AI-Monitors at SemEval-2026 Task 4: A Hybrid Embedding and LLM Ensemble for Narrative Similarity

Vishnu Tripathi[1], Prakhar Joshi[1], Azad Singh[2], Pragyanand Saho[1], Gaurav Kumar[1], Neel Mani[1], Piyush Arora[3]
School of Science & Technology, SRHU, Dehradun[1]
Department of Computer Science, DSVV, Haridwar[2]
American Express[3]
{vishnu.tripathi|prakhar.joshi|azad.singh|pragyanand.saho}@aicentre.org
gaurav.soni@aicentre.org | neelmani@srhu.edu.in | piyush.arora1@aexp.com

Abstract

Narrative similarity requires reasoning over the deeper structural properties of stories - shared themes, causal progression, and outcomes - rather than surface-level lexical overlap. We describe AI-Monitors, our system for SemEval-2026 Task 4 (Track A), which determines which of two candidate stories is more narratively similar to a given anchor.

We explore a progression of approaches - from embedding-based similarity to structured LLM prompting and ensemble construction - guided by four hypotheses about where narrative reasoning gains can be found. The final system achieves 75% test accuracy on 400 instances, ranking 3rd out of 47 systems and approaching the individual human annotator ceiling of 78%.

Our key findings are: i) structured few-shot prompting substantially outperforms dense embedding similarity; ii) selecting ensemble components by how differently they make errors - rather than by accuracy alone - produces stronger predictions; and iii) how you describe an example to the model affects its predictions.

1 Introduction

Stories can share similar words while being structurally and thematically very different — and two stories that look nothing alike on the surface can follow the same underlying narrative pattern. Standard semantic similarity models are not designed to capture this distinction. SemEval-2026 Task 4 (Hatzel et al., 2026) formalises this challenge as a comparative decision problem: given an anchor story and two candidate stories, a system must determine which candidate is more narratively similar to the anchor. Narrative similarity is operationalised along three annotation dimensions — *Abstract Theme*, *Course of Action*, and *Outcomes* — making this task fundamentally distinct from traditional semantic textual similarity benchmarks, which operate at the sentence level.

The task is intentionally challenging by design. Story summaries are sourced from Wikipedia and filtered to coherent 4–8 sentence synopses; candidate triples are then constructed so that trivial lexical overlap is minimised. Critically, only triples on which two independent commercial LLMs already *disagree* are retained for human annotation (Hatzel et al., 2026) — meaning the dataset is explicitly constructed to emphasise ambiguous, structurally nuanced cases. Under these conditions, human annotators achieve only approximately 78% accuracy against oracle labels, with an inter-annotator agreement of Krippendorff's $\alpha = 0.33$, reflecting the inherent subjectivity of narrative judgment. Organiser-provided baselines further illustrate the difficulty: token-based Jaccard similarity achieves 56.25% and a zero-shot GPT-4o-mini prompt achieves 67.0% on the 400-instance test set, leaving substantial room for structured approaches.

We participate in Track A, the direct comparison setting. We approach the task through a four-stage pipeline progressing from embedding-based similarity to structured LLM prompting and complementarity-driven ensemble construction. Code and prompt templates are available on GitHub.[1]

These four stages are each motivated by a research question, which we treat as testable hypotheses and empirically evaluate in Section 5:

H1 Do dense embedding models exhibit a performance ceiling for narrative similarity, given that they lack mechanisms for explicit reasoning over structured narrative dimensions?

H2 Does explicitly grounding prompts in the three narrative dimensions — *Abstract Theme*, *Course of Action*, and *Outcomes* — improve prediction accuracy beyond what zero-shot prompting achieves?

[1] https://github.com/aimonitors25/narrative-similarity-semeval2026

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2139–2148

H3 Do carefully selected k-shot demonstrations improve narrative comparison, and does prompt framing — such as how examples are labelled — affect how well this generalises?

H4 Do systems with strong standalone accuracy but low prediction agreement exhibit complementary error patterns, making them better ensemble candidates than accuracy-ranked selection alone would suggest?

2 Related Work

Narrative similarity and story datasets. Early work matched events across Wikipedia movie remake pairs using story-kernel methods, but relied on external structural metadata rather than human judgment (Chaturvedi et al., 2018). Hatzel and Biemann (2024b) released the Tell Me Again! corpus of cross-lingual story summaries, used to train *story-emb* (Hatzel and Biemann, 2024a) — a contrastive narrative embedding model later extended to better separate surface-similar but narratively distinct stories (Sterner et al., 2026). SemEval-2026 Task 4 (Hatzel et al., 2026) advances this line with the first large-scale multi-annotator dataset of narrative similarity judgments, grounded in direct human comparison and constructed to emphasise deep narrative structure through LLM-disagreement filtering — with downstream relevance for story recommendation, literary analysis, and plagiarism detection (Piper et al., 2021; Chun, 2024).

