Yam at SemEval-2026 Task 4: Failure-Driven Prompt Evolution for Narrative Comparison

Yen Yee Yam
National University of Singapore
Pinetree Research
e1354172@u.nus.edu

Hong Meng Yam
Stanford University
Pinetree Research
hongmeng@stanford.edu

Abstract

We present a structured, parameter-free system for SemEval-2026 Task 4 on Narrative Story Similarity. Instead of treating similarity as scalar embedding proximity, we align model reasoning with the task ontology by decomposing each story into abstract theme, course of action, and outcome, and performing contrastive comparison over these dimensions. Our primary contribution is a closed-loop, failure-driven prompt optimization procedure that iteratively refines concise guideline documents while keeping model parameters fixed and reverting updates that degrade performance, thereby isolating improvements attributable to structured reasoning rather than representation learning. Ontology-aligned decomposition alone achieves 70% accuracy on both train and test sets; with controlled guideline evolution, performance improves to 76% on train and 73% on test without additional supervision or fine-tuning. These results demonstrate that structured prompt optimization can meaningfully enhance contrastive narrative reasoning in a fully parameter-free setting.

1 Introduction

Narrative similarity involves recognizing shared abstract patterns of causality and progression across stories, beyond surface overlap. SemEval-2026 Task 4 (Hatzel et al., 2026a) formalizes this challenge through contrastive triple judgments: given an anchor and two candidates, systems must determine which candidate is narratively more similar, along three theory-informed dimensions of *course of action*, *outcome*, and *abstract theme* (Hatzel et al., 2026b). The task thus requires structured comparative reasoning rather than scalar similarity estimation.

Prior work on narrative modeling has largely focused on representation learning, from symbolic event schemas (Chambers and Jurafsky, 2009; Chambers, 2013) to distributed story embeddings (Hatzel and Biemann, 2024) and contrastive objectives (Sterner et al., 2026). While effective for capturing global structure, these approaches typically operationalize similarity as geometric proximity in embedding space, without explicitly aligning model reasoning to theory-defined narrative dimensions.

At the same time, preference-based learning has become central in NLP, from classical comparative judgment theory (Thurstone, 1994; Kiritchenko and Mohammad, 2017) to modern alignment frameworks such as RLHF (Ouyang et al., 2022; Jiang et al., 2025). However, such methods rely on parameter adaptation, making it difficult to isolate gains due to structured reasoning from gains due to representation learning.

In this work, we propose a parameter-free framework for ontology-aligned narrative comparison. Rather than learning new representations, we explicitly decompose each story into dimension-specific representations corresponding to abstract theme, course of action, and outcome, and perform contrastive reasoning over these components. We further introduce a closed-loop guideline evolution procedure that iteratively refines reasoning instructions using evaluation feedback, while keeping model parameters fixed. This design isolates improvements attributable to structured reasoning and prompt evolution.

Empirically, we show that ontology-aligned decomposition improves decision stability relative to direct prompting, and that reflective guideline evolution yields consistent accuracy gains with reduced variance. Our results demonstrate that structured, parameter-free prompt optimization provides a principled alternative to weight adaptation for modeling narrative similarity in contrastive settings.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3394–3398

2 Related Work

Narrative Structure Modeling. Early computational models represented narrative structure through symbolic abstractions such as plot units (Lehnert, 1981), narrative chains (Chambers and Jurafsky, 2009), and probabilistic event schemas (Chambers, 2013). Subsequent work incorporated graph-based decoding and event coreference to better capture temporal and script-level dependencies (Liu et al., 2018). More recent approaches embed narratives in distributed vector spaces, modeling relational structure via character networks (Lee and Jung, 2020), learning story-focused embeddings (Hatzel and Biemann, 2024), or applying contrastive objectives to capture narrative salience (Sterner et al., 2026). These methods advance representation quality but typically reduce similarity to scalar embedding distance, without explicit reasoning over theory-defined narrative dimensions.

Comparative Judgment and Preference Learning. Comparative evaluation has long been studied in psychometrics through Thurstone's law of comparative judgment (Thurstone, 1994) and Best–Worst Scaling (Kiritchenko and Mohammad, 2017). In modern NLP, pairwise preference signals underpin reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and broader alignment strategies (Jiang et al., 2025). These approaches demonstrate the effectiveness of preference supervision but rely on parameter updates and large-scale training.

