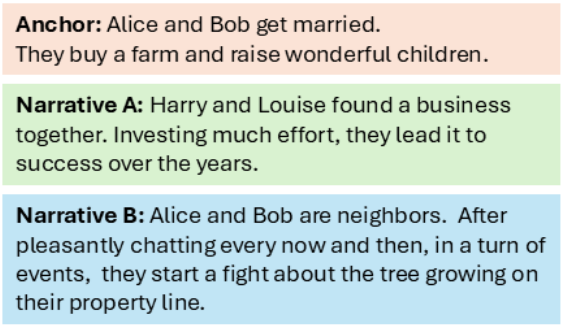

Figure 1: Example of narratives considered for comparative similarity judgment in SemEval 2026 Task 4.

Task 4 of SemEval 2026 (Hatzel et al., 2026b,a) aims to evaluate comparative narrative similarity, not simple similarity between pairs of narratives, as done before in the literature. In Task 4 of SemEval 2026 comparative narrative similarity is cast as a binary selection problem, in which given a triplet of narratives: [*Anchor Narrative, Narrative A, Narrative B*] systems have to decide which of the pair of narratives (*Narrative A, Narrative B*) is most similar to the *Anchor Narrative*. Figure 1 illustrates an example of a narrative triplet, in which *Narrative A* is judged more similar to the *Anchor* narrative than *Narrative B*. To justify the comparative similarity decision, the organizers of the task have instructed their annotators to consider three **narrative aspects**, defined as:

- *A1: Abstract Theme of a narrative:* The defining constellation of problems, central ideas, and core motifs (excluding concrete setting).
- *A2: Course of Action of a narrative:* The sequence of events, actions, conflicts, turning

[1] Code:https://anonymous.4open.science/r/HLTRISimNarrativesSemEvalTask42026/

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3508–3519

points, and the order in which they occur.

- *A3: Outcomes of the narrative:* The results of the plot at the end of the text (e.g., conflict resolution, characters' fates, moral lessons).

Furthermore, the annotators were not provided any guidance on how to weight the narrative aspects when comparing *Narrative A* to the *Anchor Narrative*, or respectively *Narrative B* to the *Anchor Narrative*. A reading of the *Anchor Narrative* and *Narrative A* illustrated in Figure 1 shows that they share the *abstract theme*: i.e. "building a successful life through partnership and sustained joint effort", whereas their course of action is different. But most importantly, the *Anchor* and *Narrative A* have the same outcome: a successful life. In contrast, the *abstract theme* of *Narrative B*: "relations between neighbors" is different than the *abstract theme* of the *Anchor*. In addition, their *outcomes* are also different. But, in their respective *course of action*, both the *Anchor Narrative* and *Narrative B* use the same characters: Alice and Bob, that is **distracting** the narrative comparative similarity assessment.

While the Task 4 of SemEval-2026 had two tracks: **Track A: comparative Narrative Similarity** and **Track B: Narrative Representation Learning**, we chose to participate only in Track A, since our interest was led by curiosity on the capabilities of current Large Language Models (LLMs) to reason with the content of narratives, rather than on their capabilities to participate in the distillation of narrative representations. Our participation in Track A was inspired by the annotation guidelines, aiming to mimic the judgments of experts for comparative similarity of narrative, which explains the title of our paper.

Our system, **UTD.HLTRI-SIM.NARRATIVES**, is considering, in addition to the *narrative aspects*, a set of *narrative similarity distractors*, which contribute the the quantification of the *difficulty* of assessing the comparative similarity for a narrative triplet. Narrative triplets are selected for generating demonstrations used in the Chain-of-Thought (CoT) (Wei et al., 2022) prompting of an LLM. The narrative triplets are presented to the LLM using an order than goes from the least difficult narrative triplet to more difficult ones. In this few-shot learning scenario the LLM prompting in the UTD.HLTRI-SIM.NARRATIVESsystem, it requires *contrastive reasoning*.

2 The NSNRL Dataset

The Narrative Story Similarity and Narrative Representation Learning (NSNRL) dataset is the official benchmark for SemEval 2026 Task 4 (Hatzel et al., 2026b). Sourced from English Wikipedia film summaries from the Tell-Me-Again corpus (Hatzel and Biemann, 2024b), this dataset comprises 1,039 narrative triplets, from which a Development Set of 200 narrative triplets was created, as well as a Test Set of 400 narrative triplets. Texts from the narrative triplets were strictly filtered to contain 4-8 sentences, thus providing a high density of narrative progression. However, we have noticed that the number of words in the narratives varies substantially, as illustrated in Table 1. This indicates than many narratives are longer, in terms of the number of words they use, than those illustrated in Figure 1.

	Development Set	**Test Set**
Shortest Narrative	36 words	35 words
Longest Narrative	304 words	355 words
Average # words per narrative	123.3 words	121.4 words

Table 1: Number of words in narratives.

An important attribute of the NSNRL dataset is that it is highly adversarial. Candidate narratives were generated using LLM prompting to create *hard negatives* namely, narratives exhibiting high surface-level semantic similarity to the anchor narrative, but drastically diverging in their narrative aspects.

3 The Method

The architecture of the UTD.HLTRI-SIM.NARRATIVES system is illustrated in Figure 2. We start by selecting a few examples which were provided by the organizers of the Task 4 of SemEval-2026 as instructions for their annotators. Since these examples are meant to be used in a few-shot learning scenario for prompting an LLM, we also considered using *curriculum learning*, such that examples are presented in order of their difficulty, enabling the LLM to first learn from simpler examples before being presented with more difficult examples. This observation proved helpful in our past experiments (Weinzierl and Harabagiu, 2024) with CoT prompting of LLMs.

