NarSiL at SemEval-2026 Task 4: A Multi-Expert, Multi-Pathway System for Narrative Story Similarity

Bogdan Octavian Grecu[1], **Costin Chiru**[2] **and Oana Cocarascu**[1]
[1]Department of Informatics, King's College London
[2]Politehnica University of Bucharest
bogdan.grecu@kcl.ac.uk, costin.chiru@upb.ro,
oana.cocarascu@kcl.ac.uk

Abstract

We present **NarSiL**[1] (**Nar**rative **Si**milarity **L**earners), our system for SemEval-2026 Task 4 Track A on Narrative Story Similarity. NarSiL employs a two-stage architecture: a Mixture-of-Experts (MoE) initial classifier that also leverages supermajority voting across three large language models (Gemma-3-12B, GPT-3.5-turbo-instruct, and Gemini-2.5-Flash) over multiple runs, followed by a structured three-pathway fallback for ambiguous cases. The three pathways correspond directly to the task's three core similarity components, abstract theme, narrative outcome, and course of action. Each path yields a similarity score corresponding to its respective component, and the scores are then combined through a weighted aggregation step. NarSiL achieves 64.25% accuracy on the official test set. An improved score of 70.25% is obtained by considering only the supermajority voting of GPT, followed by the previously described fallback.

1 Introduction

Computational models of narrative similarity have long been a subject of interest in natural language processing and computational narratology (Chaturvedi et al., 2018; Chun, 2024). Determining whether two stories share the same underlying narrative structure, despite differences in setting, characters, or temporal context, poses a fundamental challenge at the intersection of semantics, discourse understanding, and narrative theory.

SemEval-2026 Task 4 (Hatzel et al., 2026) formulates narrative similarity as a binary classification problem: given an anchor story and two candidate stories, a system must identify which candidate is more narratively similar to the anchor. Narrative similarity is defined along three dimensions: the *abstract theme* (the ideas and motives of the story), the *course of action* (the sequence of central events and turning points), and the *outcomes* (the results of the story). The dataset consists of over 1,000 Wikipedia story summary triples, with the test set for track A comprising 400 items.

Our system, NarSiL, first deploys a Mixture-of-Experts classifier that integrates three large language models (LLMs). To resolve straightforward cases quickly and with high confidence, the system requires two strict conditions to be met: each individual LLM must reach an internal supermajority (at least 4 out of 5 identical responses), and all three LLMs must unanimously agree on that result. For cases where this condition is not met, the system falls back to three specialized sub-systems (one for each narrative dimension), aggregating their outputs via a weighted scoring function. A key design principle of our system is modularity and customizability: the weights for each narrative dimension in the final aggregation step can be tuned to align with different theoretical or empirical assumptions regarding the relative importance of abstract theme, outcomes, and course of action.

During the shared task, our initial submission utilized the "base" NarSiL architecture, which required a strict three-way expert consensus before defaulting to the fallback pathways. More experiments revealed that this strict consensus created a bottleneck. Consequently, we developed an improved 'GPT-only' supermajority system which we submitted during the evaluation test, obtaining the 13th position on the leaderboard. Post evaluation results allowed us to evaluate a "final" system (achieving 71.50%), which balances high-confidence multi-model voting with improved coverage. To provide a comprehensive overview of our methodology, this paper details the "base" architecture, the officially submitted system, and our post-evaluation improvements.

[1]The legendary sword from J.R.R. Tolkien's works, famous for being reforged from shards. This reflects how NarSiL pieces together fragmented expert LLM opinions and distinct narrative dimensions to forge a final answer.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3035–3044

2 Background

The task of detecting narrative similarity has a long-standing history in NLP and computational literary studies. Early work focused on symbolic representations of narrative structure, including narrative chains and schemas (Chambers and Jurafsky, 2009) and event-based representations (Finlayson, 2012). More recently, neural embedding models trained on narrative data have enabled corpus-scale similarity retrieval (Hatzel and Biemann, 2024a).

