PLlama at SemEval-2026 Task 4: Zero-shot Prompting with Llama-3.2 for Narrative Similarity

Kanishka Jain
Indian Institute of Technology Delhi
kanishka@hss.iitd.ac.in

Abstract

This paper describes our submission to the SemEval-2026 Task 4 on Narrative Story Similarity and Narrative Representation Learning. The shared task focuses on modeling the similarity across narratives on the basis of perceived relatedness between events' causality. The task frames narrative similarity as a binary classification problem in which the models determine which of the two stories is more narratively similar to a given anchor story. Our approach leverages the pre-trained language model Llama-3.2-3B-Instruct with prompt engineering, allowing the system to assess narrative similarity without explicit fine-tuning. On the test data, our system achieved an accuracy of approximately 55% in Track A. While modest, our results establish a baseline for narrative similarity detection in large language models (LLMs) highlighting both their potential and challenges of applying computationally efficient instruction-tuned models to this task. Our analysis highlights the struggle of LLMs in capturing event causality and long range narrative dependencies.

1 Introduction

When humans read a narrative (or story), it is natural for them to relate it to other narratives that they may have heard, read, or sometimes to their own experiences. Human readers generally use different aspects of narratives such as common event(s), characters, resolutions, and much more to make connections between different stories (Chaturvedi et al., 2018). However, it is difficult to say whether mechanical readers like large language models (LLMs) are also able to understand similarities between two narratives and even more so what heuristics they may use to do so.

Narrative similarity is a complex and underexplored task in natural language processing (NLP), involving the identification of narratively similar stories based on different characteristics or features of the stories. Measuring narrative similarity can be useful in tasks like plagiarism detection, story clustering, recommendation systems, and understanding thematic or structural patterns across texts. This paper focuses on our submission to the Narrative Story Similarity task proposed in the SemEval-2026 Task 4 (Hatzel et al., 2026) for English, which aims to determine the capability of systems to compare and identify the narratively similar stories. The task has two tracks. In Track A, systems directly compare two candidate stories against an anchor story and select the one most similar to it whereas in Track B systems instead generate narrative representations as embeddings for the stories.

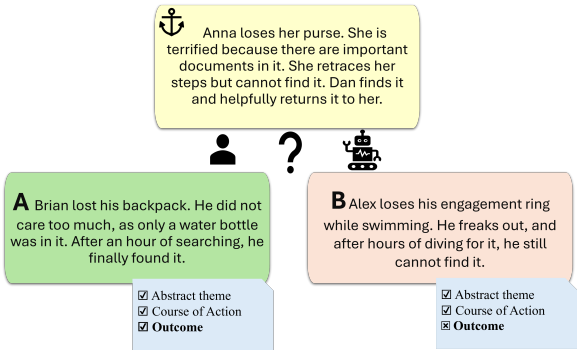

Figure 1: An example of the task. Given the anchor story, the participant needs to identify which story is more similar to it – A or B.

We submit our system for track A. Instead of relying on surface level lexical overlap, the task emphasizes exploiting causal relatedness across narratives, including the overarching theme or goal, sequence of events, and the final outcome(s). For example, in figure 1 story A is considered more similar to the anchor story. Though A, B, and the anchor all have the same theme, in the case of A, the end result matches with the anchor story, whereas this is not the case for story B. This setup

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2673–2678

provides a challenging benchmark for models to capture deep narrative understanding beyond simple semantic or syntactic cues.

To address this challenge, this work employs the Llama-3.2-3B-Instruct model (Dubey et al., 2024) in a prompt engineering framework. The model achieves a modest performance of 55.5% in zero-shot setting, as described in the rest of this paper. We also conduct a detailed analysis to understand how the model makes its judgments. We examine the model's output on various parameters such as lexical overlap, discourse markers, causal language, and prediction agreement with embedding based similarity. Our analysis reveals that only a subset of correct predictions reflects evidence of understanding of the deeper story structure. A substantial portion of model behavior appears to be driven by superficial surface level cues or unstable decision criteria, where the model's reasoning relies on arbitrary narrative features rather than consistent causal logic.

2 Background

2.1 Dataset

The dataset for Track A of this shared task is organized in the form of narrative triplets, where each data point comprises an anchor story and two candidate stories. These stories were extracted from English Wikipedia articles, however only the article summaries were used to construct the dataset, ensuring the narratives are concise yet representative of the full story wherever possible.