The structured prompting that we have used, detailed in Appendix A, defines the narratives aspects,

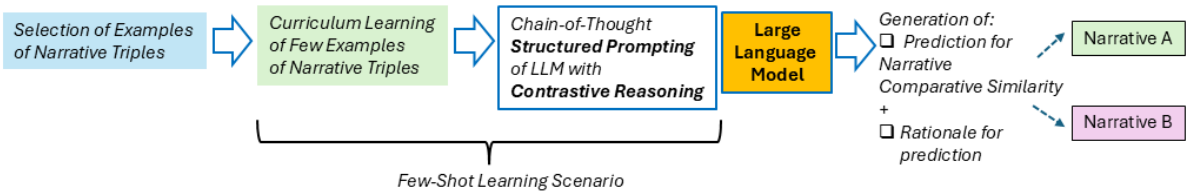

Figure 2: The Architecture of the UTD.HLTRI-SIM.NARRATIVES system.

and instructs the LLM to use *contrastive reasoning* when deciding the *comparative similarity* between (1) the *Narrative A* and the *Anchor Narrative*; and (2) the *Narrative B* and the *Anchor Narrative* of each narrative triplet. The structured prompt also requires the LLM to generate a rationale for its decision.

3.1 Selection of Examples of Narrative Triplets

The annotators for this task were instructed to consider the three possible *narrative aspects* when performing their comparative similarity judgments for the narrative triplets. They were also told that while distinguishing the three aspects of narratives can be challenging, it is important to consider each aspect independently. They were provided with five guideline examples and ten additional training examples. After carefully examining these examples, we have selected nine of them. Our selection followed principles of *complexity* and *diversity*. For example, when considering complexity, we selected examples in which (a) all three narrative aspects were matched between the most similar narrative and the anchor narrative; (b) only two narrative aspects were matched between the most similar narrative and the anchor narrative; and (c) only one narrative aspect was matched between the most similar narrative and the anchor narrative. In terms of diversity, we selected examples in which different narrative aspects were matched between the most similar narrative and the anchor narrative. The examples that we have selected are listed in Appendix B.

3.2 Curriculum Learning

The examples of narrative triplets that were selected were aimed to be presented in a few-shot learning scenario to the LLM. However, we also realized that the LLM will learn better if the examples would be presented in a *curriculum* (Bengio et al., 2009) which is: (1) ranking of examples in terms of difficulty; and (2) transitioning from easy to difficult examples during training. For comparative narrative similarity, example difficulty varies across narrative triplets due to (i) the number of narrative aspects matched against the narrative anchor and (ii) the number of *narrative similarity distractors* that are present in narratives that are compared. We have defined five different narrative similarity distractors:

- *D1: Narrative Beginning:* The anchor narrative and the least similar narrative share the same beginning.
- *D2: Common Setting:* The anchor narrative and the least similar narrative share the same spatial and temporal grounding.
- *D3: Common entities, similar events:* The anchor narrative and the least similar narrative share narrative characters or similar events.
- *D4 Common Emotions:* The anchor narrative and the least similar narrative share the same emotional framing.
- *D5: Common aspects of background:* The anchor narrative and the least similar narrative share commons aspects of their respective narrative backgrounds.

The definition of these narrative similarity distractors informs the quantification of their impact on the comparative similarity decision through a function $D(N_X, N_A)$ which computes the number of narrative similarity distractors observed between a narrative N_A (which could be either the *Narrative A* or the *Narrative B* or a narrative triplet, and N_A, the *Anchor Narrative* or the triplet. Evidently, $D(N_X, N_A) \in \{0, \ldots, 5\}$. For each of the nine selected examples, we have manually quantified the values of $D(N_X, N_A)$ for the *Narrative A* and the *Narrative B* of the example.

Since the narrative aspects of each example had to also be considered, we have defined a function $A(N_X, N_A)$ which computes the number of common narrative aspects observed between a narrative N_X (which could be either the *Narrative A* or the *Narrative B* of an example), and N_A, the *Anchor Narrative* or the example. Evidently,

$A(N_X, N_A) \in \{0, \ldots, 3.\}$ For each of the nine selected examples, we have manually quantified the values of $A(N_X, N_A)$ for the *Narrative A* and the *Narrative B* of the example.

For each narrative triplet example $\mathcal{T} = (N_A, N_{XA}, N_{XB})$, where N_A represents the *Anchor Narrative*, N_{XA} represents *Narrative A*, and N_{XB} represents *Narrative B*, we also had access to X, which indicates which narrative was most similar to N_A, where X could be either XA or XB. We also are aware of Y, which indicates which narrative was least similar to N_A, where Y could be either XA or XB. This enables us to define the difficulty score of each example narrative triplet $\mathcal{T} = (N_A, N_{XA}, N_{XB})$ as:

$$\begin{aligned} D\_Score(\mathcal{T}) \; &= \alpha\big(3 - A(N_X, N_A)\big) \\ &+ \beta\, A(N_Y, N_A) \\ &+ \gamma\, D(N_Y, N_A) \\ &+ \delta\, D(N_X, N_A) \end{aligned}$$

□*Intuitions:* When defining the D_{Score}, we used several intuitions of the difficulty of performing comparative similarity decisions. First, we believe that when fewer common narrative aspects exist between the N_A, the *Anchor Narrative* and N_X, the narrative deemed to be most similar to N_A, it is more difficult to make the comparative similarity decision, therefore we used the term $(3 - A(N_X, N_A)$, which increases by the number of common narrative aspects between N_A and N_Y that are observed. Second, if the least similar narrative N_Y also shares some narrative aspects with N_A, it should also contribute to the value of D_{Score}, hence the usage of the term $A(N_Y, N_A)$. Third, we also account for the number of narrative distractors observed between N_Y and N_A, and N_X and N_A respectively. The coefficients α, β, γ, and δ were selected empirically using the development set. We evaluated several combinations and the values $\alpha = 2$, $\beta = 2$, $\gamma = 1.5$, and $\delta = 1.5$ yielded the best validation accuracy and were therefore used in our final setup.
The difficulty score D_Score produced an ordering of the selected examples, which were presented as few shot examples to the LLM.