SemEval-2026 Task 4 (Hatzel et al., 2026) introduces a contrastive annotation setup that departs from prior similarity judgments (Fisseni and Löwe 2012; Chen et al. 2022). Framing the task as a binary preference over triples leads to more consistent annotations and allows for straightforward evaluation using accuracy. The dataset was constructed from Wikipedia story summaries drawn from the Tell-Me-Again corpus (Hatzel and Biemann, 2024b), with candidate triples sampled using a narrative embedding model and filtered to include only difficult cases.

Track A, which our system targets, evaluates narrative similarity through direct comparison by selecting one of two options most similar to a given reference story (i.e. the anchor story). Many top-performing systems on this track employ LLM ensembling. Our system also follows this approach while incorporating a principled, theory-driven fallback mechanism that decomposes the decision into the task's three explicit narrative dimensions.

3 System Description

The NarSiL[2] pipeline consists of ten sequential steps organised into two major stages, as illustrated in Figure 1. In the **first stage** (Steps 1–2), a Mixture-of-Experts classifier attempts to reach a high-confidence consensus decision. In the **second stage** (Steps 3–10), cases where consensus was not reached are processed through three independent analysis pathways (abstract theme, narrative outcome, and course of action), whose outputs are fused into a final prediction.

3.1 Stage 1: Mixture-of-Experts Classification

Steps 1-2: MoE Inference and Consensus Analysis We use three LLMs: Gemma-3-12B-it (Kamath et al., 2025) (ran via HuggingFace Transformers), GPT-3.5-turbo-instruct (OpenAI, 2022)

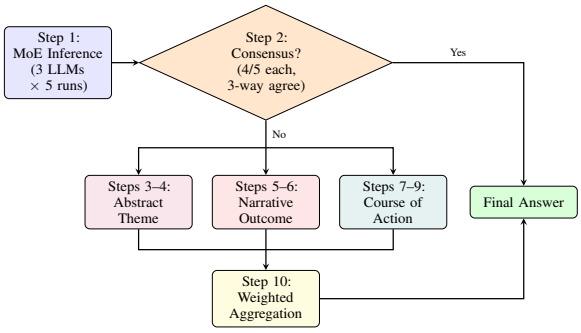

Figure 1: Overview of the NarSiL pipeline: Stage 1 (top row) attempts to resolve instances via LLM consensus, while Stage 2 (bottom rows) applies three independent pathways and aggregates their scores.

(accessed via the OpenAI API), and Gemini-2.5-flash (Anil et al., 2025) (accessed via the Google GenAI SDK). Each model is prompted five times with the same input using a temperature of 0.7, yielding a total of 15 independent predictions per instance. The system prompt instructs the model to analyse the anchor story and the two candidate stories across all three narrative dimensions, ultimately indicating which of story A or story B is closer to the anchor story.

We apply a supermajority rule to each LLM: if at least 4 out of 5 runs select the same candidate, the LLM's prediction is confirmed. When all three LLMs converge independently on the same answer, a decision is issued at this stage and the instance is not processed further.[3]

3.2 Stage 2: Three-Pathway Fallback

For instances that do not reach consensus, the system proceeds to three parallel analysis pathways, each corresponding to one of the task's three narrative dimensions. We selected Gemma-3 12B as the backbone for the fallback pathways to explore the capabilities of an open-weights model for granular narrative extraction. While commercial GPT models showed stronger performance in Stage 1, using Gemma-3 allowed for reproducible, cost-effective execution of the complex, multi-step prompting required across the three fallback dimensions.

3.2.1 Pathway 1: Abstract Theme

Steps 3-4: Theme Abstraction and Abstract Comparison For each of the three stories (anchor, candidate A, candidate B), Gemma-3 12B is prompted to extract the abstract theme. The

[2]Code available at https://github.com/grecu-bogdan-13/NarSiL

[3]Section 4 motivates the decision to rely solely on GPT's supermajority.

model is instructed to identify the protagonist's goal and primary obstacle, summarize the resolution, and generate a single-sentence abstract theme that omits proper names, specific locations, and, story-specific vocabulary. This forces the model to focus on underlying narrative patterns rather than surface-level details.

