TFB at SemEval-2026 Task 4: Diagnosing Model Failures in Narrative Understanding

Anna Colli[1]*, Benedictus Kent Rachmat[2,3]*, Eve Sauvage[2,4]*,
Delphine Battistelli[1], Thomas Gérald[2], Cyril Grouin[2], Julien Tourille[4], Zheng Zhang[3]

[1]Université Paris Nanterre, CNRS, Modyco, France
[2]Université Paris-Saclay, CNRS, LISN, Orsay, France
[3]Embedded AI Lab, SLB, Clamart, France
[4]EDF, France

acolli@parisnanterre.fr, rachmat@lisn.fr, eve.sauvage@lisn.fr

Abstract

We describe the participation of team TFB in SemEval-2026 Task 4 on narrative similarity. We explore ColBERT-inspired sentence-level late interaction to capture event reordering, compare fine-tuning with synthetic data at multiple difficulty tiers, finding that distribution proximity to the target data matters more than volume and evaluate chain-of-thought prompting. We complement our approaches with a human annotation study (Krippendorff's $\alpha = 0.32$) confirming the task's inherent difficulty, an analysis of synthetic data distribution shift explaining why fine-tuning on out-of-distribution data hurts the model's performance. Despite our experiments, we didn't surpass results of SENTENCE-T5-XXL on Track B and Qwen2.5-7B-Instruct on Track A. We finally decided to submit these two models for the task.

1 Introduction

This paper presents our participation to Track A (classification) and Track B (embedding) of SemEval-2026 Task 4 on Narrative Textual Similarity (Hatzel et al., 2026). The proposed task aim to find, given three stories (A, B and anchor), the most similar story between A and B to anchor. Narrative similarity differs here fundamentally from standard textual similarity. While sentence-level similarity benchmarks (Reimers and Gurevych, 2019; Muennighoff et al., 2023) evaluate whether two sentences express the same meaning, narrative similarity requires comparing higher-level structures: abstract themes, courses of action, and outcomes (Hatzel et al., 2026). All texts are in English.

For Track A, we submit QWEN2.5-7B (fine-tuned with LORA) and for Track B, we submit SENTENCE-T5-XXL (off-the-shelf). Beyond our submissions, for Track B, we study the late interaction mechanism of COLBERT (Khattab and Zaharia, 2020) adapted from token-level to sentence-level. and we compare fine-tuning with synthetic data at multiple difficulty tiers. For Track A, we evaluate different prompting strategies and fine-tuning with LoRA across several LLMs.

Our contributions are:

1. Submitted systems QWEN2.5-7B (fine-tuned with LORA) (Track A) and SENTENCE-T5-XXL (Track B), with additional analysis of ColBERT-inspired sentence-level max-similarity and difficulty-tiered synthetic fine-tuning;

2. An annotation study quantifying task difficulty and systematic LLM failure modes including degenerate class-prediction behavior;

3. An analysis of synthetic data quality showing that distribution proximity to target data, and not the volume, determines fine-tuning effectiveness.

2 Related Work

Modern text similarity spans multiple granularities. Lexical methods such as TF-IDF (Sparck Jones, 1988) and BM25 (Robertson and Zaragoza, 2009) measure term overlap. Learned sparse methods like SPLADE (Formal et al., 2021) extend lexical matching with contextualized term expansion. Dense approaches, notably SENTENCE-BERT (Reimers and Gurevych, 2019), produce fixed-size embeddings evaluated on benchmarks such as MTEB (Muennighoff et al., 2023). Finally, late interaction models like COLBERT (Khattab and Zaharia, 2020) retain per-token representations and compute fine-grained similarity at query time. Closer to our task, Hatzel and Biemann (2024) introduce Story Embeddings with entity pseudonymization, achieving 94.6% precision on narrative retrieval by preventing models from relying on character-name overlap. Recent work has also shown that narrative coherence requires

*Equal contribution.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2932–2938

Dataset	n_{triples}
Organizer synthetic	1,900
TFB-syn easy	1,200
TFB-syn medium	1,200
TFB-syn hard	1,200
TFB-syn hard-fewshot	600

Table 1: Number of triples per synthetic dataset. TFB-syn denotes our generated data at four difficulty levels.

modeling inter-event dependencies beyond atomic facts (Zheng et al., 2025; Castricato et al., 2021). In fact, classical narrative theory distinguishes the *fabula* the chronological sequence of events from the *syuzhet* the arrangement of those events as presented in the text (Bal and van Boheemen, 2009; Schmid, 2010; Onega and Landa, 2014). This distinction matters for similarity: two narratives may present the same events in different orders, so representations must capture the underlying event structure rather than surface ordering.