3.3 Structured Prompting of the LLM

The prompting of the LLM is first used in the few shot learning scenario, where demonstrations accounting for the selection of the most similar narrative to an *Anchor Narrative* are presented, guiding the contrastive reasoning of the LLM. Examples of such demonstrations are provided in Appendix C. After the LLM is presented with the nine examples and their demonstrations, the LLM is then prompted with examples from the Test Set, being required to *produce rationales* for its decisions. The rationales inform the error analysis of the decisions make by the LLM.

4 Experimental Results

To prepare for our submission in the Task 4 of Semeval-2026, we have considered three baselines, which served in the comparative evaluations of the UTD.HLTRI-SIM.NARRATIVES system on the Development Data and Test Data set. The metric of evaluation was *Accuracy*, which measures the number of times the most similar narrative to the *Anchor Narrative* was predicted out of all the test narrative triplets.

4.1 The Baselines

Narrative embeddings: Our baseline encodes all the narratives from a triplet as embeddings, using `all-MiniLM-L6-v2` (Sentence-Transformers, 2026; Reimers and Gurevych, 2019). The prediction of the most similar narrative is informed by the largest cosine similarity between the embedding of the *Anchor Narrative* and the embeddings of *Narrative A* and *Narrative B* respectively.

Zero-shot prompting of LLM For a simple zero-shot baseline, we follow the organisers' reference implementation[2]. Given a narrative triplet, we prompt the LLM to decide the comparative similarity between *Narrative A* and the *Anchor Narrative* or *Narrative B* and the *Anchor Narrative*. We do not provide any examples to the LLM.

Plot-centered prompting. This baseline considers the importance of discerning the plots of each narrative in order to perform comparative similarity over a narrative triplet. It consists of a sequence of three prompts, detailed in Appendix E

4.2 The results of the UTD.HLTRI-SIM.NARRATIVES system

Prompting different LLMs: We evaluate the UTD.HLTRI-SIM.NARRATIVES system when prompting two LLMs: `GPT-4o` (OpenAI, 2024b,a) and `gemini-2.5-flash-lite` (Google, 2025;

[2] https://github.com/narrative-similarity-task/semeval-2026-task-4-baselines

Table 2: Accuracy results on Development Data and Test Data.

Method	Dev	Test
MiniLM Embedding Similarity	0.55	0.59
Zero-shot Prompting	0.69	0.68
Plot-Centered Baseline	0.68	0.67
UTD.HLTRI-SIM.NARRATIVES (Gemini)	0.71	0.68
UTD.HLTRI-SIM.NARRATIVES (GPT)	0.77	0.74

Table 3: Ablation accuracy results on the Development Data.

Setting	Acc.
UTD.HLTRI-SIM.NARRATIVES	
w/o demonstrations	0.71
w/o contrastive reasoning	0.74
random order of demonstrations	0.75
UTD.HLTRI-SIM.NARRATIVES	0.77

Google Cloud, 2025) on both the development and test sets.

Table 2 lists the results on the Development and the Test Data when prompting each LLM. The Table also shows the results of the baselines. Prompting `GPT-4o` generated the best overall results, both on the Development and on the Test Data. Interestingly the results when prompting both of LLMs decreased by 3% Accuracy from the Development Data to the Test Data, indicating that the Test Data was probably more difficult than the Development Data. However, it was underwhelming to notice that at most 6% Accuracy gain was obtained when (a) performing few-shot learning; and (b) using contrastive reasoning. This indicates that LLMs have developed capabilities are are strong enough for comparative similarity reasoning, such that a few examples and their demonstrations have limited impact on their selection capability. However, there is still a lot of room to improve on LLM reasoning, as in more than 25% is the cases, their reasoning was not correct, which allowed us to perform an error analysis of the test results.

We were also interested to understand the impact on the UTD.HLTRI-SIM.NARRATIVES system of (a) the examples and their demonstrations of narrative comparative similarity; (b) the contrastive reasoning; and (c) the curriculum learning of the examples. For this reason we have conducted ablation studies.

Ablation studies. We conducted three ablation studies on the Development Data set to quantify the contribution of key components of the UTD.HLTRI-SIM.NARRATIVES system. First, we removed the in-context demonstrations of the selected examples. Second, we remove the contrastive reasoning instructions from the LLM prompts. Third, we presented the selected examples, but instead of the order produced by the D_{Score} used in the curriculum, we presented the LLM with the same examples and their demonstrations in a random order. Table 3 reports the accuracy results. Overall, the ablation results indicate that the largest gains in accuracy are granted by the usage of contrastive reasoning, followed by the usage of demonstration of selected examples, ranked by their difficulty.

