Team CV at SemEval-2026 Task 4: Prompting LLMs and Benchmarking Embedding Models for Narrative Story Similarity

Chandan Kumar R S
Mysore University School of Engineering
Karnataka, India
chandankumarrs683@gmail.com

Vinay Babu Ulli
Oogwai Analytics
Karnataka, India
ullivinaybabu@gmail.com

Abstract

This paper describes Team CV's systems for SemEval-2026 Task 4: Narrative Story Similarity and Narrative Representation Learning (Hatzel et al., 2026). For Track A (comparative judgment), we explore five prompting strategies—zero-shot, chain-of-thought, structured feature extraction, pairwise scoring, and few-shot—and QLoRA fine-tuning of smaller models. For Track B (narrative embeddings), we benchmark twelve dedicated text embedding models of varying dimensionality (384–4096) spanning open-source (E5-Large-v2, BGE, GTE, Qwen3 Embedding) and closed-source (OpenAI, Gemini, Mistral) families, and fine-tune Qwen3 Embedding 4B on task-specific triples. Few-shot prompting with Qwen-2.5 7B (64.00%) outperforms all fine-tuned variants (best 57.50%) on Track A; scaling to LLaMA-3.3-70B yields 75.00%. On Track B, OpenAI `text-embedding-3-large` (3072-d) achieves the best dev accuracy (67.00%), while fine-tuning Qwen3 Embedding 4B (2560-d) on synthetic triples slightly decreases accuracy. Our final submission—LLaMA-3.3-70B (3-shot) for Track A and `text-embedding-3-large` for Track B—achieves 70.75% and 64.50%, exceeding the GPT-4o-mini and STORY-EMB baselines respectively.

1 Introduction

Narrative similarity requires understanding deeper structural elements—themes, event sequences, and outcomes—while disregarding surface details like character names and settings (Hatzel et al., 2026). SemEval-2026 Task 4 operationalises this across two tracks. Track A frames it as a binary choice: given an anchor story and two candidates, predict which candidate is more narratively similar. Track B requires systems to produce per-story embeddings such that narratively similar stories are closer in cosine distance.

We address both tracks: for Track A we systematically compare five prompting strategies and QLoRA fine-tuning (Dettmers et al., 2023) of smaller open-source models; for Track B we benchmark twelve dedicated text embedding models across a wide range of dimensionalities (384 to 4096) and fine-tune one (Qwen3 Embedding 4B, 2560-d) on task-specific triplets. Our key findings are: (1) few-shot prompting outperforms both elaborate reasoning strategies and supervised fine-tuning on Track A; (2) high-capacity, high-dimensional embedding models like `text-embedding-3-large` (3072-d) and E5-Large-v2 (Wang et al., 2022) (1024-d) transfer well to narrative similarity on Track B; and (3) model scale and embedding dimensionality are important factors, though not strictly monotonic. All fine-tuned models and embeddings on the devset are accessible via our HuggingFace collection.[1]

2 Background

2.1 Task Description

The task defines narrative similarity through three components: *Abstract Theme* (ideas, motives, messages), *Course of Action* (central events and turning points), and *Outcomes* (resolutions and consequences) (Hatzel et al., 2026; Hatzel and Biemann, 2024a). Stories are Wikipedia plot summaries (4–8 sentences) drawn from the Tell-Me-Again Corpus (Hatzel and Biemann, 2024a). Triples were sampled via STORY-EMB (Hatzel and Biemann, 2024a) with rejection sampling retaining only cases where two LLMs disagree, yielding genuinely difficult instances. The estimated human upper bound is 89% (Hatzel et al., 2026). Track B uses the same annotations: for each triple, the system is correct if the embedding of the closer candidate has higher cosine similarity to the anchor than the farther can-

[1] https://hf.co/collections/Chandan683/narrative-similarity-task

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2981–2985

didate.