3 Methodology

3.1 Task and Data

SemEval-2026 Task 4 (Hatzel et al., 2026) defines two tracks. **Track A**: given a triple (anchor, text_a, text_b), predict which text is more narratively similar to the anchor (binary classification). **Track B**: produce a vector representation per story such that cosine similarity reflects narrative similarity. The dataset consists of manually annotated triples with a reported Krippendorff's $\alpha = 0.33$ (Hatzel et al., 2026) on two annotators. A synthetic dataset of 1,900 triples was also provided by the organizers.

3.2 Annotation Study

To assess task difficulty independently, we re-annotated a subsample of 50 triples from the Track A development set. Three non-English native annotators provided binary classifications with a confidence score on a 1–5 Likert scale. We compute inter-annotator agreement using Krippendorff's α and we additionally annotate the same triples with five LLMs to compare human and machine performance.

We use the following LLMs: LLAMA-3.1-8B-INSTRUCT[1], LLAMA-3.1-70B-INSTRUCT[2], QWEN-2.5-7B-INSTRUCT[3], GEMMA-3-12B-INSTRUCT[4] and GPT-4O[5]

3.3 Synthetic Data Quality Analysis and Generation

To assess synthetic data quality, we analyse the organizers' dataset along two axes: (1) the distributional overlap with the development set in embedding space and (2) the diversity of sentence-order permutations in the generated texts relative to the source narratives. In addition to the organizers' synthetic dataset, we generate four new synthetic datasets, detailed in Table 1, using GPT-4o (cf. Appendix B). Each prompt enforces a shared set of narrative requirements (thematic, structural, stylistic, and logical alignment) between the anchor, A, and B. The tiers differ in the level of discriminability between A and B. In **TFB-syn easy**, A is strongly aligned with the anchor, while B is clearly weaker. In **TFB-syn medium**, there is mild ambiguity between A and the anchor, with moderate misalignments in focus or pacing. In **TFB-syn hard**, both candidates are strong matches, with A only slightly closer due to subtle thematic or causal factors. **TFB-syn hard-fewshot** is identical to *hard*, but with a real development-set example prepended to calibrate the difficulty.

3.4 Track A

3.4.1 LLM Prompting and Finetuning

To solve Track A, we first experimented with different prompting strategies. Specifically, each model (same model list as presented in 3.2) receives a classification prompt presenting the anchor and both candidates, asking it to respond with a single letter (*A* or *B*). We evaluate two prompt variants: (1) a simple expert persona (PROMPT-1, cf. Appendix A); and (2) a minimal instruction with no persona (PROMPT-2, cf. Appendix A). Zero-shot results uses both prompts; few-shot results use PROMPT-2. Second, we fine-tuned[6] QWEN2.5-7B and LLAMA3.1-8B using Low-Rank Adaptation (LoRA; rank=16, α=32) on the TFB-syn hard-fewshot dataset.

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[2] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[3] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[4] https://huggingface.co/google/gemma-3-12b-it

[5] API call

[6] A classification head is trained with cross-entropy loss for three epochs (batch size 4, learning rate 2×10^{-5}). All models use 8-bit quantization; inference is deterministic (temperature 0)

3.5 Track B

3.5.1 Model testing

For Track B, we first tested different off-the-shelf models for sentence similarity : ALL-MINILM-L6-V2[7], MSMARCO-DISTILBERT[8], MULTILINGUAL-E5[9] and SENTENCE-T5-XXL[10]. We wanted models to cover different sizes, architectures, and training objectives.

3.5.2 Sentence-Level Max-Similarity

Since the *fabula* and the *syuzhet* may differ in event ordering, we tested a sentence-level adaptation of COLBERT's late interaction mechanism (Khattab and Zaharia, 2020). COLBERT encodes queries and documents independently into sequences of token-level embeddings, and computes similarity at match time via a MaxSim operator.