4.3 Error Analysis

Our inspection of the erroneous decisions of the UTD.HLTRI-SIM.NARRATIVES system and of the rationales that it generated indicates several major types of errors:

Error Type 1: The UTD.HLTRI-SIM.NARRATIVES system infers that two narratives have the same narrative aspects when in fact that do not do so. Appendix D illustrates the analysis of this type of errors.

Error Type 2: The UTD.HLTRI-SIM.NARRATIVES system does not recognize common narrative aspects correctly. missing the inference of common narrative aspects shared by the *Anchor Narrative* and the most similar narrative.

Error Type 3: The UTD.HLTRI-SIM.NARRATIVES system is confused by narrative distractors that make difficult the inference of common narrative aspects, especially when only one common narrative aspect is shared with the *Anchor Narrative*.

5 Discussion

The results of the ablation studies that we have conducted indicate that the contrastive reasoning of the LLM had the greatest impact on performance of the UTD.HLTRI-SIM.NARRATIVES system, but it is obvious that the LLM needed further assistance by being guided to a more nuanced understanding of the narrative distractors. Unfortunately, the prompts that we have used did not include any information about those distractors. In addition, we believe that we should also also defined a set of comparative narrative similarity *enforcers* which

should have been used in the prompting of the LLM system as well.

In addition, although we have conceptualized several narrative similarity distractors, we have not implemented automatic methods of discerning them, preferring to use ad-hoc human evaluations of these distractors only for curriculum generation. Future work should consider automatic means of evaluating both distractors and enforcers of comparative narrative similarity, which should be included in the LLM prompting.

Finally, we believe that further modeling of the three aspects of the narratives should improve the performance on this task.

6 Conclusion

Our participation in the Task 4 of SemEval-2026 through the UTD.HLTRI-SIM.NARRATIVES system enabled us to conclude that a simple system that uses few-shot learning in which examples are carefully selected and ordered by their level of difficulty for comparative similarity decisions has less impact on results than contrastive reasoning. Moreover, prompting an LLM to perform judgments similar those of the annotators helps in improve the results. As expected, LLM reasoning outperforms embedding-based methods for narrative comparative similarity detection. Relying on elements of narrative analysis, such as plot structure, does not help with the task of narrative comparative similarity detection. This may be explained by the fact that in the Task 4 of SemEval-2026 the narratives were in fact very short, produced as summaries, where plot structure was minimal. The complexities of the task of comparative similarity for narrative triplets require further research, as currently, LLMs generate many errors. Our results show that LLMs can closely approximate expert judgments on narrative similarity when guided by structured contrastive reasoning and curriculum-ordered in-context demonstrations. By directing the model to ignore surface cues and focus on abstract theme, course of action, and outcomes, our system achieved promising performance results. In addition, leveraging the organizers' annotation guidelines and carefully ordering demonstrations provides a clear benefit. Finally, our findings are consistent with the broader observation that LLMs capture narrative structure more reliably than embedding-based similarity methods.

Limitations

Dependence on Proprietary LLMs and Reproducibility Our approach relies heavily on commercial, closed-source Large Language Models (e.g., GPT-4o) accessed via paid APIs. This introduces inherent financial costs that scale linearly with the dataset size, potentially limiting the method's deployment in resource-constrained environments. Furthermore, because proprietary models are subject to continuous, unannounced updates by their providers, the long-term reproducibility of our exact experimental results remains a structural limitation.

Similarity distractors Although the narrative triplet examples were presented to the LLM based on a ranking informed by a difficulty score, which involved, in addition to the narrative aspects, the narrative similarity distractors, the LLM prompt did not give any instructions for these distractors to be used when performing contrastive reasoning. The error analysis shows that these distractors played in important role in many of the observed errors. This makes us conclude that the structured prompt has some limitations, as it should have used definitions and instructions about take in account the similarity distractors as well.

Absence of Task-Specific Fine-Tuning Our current methodology operates entirely at inference time via few-shot algorithms. We did not explore supervised fine-tuning (SFT) or parameter-efficient fine-tuning (e.g., LoRA) on open-weight LLMs. Fine-tuning a dedicated model using the dataset's annotated contrastive demonstrations and explicit distractor labels could potentially yield a more robust, cost-effective, and fully reproducible alternative to relying on generalized commercial models.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have I heard this story before? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 673–678, New Orleans, Louisiana. Association for Computational Linguistics.

Google. 2025. Gemini 2.5 flash-lite (gemini api documentation). https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite. Accessed: 2026-02-28.

Google Cloud. 2025. Gemini 2.5 flash-lite (vertex ai model documentation). https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash-lite. Accessed: 2026-02-28.

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

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

Hans Ole Hatzel and Chris Biemann. 2024a. 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 and Chris Biemann. 2024b. Tell me again! a large-scale dataset of multiple summaries for the same story. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2014. Using crowdsourcing to investigate perception of narrative similarity. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*.

OpenAI. 2024a. Gpt-4o model (openai api documentation). https://developers.openai.com/api/docs/models/gpt-4o. Accessed: 2026-02-28.

OpenAI. 2024b. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2026-02-28.

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.

Sentence-Transformers. 2026. sentence-transformers/all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Hugging Face model card, accessed 2026-03-02.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Discovering and articulating frames of communication from social media using chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1617–1631, St. Julian's, Malta. Association for Computational Linguistics.