The three extracted theme statements are then passed to Gemma, which compares the anchor story with candidate A and candidate B, respectively, and outputs the closer candidate along with a confidence label (High, Medium, or Low). The confidence label is later used in the aggregation step to weigh the contribution of this pathway.

3.2.2 Pathway 2: Narrative Outcome

Steps 5-6: Outcome Analysis and Outcome Similarity Scoring For each story, Gemma-3 12B is first prompted to extract a 2–3 sentence summary of the outcome, focused on the protagonist's final state and the resolution of the central conflict. It then classifies the outcome into one of seven categories from a custom outcome taxonomy: Total Victory, Compromised Success, Pyrrhic Victory, Noble Failure, Tragic Failure, No Change, Ambiguous/Open.[4] Additionally, a sentiment score for each outcome summary is computed using *siebert/sentiment-roberta-large-english* (Hartmann et al., 2023) from Hugging Face, and a semantic embedding is produced using *all-mpnet-base-v2* (Song et al., 2020).

An outcome similarity score between each candidate and the anchor is computed as a weighted combination of three distances:

$$\delta(c, a) = w_{\text{cat}} \cdot d_{\text{cat}} + w_{\text{sent}} \cdot d_{\text{sent}} + w_{\text{sem}} \cdot d_{\text{sem}} \quad (1)$$

where d_{cat} is a categorical penalty (with values in the range $[0, 1]$) derived from a hand-crafted distance matrix over the seven outcome classes, reflecting their narrative proximity; d_{sent} is the absolute difference between sentiment scores normalised to $[0, 1]$; and d_{sem} is the cosine distance between the outcome embeddings, also normalised to $[0, 1]$. The weights are set as $w_{\text{cat}} = 0.60$, $w_{\text{sent}} = 0.25$, and $w_{\text{sem}} = 0.15$. These values were determined empirically by tuning on a subset of the development set to maximize accuracy. The candidate with the lower overall distance is selected as the closer story in this pathway.

3.2.3 Pathway 3: Course of Action

Step 7: Coreference Resolution Pronoun chains in the three stories are resolved using *LingMess-Coref* (Otmazgin et al., 2023), a state-of-the-art neural coreference resolution model. Canonical entity names are selected by Gemma-3 12B from the coreference cluster mentions. This step normalises entity references (*he* → *John Smith*) prior to event extraction, reducing fragmentation in the extracted event chains.

Step 8: AMR-based Event Extraction After coreference resolution, each story is segmented into sentences using spaCy[5] and parsed into Abstract Meaning Representation (AMR) graphs using a BART-large AMR (Banarescu et al., 2013) parser *model_parse_xfm_bart_large-v0_1_0* (Jacob, 2024). From each AMR graph, we extract action predicates (i.e. PropBank (Palmer et al., 2005) rolesets, such as `run-01`) that are identified as semantically salient. Salience is determined using WordNet (Miller, 1994) for verb supersense classification and SemLink PropBank-to-VerbNet mapping (Loper et al., 2007; Schuler, 2006), filtering for predicates in categories relevant to a story's course of action, such as Possession, Change, Motion, Causal, and Aspectual. This yields an ordered sequence of action predicates for each story, each annotated with category tags, polarity, and resolved argument values.

Step 9: Narrative Classification Gemma-3 12B receives the three action predicate sequences (anchor, candidate A, candidate B) and is prompted to identify which candidate sequence is structurally closer to the anchor with respect to the course of action. The model is instructed to focus on predicate types, polarity (negation vs. positive), role alignment (ARG0, ARG1), and the narrative arc structure. The output specifies the closer story and a confidence label (High/Medium/Low).

3.2.4 Final Aggregation

Step 10: Weighted Score Fusion The outputs of the three pathways are combined into a single aggregate score for each candidate. Let t, a, and o denote the abstract theme, course of action, and outcome pathways, respectively. Each pathway yields a score in $[0, 1]$ based on its decision and associated confidence label ($\textit{High} = 1.0$, $\textit{Medium} = 0.8$,

[4] Definitions for these outcomes are given in Appendix A.