For each candidate document, we compute pairwise cosine similarities between all query sentences' embeddings and all document sentences' embeddings, forming a similarity matrix. Following the canonical MaxSim aggregation, for each query sentence we retain only its highest similarity with any document sentence, and fuse these maxima with a sum to produce a global similarity score. We additionally compare two alternative fusion strategies (**mean** and **max** of the retained similarities) and three sentence representations: **first token**, **last token**, and **mean** of all tokens within each sentence.

3.5.3 Fine-Tuning

Finally, to evaluate whether the quality of synthetic data affects model performance, we fine-tune[11] the models tested in Section 3.5.1—excluding SENTENCET5-XXL due to computational limitations—on both our TFB-syn and the dataset provided by the task organizers.

4 Results

4.1 Annotation Study

Similarly to organisers results, we obtain a mean Krippendorff's α of 0.32 with α scores ranging from 0.12 to 0.48 (cf. Figure 1). This low agreement shows the complexity of the given task due to its subjectivity.

	Annot. 1	Annot. 2	Annot. 3
Accuracy	64%	70%	66%

Table 2: Individual annotator accuracy on the 50-triple subset.

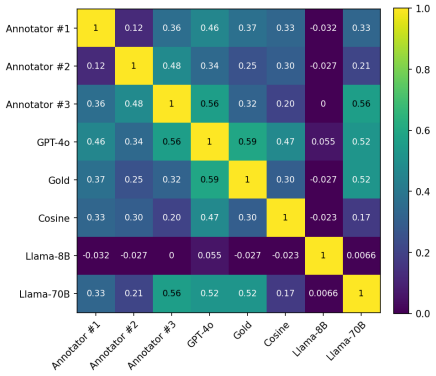


Figure 1: Inter-annotator agreement (Fleiss' κ) across humans, LLMs, gold labels, and cosine baseline on the 50-triple subset. Mean human $\kappa = 0.32$. GPT-4o shows the highest agreement with human annotators, while Llama-8B exhibits near-zero agreement with all other annotators.

Table 2 shows individual annotator accuracy against the gold labels. All three annotators agree with the gold label on only 21 texts (42%), confirming the task's high subjectivity. Additionally, confidence scores predict agreement quality: annotators are wrong in only 9% of cases when their confidence exceeds 3, but triples with confidence ≥ 4 represent 57% of the subset, indicating that many triples are perceived as unambiguous despite low overall inter-annotator agreement. We evaluate five LLMs' accuracy on the same 50 triples (results for the IAA including LLMs in Figure 1): GPT-4O (62%), LLAMA3.1-8B (52%), QWEN2.5-7B (48%), and LLAMA3.1-70B (46%). Mean LLM accuracy (54%) is significantly lower than mean human accuracy (67%; permutation test $p = 0.037$). Among LLMs, accuracy decreases with model size on this subset: LLAMA3.1-8B (52%) > LLAMA3.1-70B (46%). This ordering is consistent with larger models having memorized stronger entity-association priors from pre-training, which may interfere when narrative similarity explicitly requires ignoring surface-level entity overlap.

[7] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[8] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v2

[9] https://huggingface.co/intfloat/multilingual-e5-large

[10] https://huggingface.co/sentence-transformers/sentence-t5-xxl

[11] For all models we use the BatchHard variant of the triplet loss

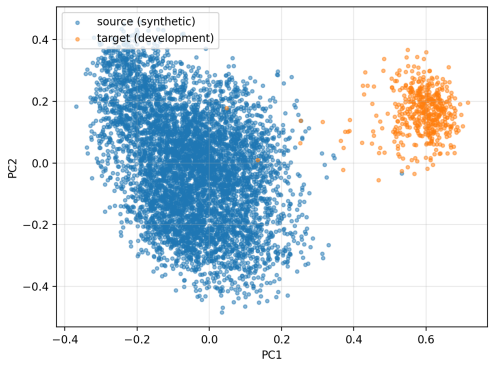
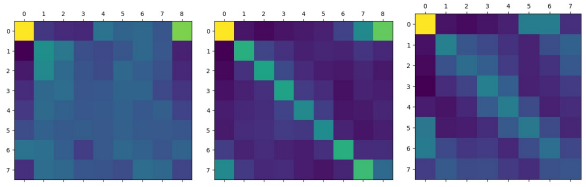


Figure 2: 2D PCA of Qwen3-Embedding-4B vectors: synthetic (blue) and development (orange) occupy disjoint regions, revealing a clear domain gap.