The construction of this dataset involved a three-stage process involving summary sourcing, candidate triple sampling, and annotation by humans. From an initial pool of 1,345 narrative triplets, approximately 300 were excluded during quality control to remove inconsistencies, resulting in a final set of 1,039 annotated triplets.

split	no. of items
Sample	39
Dev	200
Test	400

Table 1: Dataset for Track A

The official dataset split for the task is shown in Table 1. In addition, to the manually curated dataset, the organizers provide a synthetically generated set of 1,900 narrative triplets as an alternative to the standard training set. The evaluation of system submissions for this task is based on binary classification accuracy where the model selects which candidate story is narratively similar to the anchor story.

2.2 Related Works

Chaturvedi et al. (2018) studied narrative similarity by identifying correspondences between narratives, focusing on social relationships between characters and high level plot events. Their findings demonstrate that narrative similarity can be captured beyond surface-level textual information. While alignment-based approaches explicitly map character arcs and plot points, they require significant structured annotation, which limits its scalability.

To overcome these limitations, recent works have explored learning dense representations that implicitly capture narrative structure. Hatzel and Biemann (2024) proposed StoryEmb model, which moves beyond standard document embeddings by prioritizing narrative structure ("what happens") over surface level discourse ("how it is told"). They show that specialized embeddings more effectively capture abstract themes, sequences of events, and outcomes. These dimensions are central to benchmark narrative similarity tasks.

More recently, studies have shown that LLMs can achieve strong performance on reasoning and natural language understanding tasks with carefully designed prompts (Qiao et al., 2023).

We situate our study in dialogue with this framework, evaluating whether the general purpose reasoning of Llama-3.2-3b-it can recover these narrative dimensions through prompting, or whether it remains limited by spurious correlations of entities and surface level cues.

3 System Overview

Our sytem uses Llama-3.2 model (Dubey et al., 2024) (specifically *llama-3.2-3b-it*) with zero shot prompting to identify narratively similar stories.

3.1 Llama-3.2

Llama-3.2 is a decoder only transformer model from the Llama 3 family, designed to efficiently capture high level reasoning and abstraction in natural language (Dubey et al., 2024). Similar to other transformer based architectures (Vaswani et al., 2017), Llama leverages self-attention to model long

range dependencies and hierarchical structure in the text. The 3B-Instruct variant is pretrained on a massive corpus of 9 trillion tokens and supports a context window of up to 128k tokens. A defining feature of this model is its use of knowledge distillation from larger LLaMA 3.1 models (8B and 70B), allowing it to retain high-level reasoning, summarization, and abstraction capabilities.

Due to its instruction tuning and autoregressive design, LLaMA 3.2 is capable of performing tasks in zero-shot or few-shot setting. These abilities allow the model to understand and execute tasks without explicit task specific training, making it suitable for this task.

3.2 Prompt Methodology

During the development phase, we evaluated nine zero-shot prompts alongside their corresponding two-shot variants. Empirically, zero-shot prompting consistently yielded better performance than the two-shot settings. Consequently, for the final leaderboard submission we selected the best performing zero-shot prompting setting[1].

Zero-Shot (Leaderboard): The System Prompt was designed to establish a specialized "researcher" persona, priming the model to the task objective and structural analysis. The given prompt emphasizes narrative similarity over general semantic similarity. While the instruction to "read each story twice" serves as a soft reasoning cue, encouraging the model to process the full context and to mitigate any positional bias that may arise in the model's output.

Prompt: You are a researcher. In this study, you are tasked with identifying similar stories. You will be presented with three stories: a base, and two choices, a and b. You are to determine which of the candidate stories, a and b, is the most similar to the base. Specifically, you will consider the stories' narrative similarity. It is an important task so read each story twice.

Base Story: "{base_story}"
Story A: "{story_a}"
Story B: "{story_b}"

Two-Shot: We additionally experimented with a two-shot prompting approach to provide the model with explicit examples of the task. However, we observed a slight performance degradation compared to the zero-shot.

Prompt: You are a Researcher. In this study, you are tasked with identifying similar stories. You will be presented with three stories: a base, and two choices, a and b. You are to determine which of the candidate stories, a and b, is the most similar to the base. Specifically, you will consider the stories' narrative similarity.

Example 1:
User:
Base Story: "{base_story}"
Story A: "{story_a}"
Story B: "{story_b}"
Model: "{text_a_is_closer}"
Example 2: User:
Base Story: "{base_story}"
Story A: "{story_a}"
Story B: "{story_b}"
Model: "{text_a_is_closer}"
Reply with A if story A is more similar OR B if story B is more closer.

4 Experimental setup

We implemented our system using Kaggle provided GPU resources (two NVIDIA Tesla T4 GPUs). The model was accessed via the HuggingFace Transformers library (Wolf et al., 2020). To evaluate the model's raw narrative reasoning capabilities, the data was fed into Llama-3.2-3B-Instruct in its original format.