Dense embeddings for narrative representation. Sentence-BERT (Reimers and Gurevych, 2019) and large T5 encoders (Ni et al., 2022) establish strong semantic similarity baselines, further strengthened by contrastive pre-training (Wang et al., 2022) and SimCSE-style fine-tuning (Gao et al., 2021). Although *story-emb* (Hatzel and Biemann, 2024a) targets the narrative domain, evaluating it in this setting is potentially biased, as candidate triples were specifically constructed to maximise ambiguity (Hatzel et al., 2026). Critically, all these models optimise for surface-level overlap and lack mechanisms for explicit reasoning over structured narrative dimensions — directly motivating **H1**.

LLM prompting and in-context learning. Zero-shot LLM prompting without task grounding fails to fully utilise model reasoning, motivating explicit dimension-grounded prompting **(H2)**. Zhao et al. (2021) show that surface framing choices — label names and example order — cause systematic prediction biases independent of task signal. Recent work further shows that even minor prompt variations, including non-semantic tokens, label token can alter model predictions (Salinas and Morstatter, 2024; Lester et al., 2021).Whether such label priming extends to narrative comparison is what we investigate under **H3**.

Ensemble construction. Majority voting is a well-established ensemble strategy (Lam and Suen, 1997), but standard practice selects components by standalone accuracy, risking redundant combinations. We instead select by prediction disagreement rate, targeting complementary error patterns — addressing **(H4)**.

The remainder of this paper is organised as follows. Section 3 describes our methodology. Section 4 details the experimental setup. Section 5 presents results and analysis. Section 6 presents key findings & concludes.

3 Methodology

We describe the conceptual design of our multi-stage system for SemEval-2026 Task 4 (Track A), organised around four testable hypotheses (H1–H4), as illustrated in Figure 1 (Appendix A.1).

3.1 Embedding-Based Similarity (H1)

We began with dense semantic embedding models to represent stories as fixed-dimensional vectors. Each story (anchor, candidate A, candidate B) was encoded independently and similarity was computed using cosine similarity between story embeddings. The candidate with higher cosine similarity to the anchor was selected as the prediction. We use `sentence-transformers/t5-xxl` as our embedding baseline.

3.2 Contrastive Fine-Tuning (H1, continued)

Following **H1**, we fine-tuned the embedding models using a contrastive objective inspired by SimCSE (Gao et al., 2021), keeping the base embeddings frozen and training a lightweight projection layer on top (*SimCSE-style* fine-tuning). We evaluated three model variants — all-mpnet-base-v2, all-MiniLM-L6-v2, and sentence-transformers/t5-xxl (Reimers and Gurevych, 2019; Ni et al., 2022). Positive pairs were anchor–correct candidate pairs; negatives were sampled at a 2:1 ratio by pooling all possible story cross-pairings within the training

split and excluding known positives. The development set of 200 instances was partitioned explicitly into two train/validation splits — 100/100 and 150/50 — to assess whether gains held across different training sizes; fine-tuning was conducted on the train portion only in each case.

Parameter	Values Explored
Learning rate	10^{-3}, 10^{-4}, 10^{-5}
Projection dim.	768, 1024
Epochs	10, 30, 50, 100, 250
Batch size	4, 8, 16, 32

Table 1: Hyperparameter search space for SimCSE-style contrastive fine-tuning.

3.3 Prompting Strategy (H2, H3)

Given the structured definition of narrative similarity, we hypothesised that large language models may better capture alignment across narrative dimensions when explicitly guided — rather than left to interpret similarity freely. This stage evaluates H2 and H3 by grounding prompts from the annotation guidelines.