A Structured Prompting of the LLM

The system prompt used in UTD.HLTRI-SIM.NARRATIVES is a structured prompt designed to guide the LLM in determining which of two narratives (A or B) is more similar to a given Anchor narrative. The prompt first defines narrative aspects that should guide the comparison. It then instructs the model to apply contrastive reasoning by evaluating each aspect for the narrative triplet. Finally, the model is asked to select the narrative that is most similar to the Anchor and provide a final answer accompanied by a brief *rationale* that justifies its decision.

```
You are an expert annotator specializing in
    narrative similarity.
  Your goal is to determine which of two
    candidate stories (Narrative A or
    Narrative B) is most similar to an
    Anchor story based on three core aspects:
  1. Abstract Theme: The defining
    constellation of problems, central
    ideas, and core motifs (excluding
    concrete setting).
  2. Course of Action: The sequence of
    events, actions, conflicts, turning
    points, and the order in which they
    occur.
  3. Outcomes: The results of the plot at the
    end of the text, for example, the
    conflict resolution, the characters'
    fates, moral lessons, etc. It does not
    cover intermediate statuses that change
    later in the story.
  For each aspect, apply contrastive
    reasoning by evaluating the pairs
    (Anchor, Narrative A) and (Anchor,
    Narrative B).
  Follow this reasoning process:
  Aspect-wise comparision: For each aspect
    (Theme, Course of Action, Outcomes),
    compare how closely Narrative A aligns
    with the Anchor and how closely
    Narrative B aligns with the Anchor.
    Focus on structural and conceptual
    similarity rather than surface details
    like names or locations.
  Contrast the candidates: Explicitly
    contrast the similarity of Narrative A
    and Narrative B relative to the Anchor
    for each aspect. Identify which
    narrative better matches the Anchor on
    that dimension and why.
  Final Selection: Decide which candidate
    narrative overall is most similar to the
    Anchor. Base this decision on the
    aspects that most strongly align between
    the narratives.
  Identify contributing aspects: Record the
    aspects that most influenced the final
    judgment.
  Final explanation: Summarize the
    contrastive reasoning and clearly state
    why the selected narrative is more
    similar to the Anchor.
```

B Examples of Narrative Triplets selected for the Few-Shot Learning Scenario

Figure 3 shows examples of narrative triplets selected as demonstrations. These examples were chosen to ensure both complexity and diversity. Among the selected triplets, three have only one narrative aspect matched between the most similar narrative and the anchor narrative, three have two narrative aspects matched, and three have all three narrative aspects matched between the most similar narrative and the anchor narrative.

C Examples of Demonstrations presented to the UTD.HLTRI-SIM.NARRATIVES system

Figure 4 illustrates how the contrastive reasoning process works through concrete demonstrations. Each example presents an *Anchor Narrative* alongside *Narrative A* and *Narrative B*. The annotator then evaluates similarity by systematically comparing the pairs (Anchor, A) and (Anchor, B) across the three defined narrative aspects: abstract theme, course of action, and outcomes.

In Example 1, both Narrative A and Narrative B share the same abstract theme as the Anchor, but Narrative A more closely matches the course of actions and final outcome, leading to its selection. In Example 2, the stories differ in theme and outcome, yet Narrative A is still selected because it mirrors the Anchor's course of action.

D Error Analysis

Figure 5 illustrates a case where the LLM produces an incorrect prediction. The model selects Narrative B as the story most similar to the Anchor, whereas the ground truth indicates that Narrative A is the closer match. The model justifies its choice by claiming that Narrative B aligns with the Anchor in terms of abstract theme, course of action, and outcomes. A closer comparison across these aspects, however, shows that the stronger alignment is actually between the Anchor and Narrative A.

The model interprets Narrative B's focus on duty and protection as thematically similar to the Anchor's conflict. The Anchor centers on martial arts tradition, the legacy of a master, and the pupils' responsibility to confront corruption in order to uphold that legacy. Narrative A closely reflects this theme: the story revolves around martial arts

training, mentorship, and the protagonist's struggle to prove herself in a competitive fighting context. The emphasis on martial arts mastery and personal growth mirrors the Anchor's focus on disciples carrying forward a martial tradition. Narrative B, in contrast, focuses on redemption and protection. Balsa's mission is to save lives and protect a prince while confronting a political threat. Although this includes elements of duty, it does not revolve around martial arts tradition or preserving a master's legacy. Thus, the abstract theme aligns more strongly between the Anchor and Narrative A than between the Anchor and Narrative B.

The model treats Narrative B's journey as similar to the Anchor's progression. In the Anchor, the master dies, the pupils separate, grow through their experiences, and eventually reunite to confront corruption and uphold their master's legacy. Narrative A follows a related progression centered on martial arts development: Jane is discovered after a confrontation, recruited by a mentor, trained rigorously, and ultimately faces rivals in a martial arts competition. This sequence preserves the training and confrontation structure that characterizes the Anchor. Narrative B, in contrast, follows a different pattern in which a wandering warrior saves a prince, becomes his bodyguard, and undertakes a journey to protect him while uncovering a political threat. This escort-and-protection narrative diverges from the Anchor's structure of discipleship, growth, and confrontation. As a result, the course of action aligns more closely between the Anchor and Narrative A.

Finally, the model treats the presence of a political threat in Narrative B as evidence of a similar ending. In the Anchor, the pupils reunite to confront corrupt authority and defend their master's legacy, emphasizing martial tradition and collective resistance. Narrative A concludes with Jane confronting her rivals after extensive martial arts training, reflecting personal mastery and conflict resolution within a martial context. Narrative B instead resolves around protecting the prince and addressing a political or mythical threat to the kingdom. This outcome centers on safeguarding a royal figure rather than defending a martial tradition. Therefore, the outcome aligns more closely between the Anchor and Narrative A.