2.2 Dataset

The organisers provide a synthetic split of 1,900 LLM-generated triples as training data, a 200-triple dev set, and a 400-triple test set. Track A and Track B share the same underlying stories and annotations for all non-test splits.

2.3 Related Work

Computational narrative similarity was explored by Fisseni and Löwe (2012) for folk tales and by Chaturvedi et al. (2018) for movie remakes. Chen et al. (2022) extended similarity to multilingual news (SemEval-2022 Task 8). Hatzel and Biemann (2024a) proposed STORY-EMB, a narrative-focused embedding model trained on story retellings (Hatzel and Biemann, 2024b), which serves as the Track B baseline. General-purpose text embedding models—E5 (Wang et al., 2022), BGE (BAAI Research Team, 2024), GTE (Li et al., 2023), Sentence-BERT (Reimers and Gurevych, 2019)—have shown strong performance on semantic similarity benchmarks but have not been systematically evaluated on narrative similarity. On the prompting side, in-context learning (Brown et al., 2020) and chain-of-thought prompting (Wei et al., 2022) motivate our Track A approach.

3 System Overview

3.1 Track A: Prompting Strategies

We test five strategies using Qwen-2.5 7B as a controlled baseline:

S1 – Zero-Shot. The system prompt defines the three components of narrative similarity; the model outputs `A` or `B`.

S2 – Chain-of-Thought (CoT). Following Wei et al. (2022), the model analyses each story's theme, events, and outcomes before answering.

S3 – Structured Feature Extraction. A two-stage pipeline: three LLM calls extract `THEME/EVENTS/OUTCOME` per story; a fourth compares the features.

S4 – Pairwise Scoring. Each anchor–candidate pair is scored 1–10 independently; the higher-scored candidate wins.

S5 – Few-Shot (3-shot). Three labelled examples from the sample split are embedded in the system prompt for label balance and narrative diversity.

3.2 Track A: QLoRA Fine-Tuning

We fine-tune six models with QLoRA (Dettmers et al., 2023) (4-bit NF4, rank 16, alpha 32, dropout 0.05) on 1,897 synthetic triples for 3 epochs. Models: Gemma-3 1B/4B/12B, Qwen-3 4B, Qwen-2.5 7B (zero-shot and few-shot prompt formats).

3.3 Track B: Embedding Models

We evaluate twelve dedicated text embedding models as black-box encoders for Track B. Each model maps a story to a d-dimensional vector; we predict the candidate with higher cosine similarity to the anchor. The models span a wide range of embedding dimensionalities, from 384 to 4096:

Closed-source: OpenAI `text-embedding-3-large` (3072-d) and `text-embedding-3-small` (1536-d), Gemini Embedding 001 (3072-d), Mistral Embed 2312 (1024-d).

Open-source: E5-Large-v2 (Wang et al., 2022) (1024-d), BGE-Large-EN-v1.5 (BAAI Research Team, 2024) and BGE-M3 (1024-d), GTE-Large (Li et al., 2023) (1024-d), Qwen3 Embedding 4B (2560-d) and 8B (4096-d), and a small Sentence-BERT baseline `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019) (384-d).

Fine-tuned: Qwen3 Embedding 4B (2560-d) fine-tuned on the synthetic triples using a cosine-similarity objective.

4 Experimental Setup

Track A. All prompting uses temperature 0.0 and deterministic decoding. Max output tokens: 16 (zero-shot/few-shot), 2048 (CoT), 512/1024 (structured), 512 (pairwise). For Qwen-3 models (thinking mode), we set 4096 tokens and strip `<think>` blocks. Fine-tuning uses batch size 16 (4×4 gradient accumulation), max length 2048, learning rate $1e{-}4$ (Qwen) / $2e{-}4$ (Gemma), cosine schedule. Unparseable responses default to `A`.

Track B. Each story is embedded independently using the model's default pooling strategy. All embeddings are L2-normalised before cosine computation. Given a triple (a, c_1, c_2) with gold label y, we predict the candidate with higher cosine similarity to the anchor as the closer story.