We explore three prompt families: **Prompt Family 1 (Baseline - no formal grounding)**, which provides only high-level instructions for narrative comparison with loosely referenced dimensions such as theme, actions, and outcome; **Prompt Family 2 (Aspect-Grounded)**, which explicitly grounds the task in the annotation guidelines provided by the organisers (Hatzel et al., 2026), incorporating formal definitions of Abstract Theme, Course of Action, and Outcome, and constraining responses via structured JSON. The key distinction between PF1 and PF2 is the depth of dimension grounding: PF1 provides concise, high-level summaries of each narrative dimension, while PF2 incorporates the verbatim formal definitions from the official annotation guidelines (Hatzel et al., 2026); and **Prompt Family 3 (K-Shot)**, which further extends this setup by incorporating curated examples that demonstrate how these grounded dimensions should be applied in practice. All prompts are submitted as a single continuous block - no system/user/assistant turn structure is used. Full templates are provided in Appendix A.3. We refer to these prompt families consistently as PF1, PF2, and PF3 throughout the paper.

3.4 Model Selection (H4)

For clarity and consistency, we refer to models using shorthand identifiers.

We use shorthand identifiers for models: **`M1`** (t5-xxl), **`M2`** (GPT-5-mini), **`M3`** (DeepSeek-R1:32b), **`M4`** (LLaMA-3.1-8b), **`M5`** (Qwen2.5vl:7b), **`M6`** (qwen3-embedding:8b) and **`M7`** (story-emb).

We evaluate four candidate LLMs across all three prompt families:

- **M2 (GPT-5-mini)**[†] - closed model with strong instruction-following and consistent structured JSON output.
- **M3 (DeepSeek-R1:32b)** - open-weight model with explicit chain-of-thought reasoning.
- **M4 (LLaMA-3.1-8b)** - open-weight model; evaluated as a lightweight alternative.
- **M5 (Qwen2.5vl:7b)** - open-weight model; evaluated for narrative comparison capability.

Each model is evaluated on the development set under its best-performing prompt configuration. Models that do not exceed the baseline **`M1`** are eliminated from further consideration. Models that pass the threshold proceed to complementarity analysis rather than being included automatically. Passing the threshold is necessary but not sufficient — beyond accuracy, we ask whether selected models exhibit *complementary* error patterns that make them genuinely useful ensemble partners. This is addressed in Section 3.6.

3.5 Cross-Model Complementarity Analysis (H4)

Rather than selecting models solely based on standalone accuracy, we analyzed agreement patterns between systems. Agreement rate is defined as the proportion of instances for which two systems produce identical predictions. High agreement indicates redundancy, while moderate agreement combined with strong accuracy suggests complementary error patterns. Complementarity, rather than peak individual performance alone, guided final model selection.

3.6 Ensemble Construction

The final system integrates three components:

- Embedding similarity baseline - **`M1`**
- Best GPT-based configuration - **`M2`**
- Best DeepSeek-based configuration - **`M3`**

Predictions are combined using majority voting (Lam and Suen, 1997) across the three systems.

[1] [†]Accessed via Microsoft Azure OpenAI API, snapshot `gpt-5-mini-2025-08-07`. The pipeline is model-agnostic; stronger or more accessible models are expected to yield comparable or improved results.

Majority voting was chosen over weighted alternatives due to the small number of components (three systems) and the absence of reliable confidence scores across heterogeneous model types.

4 Experimental setup

We evaluate on the official SemEval-2026 Task 4 dataset (Hatzel et al., 2026). The development set consists of 200 annotated triples; the test set consists of 400 instances. All development-set experiments report accuracy as the proportion of correctly predicted candidates; final system performance is reported on the test set using the official task metric: ***accuracy***.

5 Results & Analysis

5.1 Baseline

We establish an embedding-based similarity baseline using **M1**. For each triple, we compute cosine similarity between the anchor and each candidate using mean-pooled encoder representations, and select the candidate with the higher similarity score. Using **M1** embedding we get a baseline accuracy of 71.0%.

5.2 Model Configurations and Selection Threshold

We evaluate all three prompt families across four candidate LLMs: **M2** (OpenAI, 2025), **M3** (Guo et al., 2025), M4 (Grattafiori et al., 2024), and **M5** (Yang et al., 2025). Each model is assessed under its best-performing prompt configuration, a combination of prompt family and variant (Table 4), on the development set. Table 2 reports individual model performance and selection outcomes.

Model	Best Prompt Config	Dev Acc.	Decision
M1	Cosine similarity	**71.0%**	**Baseline**
M2	PF3 (guideline + hard, **HARD EXAMPLE**)	77.5%	**Selected**
M3	PF3 (summarized hard examples)	73.7%	**Selected**
M4	PF2 (Aspect-Grounded)	58.5%	*Eliminated*
M5	PF3 (2 hard examples)	67.0%	*Eliminated*

Table 2: Individual model performance on the development set.