Overall, the incorrect prediction occurs because the model overemphasizes Narrative B's elements of duty and political danger while overlooking stronger structural and thematic similarities between the Anchor and Narrative A. Across abstract theme, course of action, and outcomes, Narrative A more closely matches the Anchor.

E Plot-Informed Baseline

This baseline decomposes the assessment of narrative comparative similarity into three sequential sub-tasks. For the first two sub-tasks the LLM is prompted to generate a transformation of the narrative, while the third sub-task prompt the LLM do decide on the narrative comparative similarity.

E.1 Sub-task 1: Simplify the narrative

```
You are a narrative normalizer for a
    narrative-similarity task.

Task:
Rewrite each input story into a simple,
    neutral, third-person narrative using a
    shared style.
Preserve the story's plot-critical meaning,
    but reduce superficial distractors by:
- Replacing proper names with stable generic
    roles (e.g., Person1, Person2, Friend,
    Thief).
- Generalizing specific time/place details to
    generic terms unless causally necessary.
- Removing sentiment-heavy phrasing; keep
    plot-relevant facts only.
- Removing background lore/subplots that do
    not affect the main plot arc.

Hard rules:
- Do NOT invent events. Do NOT remove
    plot-critical events.
- Keep the rewrite short (about 3-8
    sentences).
- Keep chronological order of events.
- Output valid JSON ONLY with keys:
    anchor_text, text_a, text_b (each a
    string).
```

E.2 Sub-task 2: Identify plot chronologies

```
You are a story segmenter for a
    narrative-similarity task.

Input:
A JSON object with keys: "anchor_text",
    "text_a", "text_b". Each value is a
    normalized story string.

Task:
For EACH story, split the story into 4-8
    chronological plot beats (segments) for
    the MAIN plot arc.
Each beat must be in story-world
    chronological order.

Beat labels (choose exactly one per beat; use
    role in the overall arc):
- PREMISE: starting situation / context
    needed for the arc
```

```
- GOAL: what the main character wants/needs
    (explicit or strongly implied)
- ACTION: a deliberate step taken
    toward/within the arc
- CONFLICT: an obstacle, pressure, or problem
    that forces a response
- TURNING_POINT: a key pivot that changes
    direction or raises stakes
- OUTCOME: the final result/resolution at the
    end of the story (EXACTLY ONE per story)

For each beat, output an object with fields:
- label: one of {PREMISE, GOAL, ACTION,
    CONFLICT, TURNING_POINT, OUTCOME}
- text: 1-2 simple sentences describing this
    beat (faithful to the input story)
- function: a 3-7 word plot-function phrase
    describing this beat's role
- gist: a compact 5-12 word event summary
    derived from the beat text
- story_effect: a short phrase describing why
    this beat matters for progression

Hard rules:
- Do NOT invent facts beyond the normalized
    story text.
- Preserve chronological order.
- Ensure EXACTLY ONE OUTCOME beat per story.
- Output valid JSON ONLY with keys:
    anchor_text, text_a, text_b.
  Each value is a list of beat objects as
    specified above. No extra text.
```

E.3 Sub-task 3: Decide comparative similarity

```
You are an expert annotator specializing in
    narrative similarity.

Goal:
Determine which of two candidate stories
    (Story A or Story B) is most similar to
    an Anchor story
based on three core aspects:

1) Abstract Theme: The defining constellation
    of problems, central ideas, and core
    motifs
   (excluding concrete setting).
2) Course of Action: The sequence of events,
    actions, conflicts, turning points, and
    the order
   in which they occur.
3) Outcomes: The results of the plot at the
    end of the text (e.g., conflict
    resolution, characters'
   fates, moral lessons). Outcomes do NOT
    include intermediate statuses that later
    change.

Ignore:
- Writing style
- Concrete setting/time period (unless it is
    required by the events)
- Character and location names
- Text length
- Level of detail in which events are
    described

Input:
Stage 2 JSON. Each story is a list of
    chronological beats with labels, gist,
    function, story_effect.

Instructions:
- Use ONLY the Stage 2 content; do not add
    new story facts.
- Compare Anchor vs Story A and Anchor vs
    Story B for EACH aspect independently:
  - For Abstract Theme: infer a concise theme
    statement for each story from the beats
    (exclude setting),
    then decide which candidate matches the
    anchor's theme more closely.
  - For Course of Action: compare the event
    sequence and key pivots/obstacles,
    including their order.
  - For Outcomes: compare the OUTCOME beats
    (final result).
- Then select the overall closer story.

Output JSON ONLY (no extra text) with this
    schema:
{
  "result": {
    "contra_reason": {
      "theme": "<contrastive theme
    reasoning>",
      "course": "<contrastive
    course-of-action reasoning>",
      "outcomes": "<contrastive outcome
    reasoning>"
    },
    "selected_story": "A" or "B",
    "contributing_aspects":
    ["A_theme","C_action","Outcomes"],
    "final_explanation": "1-3 concise
    sentences summarizing why the selected
    story is closer."
  }
}

Contributing aspects:
Include an aspect in contributing_aspects
    only if it meaningfully supports the
    selection.
If an aspect is roughly a tie, omit it.
```