Evaluation. Both tracks use accuracy (proportion of triples correctly classified) on the 200-triple dev set and 400-triple official test set.

Category	System	Acc
Prompting (Qwen-2.5 7B)		
	S1: Zero-Shot	51.00
	S2: CoT	61.50
	S3: Structured	55.00
	S4: Pairwise	62.00
	S5: Few-Shot	**64.00**
QLoRA Fine-Tuned		
	Gemma-3 1B	43.50
	Gemma-3 4B	52.00
	Gemma-3 12B	51.00
	Qwen-3 4B	42.00
	Qwen-2.5 7B (ZS)	55.00
	Qwen-2.5 7B (FS)	57.50

Table 1: Track A dev accuracy (%, 200 triples).

Model	Acc
Qwen-2.5 7B	64.00
Qwen-2.5 72B	73.50
Qwen-3 32B	69.50
Qwen-3 80B MoE	60.00
Qwen-3 235B MoE	66.50
LLaMA-3.1 70B	71.50
LLaMA-3.3 70B	**75.00**
Gemini 2.0 Flash	68.50
GPT-4o-mini	71.50
GPT-5.1	73.50

Table 2: Few-shot (3-shot) prompting across ten LLMs on Track A dev (%).

5 Results

5.1 Track A: Strategy Comparison

Table 1 shows that few-shot prompting is the best strategy (64.00%), outperforming CoT by 2.50% and all fine-tuned models by $\geq$6.50%. Zero-shot is near chance (51.00%), confirming that instructions alone are insufficient. Fine-tuned models underperform mainly due to the distributional mismatch between synthetic training data and human-annotated dev triples.

5.2 Track A: Model Scaling

Model scale and architecture both matter: LLaMA-3.3-70B achieves the best dev accuracy (75.00%), and is therefore chosen as our final Track A system.

5.3 Track B: Embedding Model Comparison

Table 3 presents the Track B results. The top three models—`text-embedding-3-large` (3072-d, 67.00%), E5-Large-v2 (1024-d, 66.50%), and Gemini Embedding 001 (3072-d, 66.00%)—form a strong tier, all exceeding the STORY-EMB baseline (63.25%). Open-source models such as Qwen3 Embedding 4B (2560-d, 64.50%) and BGE-Large-v1.5 (1024-d, 63.50%) are competitive but trail the best closed-source models by 2–4%.

Model	Dim	Acc
`text-emb-3-large`	3072	**67.00**
E5-Large-v2	1024	66.50
Gemini Emb 001	3072	66.00
Qwen3 Emb 4B	2560	64.50
Mistral Emb 2312	1024	64.00
Qwen3 Emb 4B (FT)	2560	63.50
BGE-Large-v1.5	1024	63.50
GTE-Large	1024	61.00
Qwen3 Emb 8B	4096	60.00
BGE-M3	1024	59.00
`text-emb-3-small`	1536	58.50
all-MiniLM-L6-v2	384	43.50

Table 3: Track B dev accuracy (%, 200 triples) across twelve embedding models, with embedding dimensionality.

	Track A	Track B
GPT-4o-mini baseline	67.00	—
STORY-EMB baseline	—	63.25
Team CV	**70.75**	**64.50**

Table 4: Official test accuracy (%).

Embedding dimensionality does not strictly predict performance. E5-Large-v2 achieves 66.50% with only 1024 dimensions, outperforming Qwen3 Embedding 8B (4096-d, 60.00%) and `text-embedding-3-small` (1536-d, 58.50%). Conversely, the two highest-dimensional models in the top tier—`text-embedding-3-large` and Gemini Embedding 001—both use 3072 dimensions. This suggests that while higher dimensionality provides more representational capacity, the quality of pre-training and the contrastive objective used during training matter more than raw dimensionality for capturing narrative similarity.