M1 baseline (71.0%) represents the best accuracy achievable through embedding-based approaches alone, before any prompt engineering, and serves as our selection threshold. **M4** and **M5** fall below it — both are relatively small models; **M4** likely due to insufficient instruction-following capacity at 8B scale, and **M5** because its vision-language architecture confers no advantage on text-only inputs.

M2 and **M3** exceed the threshold and are carried forward for prompt family evaluation and subsequent complementarity analysis on their best-performing configurations (Sections 5.4 – 5.6).

We evaluate our four hypotheses in order, each corresponding to a phase of the experimental pipeline.

5.3 H1: Embedding Ceiling

Table 3 reports zero-shot development accuracy across the three candidate embedding models.

Embedding Model	Dev Acc.
M1(sentence-transformers/t5-xxl)	71.0%
M6(qwen3-embedding:8b)	63.5%
M7(story_emb)	55.0%

Table 3: Zero-shot development accuracy of candidate embedding models (200 instances).

M1(T5-xxl) achieves 71.0% development accuracy via cosine similarity over mean-pooled representations — a strong starting point that substantially exceeds the organiser Jaccard baseline (56.3%) and the zero-shot GPT-4o-mini prompt (67.0%).

To assess whether supervised adaptation can improve embedding-based similarity, we further explore SimCSE-style contrastive fine-tuning (Gao et al., 2021) on train/validation partitions derived from the development set. The best configuration (lr=10^{-4}, proj. dim=1024, epochs=100) achieves 70.0% on the held-out validation split (100/100 partition), falling below the embedding baseline of 71.0%.

These findings confirm **H1**: dense embedding models capture global semantic similarity effectively but lack the mechanisms needed for explicit reasoning over structured narrative dimensions, motivating the shift to prompt-based approaches.

5.4 H2: Aspect-Grounded Prompting

M2 under **PF1** achieves 70.5% development accuracy, comparable to the **M1** embedding baseline (Table 4), suggesting that unguided LLM prompting alone does not provide a clear advantage. Switching to **PF2** improves accuracy to 72.5%, a gain of 2 points over PF1. This provides evidence for **H2**: grounding prompts in structured narrative dimensions leads to measurable, albeit modest, improvements in LLM reasoning quality over zero-shot prompting.

5.5 H3: K-Shot Sensitivity and Label Priming

Table 4 reports **M2** development accuracy across all prompt configurations explored, spanning **PF1**, **PF2**, and **PF3**. Brief descriptions of each variant are provided in Appendix A.2.

Prompt ID	Prompt Family	Prompt Variant	Dev Acc.
PF1.1	PF1	Baseline (no formal grounding)	70.5%
PF2.1	PF2	Multi-aspect(Theme,Action,Outcome)	69.0%
PF2.2	PF2	Single-aspect(Abstract Theme only)	69.0%
PF2.3	PF2	Procedural instructions	72.5%
PF3.1	PF3	K=1 (guideline example)	71.5%
PF3.2	PF3	K=2 (guideline + hard, **HARD EXAMPLE**)	**77.5%**
PF3.3	PF3	K=2 (guideline + hard, **EXAMPLE**)	74.5%
PF3.4	PF3	K=2 (summarized hard examples)	75.0%
PF3.5	PF3	Procedural Instructions + few-shot	67.0%

Table 4: Development accuracy of **M2** across prompt variants organised as progressively structured prompting approaches.

Carefully constructed **PF3** **(K-shot)** prompts outperform both **PF1** and **PF2**. The strongest configuration (K=2: guideline + hard, **HARD EXAMPLE**) achieves 77.5%.

Demonstration Label	Correct	Dev Acc.
EXAMPLE	149/200	74.5%
HARD EXAMPLE	155/200	77.5%
Difference	+6	**+3 pts**

Table 5: Label priming ablation on **M2**. The two prompts (PF 3.2 & PF 3.3) are identical except for the demonstration label.

A notable sensitivity to prompt framing was observed. Table 5 isolates the effect of *label priming*: the two variants are identical in every respect except the label used to introduce the demonstration — the neutral **EXAMPLE** versus the evaluative **HARD EXAMPLE**. This single-word change results in a 3-point increase in accuracy. This behaviour highlights the sensitivity of in-context learning to demonstration framing (Salinas and Morstatter, 2024; Lester et al., 2021).

These findings support **H3**: generalisation from k-shot demonstrations is sensitive to prompt framing, and demonstration labelling can meaningfully influence performance.

5.6 H4: Complementarity-Driven Ensemble

For each selected LLM, we measure its *agreement rate* with the **M1** embedding baseline. A large gap between a model's standalone accuracy and its agreement rate with **M1** indicates complementary error patterns.