The dataset for all phases is hosted on the official competition site[2]. All inferences were performed with a maximum generation length of 350 tokens. To ensure reproducibility we employed greedy decoding by setting the temperature to 0 and top-p to 1.0 (with do_sample=False). Additionally, no fine tuning or hyperparameter optimization was conducted, as the evaluation was designed for a controlled zero-shot prompting study.

Since our system utilizes a generative approach, the model's output typically exceeded 200 tokens, containing a detailed comparative rationale followed by its final judgment. To ensure 100% accuracy in mapping these narrative justifications to the required competition labels (story A or story B), we have manually annotated the final results. For most of the triples, the model provided a clear preference (*After analyzing the three stories, I have*

[1]All prompts and code can be found here https://github.com/kjain93/pllama-semeval26.git

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

determined that Story A is the most similar to the base story. Here is why...), the response was initially extracted via a Python script designed to identify specific choice indicating phrases. However, to ensure that the correct response was consistently mapped to the corresponding candidate, particularly in cases of complex or non-standard phrasing, we manually reviewed each output and assigned the final choice. While this manual annotation process was time consuming, it was necessary to ensure the integrity of our results and to eliminate any extraction errors that might have been introduced.

Further, in one case the model explicitly refused to output any candidate due to the narratives' explicit content. In this case we use the default choice and selected story A as the candidate. After the extraction of predictions, we exported the results in the jsonl format as required by the task organizers.

5 Results

Our zero-shot Llama-3.2-3B-Instruct model achieved an accuracy of 55.5% on the task, slightly above the 50% random baseline. On the other hand the accuracy of two-shot setting was only 51%.

While the zero-shot model performs above chance, the margin remains limited indicating its instability in binary narrative similarity judgments under a purely zero-shot prompting setup. This result highlights the difficulty of eliciting consistent structural reasoning from instruction-tuned LLMs in complex comparative narrative tasks (Brown et al., 2020).

The results are noteworthy. While the task organizers achieved a higher accuracy of 67% using GPT-4o-mini in a zero-shot prompting setting, our model's accuracy is comparatively lower. We attribute this performance gap to differences in model size and training data. GPT-4o-mini is significantly larger than our 3 billion parameter model, enabling it to better capture nuanced narrative details and the world knowledge required to track similarity relationships across multiple stories.

To further understand Llama's low accuracy and sources of errors, we conduct a detailed error analysis. By examining both the model's predictions and the reasoning it generates, we aim to identify patterns of misclassifications, potential surface level biases, and limitations in the model's structural reasoning. This analysis provides insights into the strengths and weaknesses of zero-shot prompting method for narrative similarity and informs potential directions for improving model performance.

5.1 Error Analysis

5.1.1 Hallucination

To understand model's performance, we first manually review some of the generated justifications provided by the model. Our analysis shows that many times the model correctly identifies the structural similarities between the anchor story and the correct candidate, but ultimately selected the wrong candidate due to hallucination or an over-reliance on a single shared named entity pointing towards some surface level bias in model's responses. Examples are given in Table 2. In first example, in order to show similarity between anchor and story A, the model mentions that the dentist transforms into a rich merchant as in the base story where the frog transforms into a prince. However, it is not the case. The dentist and merchant are two different people. While in second example, model reasons that the anchor story and story A are similar as both main characters adopt a new identity. Though it is true for Story A, in anchor story the main character 'Lary' doesn't take any new identity.

5.1.2 Lexical Bias and Reasoning Analysis

We also examine if the model relies on superficial cues by analyzing the lexical and entity overlap between the anchor story and the candidate stories. The task paper reports that a simple Jaccard similarity baseline achieves 56.25% accuracy on Track A, only slightly above the 50% random baseline, suggesting that lexical overlap alone is a weak signal for this task (Hatzel et al., 2026). Consistent with this, the analysis of LLama's reasoning shows that the model does not systematically prefer the candidate with higher lexical overlap. The overall lexical alignment (or similarity) rate and the lexical bias in incorrect predictions are both 0, while the model level lexical and entity bias are approximately 50.25% which is close to chance. These findings suggest that the model's predictions in zero-shot setting are not driven by simple surface features such as word overlap or shared entities.