Fine-tuning Qwen3 Embedding 4B on the synthetic triples *decreases* accuracy by 1.0% (64.50%→63.50%), mirroring the Track A finding that synthetic data introduces distributional mismatch with human annotations. The small SBERT baseline `all-MiniLM-L6-v2` (384-d, 43.50%) performs well below random, confirming that both sufficient dimensionality and strong pre-training are necessary for this task.

5.4 Official Test Results

Table 4 shows our final results. Track A (70.75%) exceeds the GPT-4o-mini baseline by 3.75%. Track B (64.50%) exceeds the STORY-EMB base-

line by 1.25%, demonstrating that general-purpose embedding models can match or surpass a purpose-built narrative encoder.

5.5 Error Analysis

We analysed the 50 Track A errors by LLaMA-3.3-70B on the dev set (31 false negatives, 19 false positives). Three systematic failure modes emerge: (1) *surface keyword matching overrides structure*—e.g., medical keyword overlap (agnosia vs. amnesia) masks a shared exploitation-of-vulnerability narrative; (2) *shared genre masks divergent outcomes*—e.g., two opera-setting stories chosen despite opposite endings; (3) *genre mismatch blocks thematic recognition*—e.g., a zombie comedy's musicians-triumphing-over-adversity arc rejected in favour of an unrelated domestic romance. These failures reveal systematic over-reliance on surface-level topical overlap at the expense of abstract narrative structure.

6 Conclusion

We presented Team CV's system for SemEval-2026 Task 4, addressing both Track A and Track B. Few-shot in-context learning with large LLMs is the most effective paradigm for comparative narrative judgment (Track A), while high-capacity text embedding models provide strong narrative representations without task-specific training (Track B). Model scale is the dominant factor on Track A, while on Track B we find that pre-training quality and embedding dimensionality jointly determine performance—though dimensionality alone is not sufficient, as demonstrated by E5-Large-v2 (1024-d) outperforming models with 2–4$\times$ more dimensions. All systems remain well below the 89% human ceiling, with persistent weaknesses in abstracting past surface features.

7 Limitations and Future Work

Our approach has several limitations. For Track A, the few-shot examples are drawn from a small sample split and performance may be sensitive to example selection. QLoRA fine-tuning experiments are limited by the synthetic training data and single-GPU compute. For Track B, we evaluate only off-the-shelf embedding models and one fine-tuned variant; we do not explore task-specific contrastive learning objectives (e.g., SimCSE-style; Gao et al., 2021) on human-annotated data or narrative-structural features. The best systems require access to large proprietary models (LLaMA-3.3-70B, `text-embedding-3-large`), limiting deployment in resource-constrained settings.

Future directions include: (1) ensemble methods combining multiple embedding models and/or LLM predictions; (2) fine-tuning embedding models on human-annotated narrative similarity data rather than synthetic triples; (3) incorporating narrative-theory features (e.g., Proppian functions, event schemas) into prompts or embedding pipelines; and (4) hybrid systems that combine LLM-based reasoning with dedicated narrative embedding models such as STORY-EMB (Hatzel and Biemann, 2024a).

8 Ethics Statement

The data consists of publicly available Wikipedia plot summaries. All APIs were used in accordance with their terms of service. We acknowledge that embedding and LLM-based systems may reflect biases in their training data.

References

BAAI Research Team. 2024. Bge-large-en-v1.5: BAAI general embedding model. https://huggingface.co/BAAI/bge-large-en-v1.5. Accessed 2026-03-03.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901.

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 A. 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)*. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural In-*

formation Processing Systems 36 (NeurIPS 2023). ArXiv:2305.14314.

Bernhard Fisseni and Benedikt Löwe. 2012. Which dimensions of narratives are relevant for human judgments of story equivalence? In *Proceedings of the Workshop on Computational Models of Narrative (CMN 2012)*.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Hans Ole Hatzel and Chris Biemann. 2024a. Story embeddings – narrative-focused representations of fictional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

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

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Fei Xia, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.