Table 6 reports accuracy and agreement rates for selected models under their best prompt configurations.

Both **M2** and **M3** show a gap of approximately 10 points between standalone accuracy and agreement rate with **M1**.

Model + Prompt ID	Accuracy	Agreement with M1	Difference w/r M1
M2 + PF3.2	77.5%	67.0%	10.5%
M3 + PF3.4	73.7%	61.5%	12.2%

Table 6: Accuracy and agreement rate with the **M1** embedding baseline for selected models under their best prompt configurations.

This confirms **H4**: selecting ensemble components based on complementarity rather than accuracy alone produces a more robust system.

5.7 Overall System Performance

System	Dev Acc.	Test Acc.
Random baseline	50.0%	50.0%
Jaccard similarity	—	56.3%
GPT-4o-mini (zero-shot)	—	67.0%
M1 (T5-xxl)	71.0%	63.0%
M3 (DeepSeek-R1:32b) + PF 3.4	73.7%	69.0%
M2 (GPT-5-mini) + PF 3.2	77.5%	74.0%
Ensemble	**81.0%**	**75.0%**
Human annotator ceiling	—	∼78.0%

Table 7: System performance on development and test sets against organiser baselines and human ceiling from Hatzel et al. (2026).

The three-model majority-vote ensemble achieves 81.0% development accuracy and **75.0% test accuracy** on the 400-instance test set, ranking **3rd out of 47 submitted systems**. The ensemble improves over the strongest individual component (**M2**) by 3.5 points on development and 1.0 point on test, confirming that combining complementary reasoning mechanisms yields consistent gains.

Our result approaches the individual human annotator ceiling of approximately 78% (Hatzel et al., 2026), falling only 3 points below.

5.8 Qualitative Analysis

To understand where and why the system succeeds and fails, we examine four representative test-set instances — one per prediction scenario — across the three task dimensions: Abstract Theme, Course of Action, and Outcomes. Story triples for all four cases are provided in Appendix A.3.

Case	Gold	M1	M2	M3	Ensemble	Key narrative dimension	Failure mode
Case 1	A	Y	Y	Y	Correct	All three dimensions align unambiguously; the theme of desire-driven deception leading to social redemption is decisive.	None — unanimous agreement across all models.
Case 2	B	Y	N	Y	Recovered	Abstract Theme (romantic entanglement generating danger via a false assumption) is overridden by a surface plot cue in M2.	Shared setting of "murder and investigation" misleads M2; M1 and M3 prioritise causal theme and correct via majority vote.
Case 3	A	N	Y	N	Wrong	Generational memory and temporal passage (Abstract Theme) must be distinguished from individual-vs-institution conflict.	M1 and M3 conflate the surface role of "young man with artistic aspirations" with the anchor's deeper structural theme. Majority voting amplifies this co-failure.
Case 4	B	N	N	N	Wrong	The structural relationship between character movement and setting (city as subject vs. backdrop) is the decisive signal.	Strong shared setting (Paris debut) creates surface pull toward the wrong candidate across all models. The correct Abstract Theme signal - city as subject of discovery vs. backdrop to a personal goal — is overridden by this surface overlap, which no model corrects independently

Table 8: Per-case model predictions and ensemble outcome across four representative test-set instances. Y = correct prediction; N = incorrect prediction. Story triples for all cases are provided in Appendix A.3.

Table 8 summarises the per-model predictions, ensemble outcome, the decisive narrative dimension for each case, and the dominant failure mode where applicable.

Surface overlap — shared settings, character roles, or plot elements — is the dominant failure mode across all cases. The ensemble corrects independent model errors (Case 2) but amplifies co-failures when two models share the same incorrect reasoning (Case 3). The hardest instances (Case 4) require reasoning about the structural relationship between character and setting, which current prompting strategies alone cannot resolve. Confidence-based routing could mitigate the co-failure pattern observed in Case 3, where majority voting cannot distinguish a shared error from genuine consensus.

6 Conclusion & Future Work

We describe AI-Monitors, our system for SemEval-2026 Task 4 (Track A), achieving 75% test accuracy & ranking 3rd out of 47 systems. Below we provide the key findings of this work:

- Embedding models provide a strong semantic floor but are limited by the absence of explicit narrative reasoning.
- Aspect-grounded prompting overcomes embedding ceiling, and carefully constructed few-shot demonstrations further improve generalisation — provided that prompt framing remains neutral.
- Ensemble design based on how differently models make errors, rather than accuracy-only selection, produces the most robust system.