Example 1
Achor: Anna loses her purse. She retraces her steps but cannot find it. Dan finds it and helpfully returns it to her.
Narrative A: Brian lost his backpack. He was terrified because there were important documents in it. After an hour of intense search he finally found it.
Narrative B: Alex loses his engagement ring while swimming. He freaks out, and after hours of diving for it, he still cannot find it.
Example 2
Achor: Anna loses her purse. She retraces her steps but cannot find it. Dan finds it and helpfully returns it to her.
Narrative A:Brian lost his backpack. He was terrified because there were important documents in it. After an hour of intense search he finally found it.
Narrative B: Alex lost his engagement ring while swimming. After hours of looking, he still can not find it. Karen finds the ring while magnet fishing and, based on the engraved name, manages to return it.
Example 3
Achor: In the trenches of World War I, Greg is hit by a grenade splinter. He is in tremendous pain, but his comrades manage to evacuate him from the position. After spending weeks fighting the infection in his leg, he succumbs to his injuries.
Narrative A: Jill was driving home when another car suddenly crashed into hers. After receiving medical attention, she recovered within just days and now advocates for traffic safety.
Narrative B: As Major Miller gives the command to charge, he is not sure if his men can manage it. In a heroic effort, they capture the next position. Only one day later, though, it is again lost to the enemy.
Example 4
Achor: Maven is a magician; ever since finishing his apprenticeship, he has worked on developing novel magic in his tower. As the years go on, he gets fewer and fewer visitors, focusing his life only on his work. After he dies alone in his tower, nobody finds his body for many decades.
Narrative A: The expedition is not going as planned. Some party members have abandoned the mission and tried to return home. One day, the other three remaining members decide to head home, and Ellie realizes she is the last one in the expedition party. Working through snow and ice, she underestimates the storm and freezes to death; nobody ever finds her.
Narrative B: Three friends go on a fishing trip. They catch nothing. They still had a great time.
Example 5
Achor: Andrew goes to the shop to buy food and drinks. He then heads home and prepares everything for his family's arrival. As aunts and uncles arrive, he can impress them with homemade cookies and fancy drinks.
Narrative A: Zoie buys ammunition and guns; she will need them. Back home, she prepares well, setting up traps and protected firing positions. When the Zombies rush her doors, she is prepared and can deal out destruction. Nonetheless, she cannot win against the unending hoards of undead.
Narrative B: Erica is great at building paper planes. One day, to her surprise, she attends a competition, and despite little preparation, she wins!
Example 6
Achor: Alice and Bob get married. They buy a farm and raise wonderful children.
Narrative A: Harry and Louise found a business together. Investing much effort, they lead it to success over the years.
Narrative B: Alice and Bob are neighbors. After pleasantly chatting every now and then, in a turn of events, they start a fight about the tree growing on their property line.
Example 7
Achor: After an accident, Neo, a lonely man in his 40s, loses his eyesight. He learns to handle the new challenges the world now has for him. He makes a new friend, Anna, who is very willing to help him out, finding connections that he never thought possible before.
Narrative A: Adam has a large circle of friends. One day, he loses his hearing in an accident. He struggles with the new challenges in communication, alienating his friends, and finally becomes very lonely.
Narrative B: Brian is a London-based artist who creates street art. He leads a lonely life. One day, he encounters a fellow artist who is painting what looks to be an identical picture to the one he is working on. The two make a deep connection over their shared love for art and remain friends to this day.
Example 8
Achor: In 1836 in southern California near Santa Barbara shortly after California became part of the United States, American settlers and the U.S. government discriminate against the Mexican landowners and frequently take their land by force or legal skullduggery. Wealthy Latino ranchers whose land and wealth are at risk decide to misdirect a U.S. government ship carrying gold so that it will be wrecked and plundered. To prevent themselves from being caught, they plan to massacre the local Chumash Indians. The hero is the now-estranged adoptive son Finley (Tom Laughlin), a master swordsman and gunfighter, who tries to prevent this while still saving his family.
Narrative A: Joaquin Murieta goes to California with the hope of finding gold there. On the way, he meets a beautiful girl whom he marries, but their happiness was short-lived. Suddenly his wife dies.
Narrative B: Timothy Benson, a young intellectual and greenhorn inherits a gold mine. Rodrigo Rodriguez, a bandido leader will stop at nothing to get Timothy to deed him the mine. It is up to Timothy's foreman and his friends to teach Timothy to be a tough hombre and with their help, stop the bandit, Rodrigo Rodriguez.
Example 9
Achor: Two travelers, Eli and Stan, in medieval Europe are stopped by a roadblock. A group of robbers descends upon them. The two defend themselves. The robbers capture Stan. Eli is forced to make a decision: flee the scene or give up and stay with Stan. He surrenders, and the robbers take their belongings but let the two live.
Narrative A: A small group of travelers is heading west by train in 1800s USA. At night, the train stops, and as passengers peer out the windows, they spot hoards of zombies. The group makes their way out, fighting off multiple zombie attacks as they go. It is only as they leave the train that they notice one of their own, Dan, is missing. Deciding not to leave Dan behind, they perform a daring rescue, saving him just before the zombies close in on him and still make it out. In a multi-day effort, the group manages to lose zombies and make it out unharmed.
Narrative B: Two travelers, Eli and Stan, make their way from their home village to the kingdom's capital. The roads are dangerous. The two hear countless tales of robbers, but sticking to the main roads and only traveling by daylight, they make it to their destination and are able to explore the kingdom's center of power for themselves.