Prompt Optimization and Parameter-Free Adaptation. An emerging line of work explores improving model behavior without updating model weights. Methods such as AutoPrompt (Shin et al., 2020), TextGrad (Yuksekgonul et al., 2024), and GEPA (Agrawal et al., 2025) treat prompts as optimization targets while keeping underlying parameters fixed. AutoPrompt performs gradient-based search over input tokens by computing gradients with respect to discrete trigger sequences, but does not modify model weights. TextGrad implements automatic differentiation over textual variables, using LLM-generated feedback as gradient-like signals within a computational graph defined over prompts and intermediate outputs. GEPA applies evolutionary reflection over execution traces, refining instructions through natural-language feedback and Pareto-based candidate selection. These approaches demonstrate that substantial behavioral improvements can be achieved through parameter-free adaptation.

However, prior prompt optimization research has primarily focused on improving aggregate task accuracy, without enforcing alignment between model reasoning and an explicit theoretical structure. In contrast, our framework constrains reasoning to follow the task-defined narrative ontology: each story is decomposed into representations of abstract theme, course of action, and outcome, and similarity is evaluated dimension-by-dimension. This explicit alignment between reasoning structure and evaluation criteria distinguishes our approach from generic prompt optimization.

Our framework integrates insights from narrative theory, comparative judgment, and prompt optimization. Unlike embedding-based similarity models, we treat narrative comparison as structured reasoning over explicit theoretical dimensions. Unlike reinforcement-based preference learning, we do not adapt model parameters. And unlike generic prompt optimization approaches, we explicitly align reflective prompt evolution with a narrative ontology, enabling controlled analysis of dimension-level reasoning in contrastive settings.

3 System Overview

Our system is a structured, prompt-only framework designed to align large language model reasoning with the task ontology of SemEval-2026 Task 4. It consists of two components: (i) ontology-aligned narrative decomposition and (ii) closed-loop failure-driven guideline evolution.

3.1 Ontology-Aligned Narrative Decomposition

Rather than directly comparing story summaries, we decompose each story into three explicit representations corresponding to the task definition:

- **Abstract theme**
- **Course of action**
- **Outcome**

Each story (anchor, A, B) is independently analyzed to produce structured summaries of these dimensions. The final similarity decision is then made through explicit dimension-level comparison.

This design enforces reasoning consistency with the evaluation criteria. By separating representation (extraction of narrative elements) from comparison (contrastive similarity judgment), we reduce

entanglement between surface-level lexical overlap and deeper narrative structure. The model is required to justify similarity along each dimension before producing a final decision.

Importantly, no model parameters are updated. All improvements arise from modifications to the prompting context.

3.2 Failure-Driven Closed-Loop Guideline Evolution

Our primary contribution is a controlled prompt optimization procedure that refines concise guideline documents using evaluation feedback.

After each evaluation pass:

1. Incorrect predictions are collected and analyzed.

2. Recurrent failure patterns are identified at two levels:

    - Errors in intermediate narrative extraction.
    - Errors in final similarity comparison.

3. Three compact guideline documents are regenerated:

    - Intermediate extraction guidance.
    - Decision-level comparison guidance.
    - Shared general reasoning principles.

To prevent uncontrolled drift:

- Revisions are accepted only if evaluation accuracy improves.

- If accuracy declines, guidelines are automatically reverted.

- Each document is constrained in length to encourage high-signal rules rather than overfitting heuristics.

This procedure treats prompts as structured artifacts that can be iteratively improved under evaluation constraints. Unlike reinforcement learning or fine-tuning, the underlying model remains fixed throughout. Improvements therefore reflect changes in reasoning guidance rather than representation learning.

3.3 Experiments

All experiments are conducted using `GPT-5.2` via API access. Model parameters are held strictly fixed across all conditions; no fine-tuning, reinforcement learning, or weight adaptation of any form is performed. All reported improvements arise solely from prompt structure and guideline evolution.

We observed non-trivial variance across individual inference calls prior to stabilization. As such, within each evaluation pass, decisions are stabilized through majority voting across repeated inference calls.

We conduct three experimental conditions:

1. **Direct Prompt.** The model directly compares the anchor with Story A and Story B without explicit narrative decomposition.

2. **Ontology-Aligned Decomposition.** Each story is first decomposed into abstract theme, course of action, and outcome before similarity comparison.

3. **Ontology + Self-Improvement.** The structured system is extended with closed-loop guideline evolution driven by failure analysis and automatic rollback.

This experimental design isolates the contributions of narrative structuring and prompt evolution while holding the underlying model constant.