For future work, our qualitative analysis points to three concrete directions. First, Case 4 shows that even when the correct signal exists within Abstract Theme, strong surface overlap pulls all models toward the wrong answer — and majority voting cannot help when all three models fail together; better ways to handle this remain an open problem. Second, Case 3 shows that when two models make the same mistake, majority voting makes it worse — in such cases, relying on the most confident single model would give a better result. Third, Cases 2 and 3 together show that models which often disagree with each other tend to catch different errors — giving more weight to such models in the vote, rather than treating all votes equally, would make the ensemble stronger.

Acknowledgements

We sincerely thank the anonymous reviewers for their constructive feedback, and the SemEval-2026 Task 4 organizers for curating a challenging dataset and facilitating this shared task.

References

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.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. 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 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.

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)*, pages 15732–15741.

Louisa Lam and SY Suen. 1997. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(5):553–568.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3045–3059.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and 1 others. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

OpenAI. 2025. GPT-5 mini: A faster, cost-efficient version of GPT-5. https://platform.openai.com/docs/models/gpt-5-mini. Model snapshot: gpt-5-mini-2025-08-07. Accessed via Microsoft Azure OpenAI API.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 298–311.

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.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4629–4651.

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr.

A Appendix - Prompt Details

A.1 Overall Approach

Figure 1 presents an overview of the methodoloy.

A.2 Prompt Used

New additions in each family are highlighted in blue.

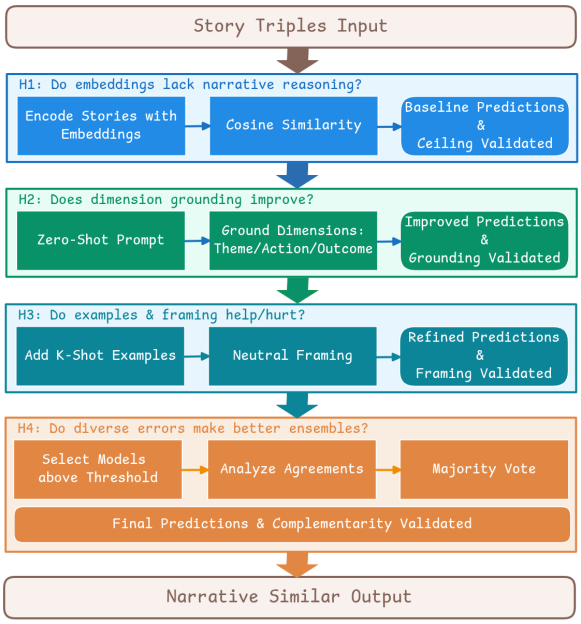

Figure 1: Hypothesis-driven pipeline of the AI-Monitors system, showing progression from embeddings (H1) to ensemble construction (H4).

Prompt Family 1 | Baseline - no formal grounding

`You are an expert at analyzing narrative similarity between stories. Your task is to determine which of two candidate stories (A or B) is more narratively similar to an anchor story.`

Narrative dimensions:

- `Abstract Theme` — central ideas and core motifs; excludes concrete setting.
- `Course of Action` — sequence of events, conflicts, and turning points in order.
- `Outcomes` — final resolution; excludes intermediate statuses.

Ignore: `writing style / setting / character names / time period / text length`

Input: `Anchor: {...} Story A: {...} Story B: {...}`

Output: `Choice: [A or B] | Abstract Theme: [Yes/No] | Course of Action: [Yes/No] | Outcomes: [Yes/No] | Explanation: [2–3 sentences]`

Prompt Family 2 | Aspect-Grounded

`Role, task, ignore list, and input:` same as Family 1.

⊕ **New — Official annotation guideline definitions (Hatzel et al., 2026):**

- `Abstract Theme:` *"The defining constellation of problems, central ideas and core motifs of a story. Does not cover the concrete setting."*
- `Course of Action:` *"The major sequence of events, conflicts, decisions, and turning points, including their order."*
- `Outcome:` *"The final resolution of the story: who succeeds or fails, who survives or dies, what ultimately happens at the end."*

Dimensions are considered independently, then weighed — not simply counted.

⊕ **New — Structured JSON output enforced:**

```
{"choice": "A or B",
"aspects": {"abstract_theme": true/false,
"course_of_action": true/false,
"outcome": true/false},
"explanation": "..."}
```

Prompt Family 3 | K-Shot

`Role, task, dimensions, ignore list, input, and JSON output:` same as Family 2.