Figure 3: Examples of Narrative Triplets selected for few-shot learning in UTD.HLTRI-SIM.NARRATIVES

Example 1
Achor: Anna loses her purse. She retraces her steps but cannot find it. Dan finds it and helpfully returns it to her.
Narrative A: Brian lost his backpack. He was terrified because there were important documents in it. After an hour of intense search he finally found it.
Narrative B: Alex loses his engagement ring while swimming. He freaks out, and after hours of diving for it, he still cannot find it.
Contrastive Reasoning for Theme: Both Story A and Story B share the Anchor's abstract theme of a lost item and the emotional distress associated with it.
Contrastive Reasoning for Course of Action: Story A follows a similar sequence to the Anchor: losing an item, searching for it, and the search concluding with the item being found. Story B follows the Anchor only through the loss and the failed search but lacks the final structural step of discovery.
Contrastive Reasoning for Outcomes: Story A results in a successful recovery of the property, aligning with the Anchor's positive resolution. Story B results in a permanent loss, which is the opposite outcome of the Anchor.
Selected Narrative: A
Contributing Aspects: Abstract theme, Course of Action, Outcomes
Final Explanation: Story A is more similar because all three aspects match with the Anchor whereas Story B only matches in the abstract theme.

Example 2
Achor: Andrew goes to the shop to buy food and drinks. He then heads home and prepares everything for his family's arrival. As aunts and uncles arrive, he can impress them with homemade cookies and fancy drinks.
Narrative A: Zoie buys ammunition and guns; she will need them. Back home, she prepares well, setting up traps and protected firing positions. When the Zombies rush her doors, she is prepared and can deal out destruction. Nonetheless, she cannot win against the unending hoards of undead.
Narrative B: Erica is great at building paper planes. One day, to her surprise, she attends a competition, and despite little preparation, she wins!
Contrastive Reasoning for Theme: The abstract themes of all three stories are unrelated. The Anchor focuses on domestic hospitality, Story A focuses on survivalist combat, and Story B focuses on competitive talent and success.
Contrastive Reasoning for Course of Action: Story A and the Anchor share a nearly identical course of action: a three-step process of purchasing necessary items, returning home to perform preparations, and finally utilizing those preparations in a culminating event. Story B follows a different sequence of spontaneous entry into a competition followed by an immediate win.
Contrastive Reasoning for Outcomes: The outcomes are not similar. The Anchor ends with successful social impression, Story A ends in an inevitable defeat against a horde, and Story B ends in a surprising victory
Selected Narrative: A
Contributing Aspects: Course of Action
Final Explanation: While the abstract theme and outcome of the Anchor and Story A have little in common, they both describe an identical process of purchasing, preparing, and then executing those preparations.

Figure 4: Example of demonstrations used in the UTD.HLTRI-SIM.NARRATIVES system.

Anchor: The events take place in 1932 in Japanese-occupied Manchuria, in which the corrupt leaders of the Japanese army are trying to take over all the Karate dojos /training halls for their own benefit. The master Eiken Shibahara (Yosuke Natsuki), from one of these dojo located on the southernmost Japanese island of Kyushu, dies before passing on the Kuroobi/ black belt to his successor. Three of his pupils: Taikan, Giryu, and Choei, have the task of deciding amongst themselves who deserves it most.\nAfter they bury their master, they are forced to leave the dojo and join the Japanese army. At this point, their journeys lead them on different paths both in life and in the understanding of their master's teachings of martial arts. They are reunited in the end, to battle together against corruption and uphold tradition.
Narrative A: When Jane, a beautiful but troubled American backpacking her way through Hong Kong, successfully fends off three thugs trying to rob her, it draws the attention of Shu, a female fighting champion. Shu recruits and trains Jane to fight in the vicious, all-female underground martial arts tournament known as The Kumite. After months of rigorous training, Jane is ready to face off against her killer rivals, including the apprentice of Shu's nemesis, a Shaolin master. As other nefarious forces emerge from the shadows, Jane's journey through The Kumite turns deadly as she risks everything to become the best female fighter in the world.
Narrative B: Balsa, spear wielder and bodyguard, is a wandering warrior who has vowed to atone for eight deaths in her past by saving an equivalent number of lives. On her journey, she saves Prince Chagum and is tasked with becoming his bodyguard. His own father, the Mikado, has ordered his assassination. The two begin a perilous journey to ensure the survival of the prince. Balsa's complicated past begins to come to light, and they uncover Chagum's mysterious connection to a legendary water spirit with the power to destroy the kingdom.
CR for Theme: Narrative A centers on martial-arts training, mentorship, and individual competition. Narrative B centers on duty, protection, political threat, and a perilous journey to safeguard an important figure. The Anchor combines martial-arts tradition with political corruption and the duty of pupils to defend their master's legacy.
CR for Course of Action: The Anchor's sequence: master dies -> pupils must decide succession -> forced into military service -> separate paths/journeys -> reunite to fight corruption. Narrative A follows unplanned confrontation -> recruitment by a mentor -> training -> tournament showdown (individual-focused). Story B follows: a wandering protector saves/accepts duty for a charge -> embarks on a perilous escort journey -> uncovers a political/legendary conspiracy (group protection against a larger political threat).",
CR for Outcomes: The Anchor ends with the pupils reuniting to resist corrupt authority and uphold tradition. Narrative A ends with the protagonist pursuing victory in a fighting arena (individual achievement) and facing nefarious forces. Narrative B ends with prolonged protection and exposure of a political/legendary threat, aligning with the Anchor's outcome of confronting political danger and defending a larger legacy.
LLM Prediction: Narrative B
Gold Label: Narrative A
Contributing aspects: Abstract theme, Course of action, Outcomes
LLM-generated final explanation: Although Narrative A shares the martial-arts motif and mentorship with the Anchor, Narrative B better matches the Anchor's combination of duty, political threat, and a protective/journey narrative culminating in confronting a larger corrupt or dangerous power. Overall, B aligns more closely across theme, sequence, and outcome.

Figure 5: Example of an incorrect LLM prediction in narrative similarity. CR represents Contrastive Reasoning.