[5] https://github.com/explosion/spaCy

Low = 0.6). The final weighted distance for candidate X is:

$$D(X) = W_t \cdot s_t(X) + W_a \cdot s_a(X) + W_o \cdot s_o(X) \quad (2)$$

where $W_t = 1.0$, $W_a = 1.2$, and $W_o = 1.5$. These values were determined empirically through testing on development set samples. Manual error analysis indicated that the narrative outcome and course of action extractions were generally more reliable than the theme pathway, prompting us to increase W_o and W_a. The candidate with the lower aggregate distance is selected as the final answer.

3.3 Customizability and Design Principles

A deliberate design principle of NarSiL is that the system is fully modular and weight-configurable. The aggregation weights W_t, W_a, and W_o in Equation 2, as well as the component weights in the outcome scorer (Equation 1), can be tuned independently to reflect different theoretical or domain-specific priors. While our empirical results (Section 5) indicate that this explicit three-pathway fallback underperforms when compared to the LLM prompting results, its modular design serves an additional, perhaps more important purpose: acting as a diagnostic probe. By forcing the system to explicitly decompose narrative similarity into abstract theme, outcome, and action, we can isolate exactly where current structural NLP techniques succeed and fail. Ultimately, this architecture demonstrates that explicitly aligning narrative components via structured extraction (e.g., AMR parsing, thematic summarization) is currently much more brittle than relying on the implicit, black-box judgments of LLMs. However, NarSiL's configurable framework provides a foundation for diagnosing and improving these individual extraction bottlenecks in future work.

4 Further Improvements to NarSiL

Our analysis of the base system revealed a key limitation in the MoE consensus design. The three-way consensus condition is highly precise: when all three experts (Gemma-3 12B, GPT-3.5-turbo-instruct, and Gemini-2.5-Flash) independently achieve a 4/5 supermajority and give the same answer, accuracy reaches 84.57%. However, this applies to only 43.75% of the test instances (175/400), meaning that more than half of all instances are delegated to the three-pathway fallback, which achieves only 48.44% accuracy. The base system is therefore restricted by its fallback stage rather than its consensus stage.

This observation led to a change in the system architecture: rather than requiring three-way expert agreement, the final submitted system conditions only on GPT-3.5-turbo-instruct's supermajority. If at least 4 out of 5 GPT runs agree on the same candidate, that answer is committed directly. Only the instances where GPT fails to reach a 4/5 agreement are passed to the three-pathway fallback.

This change has a substantial impact on the pipeline's behaviour. The GPT-only supermajority is triggered for 84.5% test instances (338/400), reducing the fallback load from 225 instances to just 62. Although GPT-3.5-turbo-instruct on its own is less accurate than the full three-way consensus (76.04% vs. 84.57% when the MoE supermajority applies), the main benefit is that the weaker fallback pathways are invoked much less frequently. On the official test set, this configuration achieves 70.25%, a substantial improvement over the GPT-4o-mini baseline (67.00%) and over the NarSiL base system (64.25%). The improvement demonstrates that, for this task, coverage of the stronger classifier matters more than the precision of a stricter consensus rule.

Performance can be further enhanced by retaining the three-way consensus and supermajority, while introducing the GPT supermajority as an intermediate step before the three-pathway fallback. In this case, performance improves to 71.5%. We will refer to this approach as our "final" NarSiL system.

5 Evaluation

We evaluate NarSiL on the official SemEval-2026 Task 4 Track A dataset (Hatzel et al., 2026), which provides 200 development triples and 400 test triples. We report accuracy on the official test set.

5.1 Overall Performance

Table 1 shows that NarSiL achieves 70.25% accuracy on the test set, ranking 13th (tied) out of 44 participating teams. This result surpasses the task baseline (GPT-4o-mini with 67.00%) and is well above the median submission accuracy. In comparison, the base system, without the improvements described in Section 4, achieves 64.25% accuracy on the test set.

5.2 Stage 1: MoE Consensus Analysis

Table 2 compares our systems' performance on the test set. The base system's three-way consensus is

System	Test (%)
Random baseline	50.00
Jaccard Similarity	56.25
GPT-4o-mini baseline	67.00
NarSiL (base)	64.25
NarSiL (GPT-only)	**70.25**
NarSiL (final)	**71.50**
Best system (COGNAC)	78.00

Table 1: Accuracy of NarSiL on the test set, compared to the official baselines and the top-performing team. NarSiL ranked 13th out of 44 teams on Track A.