⊕ **New — Reference examples added before input:**

Each example isolates one narrative dimension. Shown below is the *Course of Action* example; remaining examples are in Appendix A.3.

Example — Course of Action only:

`Anchor:` Andrew buys food, prepares it at home, and impresses his family with a meal.
`Text A:` Zoie buys guns, prepares defences, and destroys a zombie horde.
`Text B:` Erica builds paper planes and wins a competition with little preparation.

`Decision: A`
`Logic:` Purchasing → preparing → utilizing — a shared course of action regardless of theme or outcome.

Example sources evaluated:

- LLM-generated synthetic examples
- Annotation guideline examples
- Manually selected hard cases from the development set

Label sensitivity finding: Labelling a demonstration as `HARD EXAMPLE` rather than the neutral `EXAMPLE` improves development accuracy by 3 points. This reflects a label priming effect, where evaluative descriptors influence how demonstrations are utilised during in-context learning. This suggests that signalling example difficulty can affect how models interpret and apply in-context examples (Section 5.5).

A.3 Prompt Formulations

To ensure clarity and reproducibility, we define all prompt variants referenced in Table 3. Variants are grouped by prompt family (PF1–PF3) and differ only in how narrative dimensions are presented or

demonstrated.

Baseline (no formal grounding) — PF1. This approach provides only high-level task instructions for narrative comparison without enforcing explicit structure or grounding.

- **Baseline (no formal grounding):** Direct comparison of candidate stories based on overall narrative similarity, without explicit constraints or structured reasoning.

Aspect-grounded — PF2. These variants introduce narrative dimensions (Abstract Theme, Course of Action, Outcome) derived from the annotation guidelines and enforce structured reasoning.

- **Multi-aspect (Theme, Action, Outcome):** Evaluates similarity jointly across multiple narrative dimensions within a single response.
- **Single-aspect (Abstract Theme only):** Restricts evaluation to a single narrative dimension by providing only the Abstract Theme definition, isolating its contribution to narrative similarity.
- **Procedural instructions:** A single-prompt variant that provides step-by-step instructions to guide reasoning using formally defined narrative dimensions.

K-shot — PF3. These variants extend grounded prompting by incorporating curated examples demonstrating how narrative similarity should be evaluated.

- $K{=}1$ **(guideline example):** A single example drawn from the annotation guidelines, illustrating the expected reasoning process and output format.
- $K{=}2$ **(guideline + hard, HARD EXAMPLE):** Combines one example from the annotation guidelines with a challenging case selected from the development set after error analysis. The challenging instance is introduced with the evaluative label `HARD EXAMPLE`, highlighting its difficulty.
- $K{=}2$ **(guideline + hard, EXAMPLE):** Identical to the previous variant, except that the challenging instance is introduced with the neutral label `EXAMPLE`. This isolates the effect of demonstration labelling (label priming) on model behaviour.
- $K{=}2$ **(summarized hard examples):** Uses summarized versions of the stories in the k-shot examples, reducing narrative detail while preserving key distinctions for comparison.
- **Procedural + few-shot:** Combines structured reasoning instructions with example-based demonstrations in a single prompt.

All variants are evaluated under the same dataset and evaluation protocol, ensuring comparability across different prompting strategies.

Appendix B Story Triples — Qualitative Analysis

The four story triples below correspond to the cases examined in Section 5.8 (Table 8). Each triple consists of an anchor and two candidate stories (A and B); the gold label is indicated.

Case 1 — Both correct (M2 ✓, Ensemble ✓)

Anchor: In 1911, Vincenzo Peruggia, a poverty-stricken Italian glazier, steals the Mona Lisa from the Louvre to impress a woman he loves. When she proves fickle, he confesses and is arrested, but publicly reframes his motive as patriotism and is celebrated as a national hero.
Story A (Gold): Fay Cheyney, posing as a wealthy widow, attempts to steal a pearl necklace. She is caught, confesses voluntarily, and — through a turn of social leverage — is ultimately accepted back into the social circle rather than punished.
Story B: Lieutenant Hofmiller develops compassion for a paralysed woman, privately promises to marry her, then publicly denies the engagement out of fear of ridicule. She takes her own life; he is consumed by guilt.

Analysis. The anchor and Story A both follow a pattern of *theft* $\rightarrow$ *discovery* $\rightarrow$ *confession* $\rightarrow$ *reframing*,sharing the Abstract Theme of desire-driven deception leading to unexpected social redemption. Story B involves concealment but ends in death and guilt — diverging on both Course of Action and Outcome. All three models correctly select Story A; the alignment across all three dimensions is unambiguous and leaves no room for surface-level confusion.