5.1.3 Linguistic Markers of Error

We further analyze the density of causal connectives and hedging expressions in generated reasoning of the model. The goal is to understand if the model tries to persuade the user to believe its judgements in case of incorrect predictions or not. The causal connectives can be defined as ex-

Anchor Story	Candidate Story	Llama's Justification
(.....)In the original Grimm version of the story, the frog's spell was broken when the princess threw the frog against the wall, at which he transformed back into a prince, while in modern versions the transformation is triggered by the princess kissing the frog. (.....)	**StoryA:** A lovely young woman named Wanda finds herself penniless after leaving a hospital. She decides to visit a man she recently met in a café (....) to ask for money. The man, who proves to be a dentist, asks her to take her place in a chair and starts examining the state of her teeth. (....) Happily, next day she meets another fine gentleman in the same café, a rich merchant from Kazan, and gets her financial position reasonably improved.	After carefully analyzing the three stories, I have determined that Story A is the most similar to the base story. Here's why, (....) 2. **Transformation**: In the base story, the Frog Prince transforms back into a prince after the princess's actions (kissing or throwing him against the wall). In Story A, the dentist transforms into a rich merchant after Wanda's actions (.....)
Larry (Stanley Tucci), an expat piano player, settled in a remote island village seven years ago. Now he runs a small boutique hotel with his girlfriend, (.....) Things soon start to go awry, as his long lost daughter Frankie (Jena Malone) arrives out of the blue, looking for him, another social-climber, Sebastian Burrows (Hugh Bonneville), latches on to the project convincing Veronica to give it to him and Larry's girlfriend grows suspicious of his relationship with the composer's former muse, Helena (.....)	**Story A:** "Paul Exben is a remarkable success story: (.....) Paul decides to seize a radical opportunity. Assuming the identity of the deceased man, he escapes to a remote region in former Yugoslavia, (.....)	"After carefully reading each story twice, I have analyzed the narrative similarities between the base story and the two candidate stories, A and B. (.....) However, the most striking similarity between the base story and Story A is the theme of identity and the protagonist's desire to reinvent themselves. Both Larry and Paul assume new identities to escape their past and start anew.(.....)

Table 2: Examples of hallucination in Llama's justification while makings its choice

plicit discourse markers signaling logical entailment (e.g., because, therefore, as a result). In this work we select a list of 15 such expressions from the Penn Discourse Treebank 3.0 (Webber et al., 2019) following the 'Contingency.Cause' sense taxonomy. Whereas for hedging expressions we include contrastive conjunctions (like although, however, on the other hand, etc) signaling uncertainty in model's responses.

The incorrect predictions contained slightly more causal markers on average (1.07) than the correct predictions (0.91). Similarly, the number of hedge words used for the incorrect responses (4.50) were little higher than the correct response (4.39). Although the differences are small, they still indicate towards model's behavior of generating performative reasoning using an increased frequency of such markers to produce persuasive justifications for its judgement.

We can interpret this pattern as evidence that the model compensates for deficiencies in underlying reasoning by amplifying surface level signals of coherence and sophistication. As a result, explanations accompanying incorrect predictions may appear equally or even more persuasive than those supporting correct answers.

5.2 Ablation Study

The presented approach resulted in lower accuracy, therefore after the submission we have tested a small embedding based model all-MiniLM-L6-v2 (Wang et al., 2020) to see if we can still employ a smaller model to classify the narratives. We calculate similarity using the cosine similarity between the anchor embedding (u) and each candidate embedding (v_a, v_b).

To handle the model's sequence length constraints, stories were truncated to the first 512 tokens, and we applied mean pooling across the output layer to derive a fixed 384 dimensional vector representation for each narrative. The model, without any additional training, achieved 59.75% accuracy, outperforming the prompting approach.

The 4.25% performance gap between the embedding model and the Llama model is notable, suggesting that despite access to high level reasoning and pretraining on 9 trillion tokens, the LLM is outperformed by a much smaller (22M parameter) bi-encoder. This observation aligns with prior work showing that specialized embedding models can outperform general purpose LLMs on certain structured reasoning benchmarks (Hatzel and Biemann, 2024).

6 Conclusion

This work examined the effectiveness of LLaMA-3.2-3B-it for zero-shot narrative similarity classification task. The model achieved performance above the random baseline (55.5%), indicating that it captures some aspects of high level narrative alignment. However, the relatively modest margin suggests that zero-shot generative prompting alone does not yield stable binary similarity judgments in this setting. The qualitative analysis of model's generation provides a more nuanced picture of its behavior. In multiple instances, the model's generated explanations correctly identified event correspondences such as shared theme, event chains, or outcomes, but hallucinated in between. This reveals a gap in model's understanding. Despite its struggle, the model doesn't rely on simple lexical overlap for this task. Further analysis of linguistic markers indicates that errors are not strongly driven by surface level cues. Instead, incorrect predictions often exhibit equally sophisticated reasoning patterns suggesting that well formed persuasive justification does not necessarily translate in accurate classification. Overall, the findings suggest that small instruction-tuned LLMs can approximate high level narrative understanding in a zero-shot setting, but their decision making remains unstable for structured binary evaluation tasks.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *CoRR*.

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.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.