Configuration	Test Acc. (%)
Gemma-3 12B (4/5 supermajority)	63.52
GPT-3.5-turbo-instruct (4/5 supermajority)	76.04
Gemini 2.5 Flash (4/5 supermajority)	71.83
Base system (Three-way agreement)	
Consensus cases (175/400, 43.75%)	84.57
Fallback cases (225/400, 56.25%)	48.44
Overall	**64.25**
GPT-only supermajority system	
Supermajority cases (338/400, 84.5%)	76.04
Fallback cases (62/400, 15.5%)	50.00
Overall	**70.25**
"Final" system	
Consensus cases (175/400, 43.75%)	84.57
GPT-only supermajority (163/400, 40.75%)	65.64
Fallback cases (62/400, 15.5%)	50.00
Overall	**71.50**

Table 2: Accuracy breakdown by MoE configuration on the test set. Note that the standalone MoE expert accuracies (first three rows) reflect performance *only* on the subset of instances where that specific model achieved a 4/5 supermajority, not on the full dataset.

highly precise when it is triggered (84.57%), but does so on only 43.75% of instances, leaving the majority of cases to the weaker fallback pathways. The GPT-only system, instead, relies solely on GPT-3.5-turbo-instruct's supermajority (4/5 votes), which is applied to 84.5% of instances and substantially reduces the fraction of cases that reach the fallback. The "final" system improves the performance of the GPT-only supermajority system, by using the three-way agreement where possible. Among the individual MoE experts evaluated in isolation, GPT-3.5-turbo-instruct (76.04%) outperforms Gemini 2.5 Flash (71.83%) and Gemma-3 12B (63.52%).

Pathway	Acc. on non-consensus (%)	
	Base System	Improved System
Abstract Theme (Steps 3-4)	48.89	43.55
Narrative Outcome (Steps 5-6)	50.67	51.61
Course of Action (Steps 7-9)	49.78	51.61
Weighted Aggregation (Step 10)	48.44	50.00

Table 3: Accuracy of individual fallback pathways and the final aggregation.

5.3 Stage 2: Three-Pathway Fallback Analysis

Table 3 reports the accuracy of each individual fallback pathway, as well as their final weighted aggregation, across both base and "final" systems. The base system was evaluated on 225 non-consensus test instances, while the improved system was evaluated on the 62 instances where GPT did not reach a supermajority. In both setups, the narrative outcome pathway achieves the highest individual accuracy (tying with the course of action pathway in the improved system), while the abstract theme pathway consistently performs the worst. Furthermore, the weighted aggregation of all three pathways fails to improve over the best-performing individual pathways in either system, highlighting the need for more effective fusion strategies.

5.4 Error Analysis of Fallback Pathways

To better understand the performance of the individual fallback pathways (Table 3), we conducted a qualitative analysis. Importantly, these pathways were evaluated exclusively on the subset of triples that failed to reach a supermajority consensus during Stage 1 MoE phase. These represent inherently ambiguous and difficult narrative comparisons where LLMs struggled to make a decision. Within this challenging context, our findings show that explicitly parsing narrative structure introduces rigid bottlenecks that end-to-end LLM reasoning naturally avoids.

Abstract Theme The theme pathway highlights the inherent difficulty of balancing narrative abstraction with distinguishing detail. By instructing the model in Step 3 to remove proper names, locations, and domain-specific vocabulary to find the core narrative, the resulting abstractions sometimes became too broad. Consequently, when the comparative model in Step 4 evaluated the candidates, it occasionally struggled to differentiate them. The nuanced thematic parallels present in the raw text were abstracted away, demonstrating that while removing surface details is necessary to find struc-

tural similarities, over-abstraction can inadvertently remove the exact signals needed to distinguish between closely related narrative candidates. This issue is especially pronounced for the triples that are difficult to differentiate.