Case 2 — Ensemble correct, M2 wrong (M2 ×, Ensemble ✓)

Anchor: New York City janitor Daryll Deever becomes entangled in a murder investigation because reporter Toni Sokolow believes he has information. He allows her pursuit because he is romantically interested in her. Their cat-and-mouse dynamic convinces the real killers that Daryll genuinely knows

something, placing both in danger due to a false assumption.

Story A: A Sicilian truck driver is murdered near a construction site. Police captain Bellodi investigates whether the motive was corruption or a romantic affair. Witnesses lie to protect the local Mafia don; Bellodi resorts to unorthodox methods but remains unable to resolve the case conclusively.

Story B (Gold): Dédée, a woman controlled by her pimp Marco, falls for Francesco, an Italian ship captain who wants to take her away. Marco murders Francesco out of jealousy. Dédée and bar owner René track Marco down and kill him in retribution.

Analysis. Gold is B. `M2` selects A; `M3` and `M1` both select B, and the ensemble recovers through majority vote. The anchor's theme is not investigation but romantic entanglement that generates danger through a false assumption — a pattern Story B mirrors precisely, where jealousy triggers violence and danger arises from proximity rather than guilt. `M2` is misled by the surface overlap of "murder and investigation" shared with Story A, overriding the deeper thematic and causal alignment with B. The ensemble compensates because `M3` and `M1` independently prioritise the correct theme, overruling `M2`'s misdirection.

Case 3 — M2 correct, Ensemble wrong (M2 ✓, Ensemble ×)

Anchor: Federico Fellini's film recounts his youth in Rome across two historical eras (1930s and 1970s), moving through episodic scenes — a guesthouse, brothels, a vaudeville theatre, excavated catacombs — as a portrait of a city and a generation. It ends on a melancholic note with the death of actress Anna Magnani, filmed by chance.

Story A (Gold): A bourgeois Italian family is seen across multiple generations through the memoirs of Carlo, an elderly retired professor. The film spans from the Belle Époque to the 1980s, centred on the family's apartment and its dynastic traditions.

Story B: Emrah, an aspiring young director in Turkey, fights bureaucracy and censorship authorities to shoot his first film, ultimately confronting a petty official who enforces a senseless law.

Analysis. Gold is A. `M2` correctly selects A; both `M3` and `M1` select B, and the majority vote overrules `M2` — the ensemble is wrong. The anchor's Abstract Theme is generational memory and the passage of time, reflected in Story A's multi-decade family memoir rooted in a single location. Story B centres on an individual's struggle against institutional authority — a fundamentally different theme. `M3` and `M1` conflate the surface role of "young man with artistic aspirations" present in both the anchor and Story B with the deeper structural theme of temporal-generational reflection. This case exposes a key vulnerability of majority voting: when two models share the same incorrect reasoning pattern, their agreement amplifies the error. The ensemble cannot correct systematic co-failures — a limitation that confidence-based routing could address.

Case 4 — Both wrong (M2 ×, Ensemble ×)

Anchor: A young woman in Paris's Bastille neighbourhood loses her cat. The pursuit of the cat becomes a lens through which the film explores the contrasts of a changing urban district — old residents and new — revealing the dynamics of a village within a large city.

Story A: Arlette, a country girl and illegitimate daughter of a government minister, arrives in Paris for the first time hoping to build a music career. She evades a minder sent by her father, falls in with bohemians in Montmartre, gets arrested while busking, and ultimately falls in love. Her father eventually acknowledges her.

Story B (Gold): Bernie Noël, a 29-year-old man raised entirely in a suburban orphanage outside Paris, ventures into the city for the first time as an adult. Knowing the world only through television, he navigates a hostile nocturnal Paris while searching for his biological parents, eventually finding them and constructing an imaginary narrative around them.

Analysis. Gold is B; all three models predict A, and the ensemble follows. Both Story A and Story B involve an outsider navigating Paris for the first time, creating a strong surface pull. However, the anchor's Abstract Theme is how a city reveals itself through one person's movement and curiosity — a structure Story B mirrors faithfully through Bernie's nocturnal discovery of Paris as subject. Story A is goal-directed (career, paternal recognition) and treats Paris as backdrop, not subject. The failure reflects a genuine challenge: when the distinguishing signal lies in the structural relationship between character movement and setting rather than in plot events, current prompting approaches are insufficient.