4 Results

We evaluate using accuracy on the SemEval Track A narrative similarity dataset. Table 1 presents ablations isolating the effects of ontology-aligned decomposition and closed-loop guideline evolution.

Method	Train	Test
Direct Prompt	69%	68%
Ontology Decomposition	70%	70%
+ Self-Improvement	76%	73%

Table 1: Ablation results on SemEval-2026 Task 4 (Track A).

Effect of Ontology Alignment. Introducing explicit decomposition into abstract theme, course of action, and outcome improves test accuracy from 68% to 70%. Although modest, this gain is consistent across splits and indicates that structuring reasoning around theory-defined dimensions reduces reliance on superficial lexical similarity.

Effect of Failure-Driven Guideline Evolution. Adding closed-loop prompt refinement yields a substantial improvement, increasing train accuracy from 70% to 76% and test accuracy from 70% to 73%. Unlike the direct prompt baseline, this approach iteratively sharpens decision boundaries through structured failure analysis while maintaining generalization performance.

Interpretation. The ablations highlight three key observations:

1. Explicit ontology alignment yields measurable improvements over direct prompting.
2. Iterative, failure-driven prompt evolution produces further gains without updating model parameters.
3. A non-trivial portion of performance improvement in comparative narrative reasoning can be achieved through structured prompt optimization alone.

Taken together, these results suggest that careful reasoning alignment and controlled prompt refinement can meaningfully enhance performance in theory-grounded similarity tasks, narrowing the gap between static prompting and methods that rely on parameter adaptation.

5 Conclusion

We presented a structured, parameter-free system for narrative similarity that aligns large language model reasoning with an explicit task ontology. By decomposing stories into abstract theme, course of action, and outcome, and performing contrastive reasoning over these dimensions, we convert similarity prediction into structured comparative analysis.

Our primary contribution is a closed-loop, failure-driven prompt optimization procedure that iteratively refines concise guideline documents under strict evaluation constraints. Without updating model parameters or introducing additional supervision, this approach improves testing accuracy from 70% to 73%.

These results demonstrate that structured prompt design and controlled refinement can produce meaningful gains in comparative narrative reasoning. More broadly, they suggest that a portion of improvements commonly attributed to weight adaptation may instead arise from improved reasoning structure.

Future Work. An important direction is systematic comparison against fine-tuning and reinforcement-based alignment methods. Evaluating whether structured prompt optimization can approximate or complement fine-tuned performance would clarify the relative contributions of reasoning guidance versus representation adaptation, especially on a small model where we can run multiple ablation studies. Additionally, extending ontology-aligned prompt evolution to other theory-grounded reasoning tasks may reveal broader applicability of parameter-free optimization frameworks.

Limitations Our study deliberately isolates the contribution of structured reasoning and prompt evolution by holding model parameters fixed throughout. This design choice enables clean attribution of performance gains to reasoning structure rather than representation learning, but consequently does not quantify the gap between parameter-free optimization and fine-tuning. We view this comparison as a complementary research direction rather than a baseline omission: the questions of *whether* prompt optimization helps and *how much* it can substitute for weight adaptation are methodologically distinct, and conflating them would obscure the present finding. Additional limitations include the reliance on a large proprietary language model, whose effectiveness on smaller or open-weight architectures remains to be established, and the constrained search space of the guideline evolution procedure, which does not guarantee convergence to globally optimal reasoning strategies. Cross-domain robustness similarly warrants further evaluation. We regard direct comparison against fine-tuned and reinforcement-based baselines, as well as evaluation on open-weight models, as the most valuable next steps for situating parameter-free reasoning alignment within the broader landscape of adaptation methods.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, and 1 others. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807,

Seattle, Washington, USA. 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.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026a. 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, 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)*. To appear.

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.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2025. A survey on human preference learning for aligning large language models. *ACM Computing Surveys*, 58(6):1–39.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470.

O-Joun Lee and Jason J. Jung. 2020. Story embedding: Learning distributed representations of stories based on character networks. *Artificial Intelligence*, 281:103235.

Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive science*, 5(4):293–331.

Zhengzhong Liu, Teruko Mitamura, and Eduard Hovy. 2018. Graph based decoding for event sequencing and coreference resolution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3645–3657, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 4222–4235.

Igor Sterner, Alex Lascarides, and Frank Keller. 2026. Contrastive learning with narrative twins for modeling story salience. *arXiv preprint arXiv:2601.07765*.

Louis L Thurstone. 1994. A law of comparative judgment. *Psychological review*, 101(2):266.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.