Course of Action The AMR-based pipeline (Steps 7–9) illustrated the difficulty of capturing narrative equivalence across diverse surface realizations. We identified two primary challenges:

1. **Lexical Diversity in Parsing:** Because the system relies on strict PropBank roleset alignment, it can struggle to align structurally similar events expressed through different verbs. For example, an anchor story where a character *escapes* and a candidate story where a character *flees* might yield divergent predicate chains. The symbolic representation tends to penalize these lexical differences, whereas an embedding-based or LLM approach naturally absorbs the semantic overlap.

2. **Predicate Sparsity:** The Wikipedia summaries in the dataset are dense but relatively short. After applying our semantic salience filter (retaining specific VerbNet categories like Motion or Causal), the resulting event chains were occasionally sparse. This reduced structural context can make it difficult for the comparative model (Step 9) to reliably reconstruct and compare the full narrative arcs.

Ultimately, these observations demonstrate that while human annotators may easily isolate concepts like "abstract theme" or "course of action," computationally reverse-engineering these dimensions via structured, multi-step extraction remains a complex challenge, often trading necessary narrative nuance for structural rigidity.

5.5 Discussion

Despite the overall success of the system, our analysis reveals several limitations within the individual fallback pathways. First, the poor performance of the abstract theme pathway is somewhat surprising given that abstract theme was reported as the most frequently decisive factor in human annotations (selected in 98.60% of cases). One explanation is that our two-step abstraction and comparison pipeline is prone to compounding errors: an imperfect abstraction at Step 3 inherently degrades the comparison at Step 4.

Second, the results obtained in the AMR-based course of action pathway suggest that, while the symbolic event representation is linguistically grounded, it struggles to capture narrative-level patterns. Sparsity in the extracted predicate chains, particularly for shorter summaries, and the challenge of aligning predicate sequences across stories with different surface realizations poses a significant bottleneck.

Finally, while the outcome pathway achieved the highest individual performance (validating our combined analysis of outcome category, sentiment score, and semantic similarity), the system currently relies on static aggregation. The inability to dynamically adapt the weights of these three pathways limits the overall fusion strategy.

6 Conclusion

We presented NarSiL, a modular multi-expert, multi-pathway system for narrative story similarity. Our system combines LLM ensemble voting with a theory-grounded three-pathway fallback. While the fallback addresses the task's explicit narrative dimensions (theme, outcome, action), its lower performance acts as a diagnostic probe, revealing the current brittleness of explicit narrative decomposition compared to LLM classifiers. NarSiL achieves 70.25% accuracy on the official test set of SemEval-2026 Task 4 Track A, ranking 13th among 44 teams and surpassing the GPT-4o-mini baseline by 3.2%. An additional enhancement was introduced, increasing the score to 71.5%.

Future work will focus on addressing the current limitations of our individual fallback pathways. To mitigate compounding errors within the abstract theme pipeline, we plan to explore direct comparisons of raw story texts using theme-focused prompts. For the course of action pathway, further research will target overcoming the sparsity of extracted predicate chains and improving alignment across diverse surface realizations. Additionally, capitalizing on NarSiL's modular design, we intend to experiment with replacing or extending the outcome taxonomy and categorical distance matrix to support domain-specific narrative ontologies. Finally, we plan to transition from static aggregation weights to a fully dynamic optimization approach, learning to weigh the three narrative dimensions directly from data.

References

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, and 1331 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

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.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, Washington, USA. Association for Computational Linguistics.

Jon Chun. 2024. AIStorySimilarity: Quantifying story similarity using narrative for search, IP infringement, and guided creativity. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 161–177, Miami, FL, USA. Association for Computational Linguistics.

Mark Alan Finlayson. 2012. *Learning Narrative Structure from Annotated Folktales*. Ph.D. thesis, Massachusetts Institute of Technology.

Bernhard Fisseni and Benedikt Löwe. 2012. Which dimensions of narrative are relevant for human judgments of story equivalence?

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

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. 2024a. Story embeddings – narrative-focused representations of fictional stories. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 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 Fourteenth Language Resources and Evaluation Conference*, Turin, Italy.

Ben Jacob. 2024. amrlib: A python library that makes AMR processing simple. `https://github.com/bjascob/amrlib`.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Edward Loper, Szu-ting Yi, and Martha Palmer. 2007. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Linguistic Annotation Workshop*, pages 9–15, Prague, Czech Republic. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

OpenAI. 2022. Introducing ChatGPT. `https://openai.com/index/chatgpt/`. Accessed: 2026-03-01.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. Lingmess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 15680–15690.

A Typology of Narrative Outcomes.

Here we provide the details for the taxonomy of the seven narrative outcomes evaluated by our pipeline. These descriptions represent the exact instructional definitions provided to the LLMs during the inference and consensus steps.

- Total Victory - The protagonist achieves their primary goal with minimal permanent loss. The antagonist is defeated, and the world or the hero's life is objectively better than when the story began.
- Compromised Success - The main goal is achieved, but at a significant cost. The hero wins, but they had to sacrifice a secondary objective, a relationship, or a core part of their identity to get there.
- Pyrrhic[6] Victory - The protagonist wins, but loses in the process the very things that made the victory worth fighting for.
- Noble Failure - The protagonist fails to reach their goal, but they maintain their integrity or inspire others in the process. They lose the battle but "win" the moral high ground or spark a future change.
- Tragic Failure - The protagonist fails to achieve their goal due to a personal flaw, a mistake, or overwhelming external forces. This usually results in a worse state of affairs than the beginning of the story.
- No Change - The status quo is maintained. Despite the conflict, the world and the characters end up where they started.
- Ambiguous/Open - The ending is left to the reader's interpretation. It is unclear if the goal was met, or the story ends on a cliffhanger where the final outcome is yet to be determined.

[6]The term Pyrrhic comes from King Pyrrhus of Epirus (319–272 BC), a Greek general and statesman of the Hellenistic period. According to the historian Plutarch, he famously commented on his victory at Asculum: "If we are victorious in one more battle with the Romans, we shall be utterly ruined."

B Categorical Distance Matrix

In order to clarify the categorical distance penalty (d_{cat}) used in the narrative outcome pathway (Equation 1), we present the full symmetric distance matrix used to compute the penalty between any two identified narrative outcomes in Table 4.

The distances, ranging from 0.0 to 1.0, were defined based on the structural and emotional proximity of the outcomes. Exact matches incur no penalty (0.0). The matrix is designed along a spectrum from highly positive outcomes (Total Victory) to highly negative outcomes (Tragic Failure). For example, stepping from a Total Victory to a Compromised Success incurs a small penalty (0.20), while comparing polar opposites (Total Victory versus Tragic Failure) incurs the maximum penalty (1.0). Intermediate states, such as No Change or Ambiguous/Open, are assigned moderate penalties reflecting their neutral narrative positioning.

	TV	CS	PV	NC	AO	NF	TF
Total Victory (TV)	0.00	0.20	0.60	0.35	0.45	0.80	1.00
Compromised Success (CS)	0.20	0.00	0.40	0.15	0.25	0.60	0.80
Pyrrhic Victory (PV)	0.60	0.40	0.00	0.25	0.15	0.20	0.40
No Change (NC)	0.35	0.15	0.25	0.00	0.10	0.45	0.65
Ambiguous/Open (AO)	0.45	0.25	0.15	0.10	0.00	0.35	0.55
Noble Failure (NF)	0.80	0.60	0.20	0.45	0.35	0.00	0.20
Tragic Failure (TF)	1.00	0.80	0.40	0.65	0.55	0.20	0.00

Table 4: The predefined categorical distance matrix (d_{cat}) used to compute the penalty between different narrative outcome classifications.

C System Prompts

To ensure the full reproducibility of our methodology and to provide deeper insight into the system's operational mechanics, we present the core system prompts used across the NarSiL pipeline. The prompts below detail the exact instructions provided to the large language models for both the initial Mixture-of-Experts consensus stage (Section 3.1) and the specialized extraction tasks within the three fallback pathways (Section 3.2). In the prompts, injected variables are denoted by squiggly brackets.

Narrative Similarity Inference Prompt

You are an expert in narrative analysis. You are asked to identify narratively similar stories. We define Narrative similarity by three core similarity components: the ab-

stract theme, the course of action, and the outcomes of a story. We define these three aspects as follows:

- **Abstract Theme:** Describes the defining constellation of problems, central ideas, and core motifs of a story.
- **Course of Action:** Describes sequences of events, actions, conflicts, and turning points.
- **Outcomes:** Describe the results of the plot at the end of the text.

Your Task:
You will be provided with an "Anchor Story" and two candidate stories, "Text A" and "Text B". Determine which candidate (Text A or Text B) is more narratively similar to the Anchor Story. **Output Format:** 1. Provide a brief analysis (2-3 sentences). 2. End your response with a single line containing exactly: "Selection: Text A" or "Selection: Text B".

Anchor Story:
{anchor_text}
Text A:
{text_a}
Text B:
{text_b}

Narrative Abstraction Prompt

Analyze the story provided. 1. First, list the Protagonist's Goal and the Primary Obstacle. 2. Second, summarize the Resolution. 3. Finally, write an Abstract Theme Statement (1 sentence) based on the above.
Constraints for the Theme Statement:

- Describe the central human problem and the moral realization.
- Strictly avoid proper names, specific locations, or unique terminology (e.g., 'John' $\rightarrow$ 'a man').

Format your response exactly as follows:
Analysis: [Your analysis here]
Theme: [Your single sentence abstract here]

STORY: {story_text}
ABSTRACT:

Abstract Theme Comparison Prompt

You are an expert literary critic. You are given the abstract theme of an anchor story, and the abstract theme of other two stories, Story A and Story B.
Task: Identify which of the two candidates (A or B) shares a closer Abstract Theme with the Anchor.
Constraint: Ignore similarity in setting (e.g., space, medieval) or specific plot events. Focus ONLY on the central moral, philosophical question, or character arc type.
Reasoning: First, compare Anchor vs A. Then, compare Anchor vs B. Finally, output the winner. After determining the winner, classify your confidence in this decision as:

- **High:** The thematic overlap is obvious and distinct.
- **Medium:** The winner is closer, but the distinction is subtle.
- **Low:** Both are equally similar or dissimilar; the choice is a guess.

On the last line format your final answer exactly like this: {final_answer: , confidence: }

Anchor Abstract: {anchor_abs}
Story A Abstract: {a_abs}
Story B Abstract: {b_abs}
Comparison:

Narrative Outcome: Extraction Prompt

You are an expert narrative analyst. Your task is to read the story provided and extract the narrative outcome. Focus strictly on the final state of the protagonist and the resolution of the central conflict. Output a concise summary of 2-3 sentences.

STORY: {text}
OUTCOME SUMMARY:

Narrative Outcome: Classification Prompt

You are a literary classifier. Classify the provided outcome summary into exactly one of the following categories:

- Total Victory
- Compromised Success
- Pyrrhic Victory
- Noble Failure
- Tragic Failure
- No change
- Ambiguous/Open

Return ONLY the category name. Do not explain.

SUMMARY: {summary}
CLASSIFICATION:

Character Name Resolution Prompt

You are a text processing assistant.

Identify the best full proper name for a character referred to by these terms: [{candidates_str}].
Reply with ONLY the name inside double quotes (e.g., "John Smith").

Narrative Structural Similarity Prompt

You are an expert in computational narratology. Your task is to identify which of two candidate stories (Story A or Story B) is structurally closer to an Anchor Story based on the 'Course of Action'.
CRITERIA:

- **1.** Focus on the sequence of predicates (e.g., 'run-01') and frame tags.

- **2.** Focus on polarity (e.g., NOT_ vs positive) and role alignment (ARG0, ARG1).
- **3.** Analyze the narrative arc (e.g., MOTION → POSSESSION → CHANGE).

OUTPUT FORMAT:
You must return ONLY a JSON object with keys: 'analysis', 'closer_story' (A or B), and 'confidence' (High/Medium/Low).

ANCHOR STORY EVENTS:
{anchor_events_json}
STORY A EVENTS:
{story_a_events_json}
STORY B EVENTS:
{story_b_events_json}