(a) Dev sentence overlap (b) Synthetic sentence overlap (c) TFB-syn hard-fewshot

Figure 3: Sentence-order similarity matrices computed with multilingual-e5 embeddings.

4.2 Synthetic Data Quality and Generation

Figure 2 shows that the synthetic data provided by the organizers (blue) has minimal overlap with the development distribution (orange). This distribution shift means that models fine-tuned on synthetic data are not adapted to the target distribution.

Figure 3 further reveals that the provided synthetic narratives (c) largely preserve the sentence order of the source text: the similarity matrices between anchor and generated texts show strong diagonal patterns. In contrast, dev data (a) shows diffuse similarity patterns and our TFB-syn hard-fewshot data (c) shows intermediate patterns. We also evaluate the entity density gap in order to provide a measurable account of the distribution shift. We note that dev and test data contain 2.5–11× more named entities per 100 tokens than any synthetic tier (cf. Appendix, table 8). This likely reflects their Wikipedia origin, which features dense proper-noun usage, whereas the synthetic data uses more generic language. Consequently, models are trained in a low entity-density regime but evaluated on structurally different data.

Model	Method	Acc.	F1
GPT-4o	Prompt-1	50.5	0.65
QWEN2.5-7B	**LoRA (P1)**	**68.5**	0.71
	Prompt-2	67	0.70
	Prompt-1	65.5	0.70
	few-shot (P2)	61.5	0.71
LLAMA3.1-8B	LoRA (P1)	62.5	0.59
	Prompt-1	60.5	0.69
	Prompt-2	60.5	0.69
	few-shot (P2)	53.5	0.67
LLAMA3.1-70B	Prompt-1	57.0	0.68
	Prompt-2	57.0	0.68
	few-shot (P2)	51.0	0.67
GEMMA3-12B	Prompt-1	52.0	0.66
	Prompt-2	52.0	0.66
	few-shot (P2)	51.0	0.65

Table 3: Track A results on the development set (200 triples). Best accuracy in **bold**. Submitted model is bolded in red.

	Acc.on dev	Acc.on test
ALL-MINILM-L6-V2	55.0	59.75
MSMARCO-DISTILBERT	62.0	59.75
MULTILINGUAL-E5	60.0	61.75
SENTENCET5-XXL	**69.0**	61.50

Table 4: Off-the-shelf model testing results. Best accuracy in **bold**. Submitted model is bolded in red.

4.3 Track A

4.3.1 LLM Prompting and Finetuning Results

Table 3 presents results for all LLM configurations on the full development set (200 triples). Overall, the highest accuracy is achieved by QWEN2.5-7B fine-tuned with LoRA on TFB-syn hard-fewshot, reaching 65.5% (+3pp over its zero-shot performance). GPT-4O collapses on the full development set (50.5%), consistent with constant-class prediction, but achieves 62% on the 50-triple subset (§4.1), indicating that degeneration emerges only at larger scale. Conversely, QWEN2.5-7B drops from 65.5% on the full set to 48% on the 50-triple subset, likely reflecting high variance in the small sample, where a few systematic errors substantially affect accuracy.

4.4 Track B

4.4.1 Model Testing Results

Table 4 presents the performance of our baseline models. SENTENCET5-XXL achieves the highest accuracy of 69.0 on the dev set (61.5 on the final test set) on the final test set, outperforming all other models. No other model—whether using the max-similarity strategy (Section 4.4.2) or fine-tuning (Section 4.4.3)—surpassed this result.

Model	Plain	Fusion method	Sentence representation		
			First token	Last token	Mean
ALL-MINILM-L6-V2	55.0	mean	44.0	51.0	51.5
		sum	55.0	60.0	57.0
		max	53.0	54.0	54.0
MSMARCO-DISTILBERT	62.0	mean	49.0	57.0	58.0
		sum	**62.5**	60.5	61.5
		max	58.0	58.0	60.0
MULTILINGUAL-E5	60.0	mean	60.0	**62.5**	**66.5**
		sum	59.0	59.5	61.0
		max	62.0	58.0	64.5

Table 5: Max-similarity results on the development set with different fusion strategies and sentence representations (see §3.5.2). *Plain* = standard full-document cosine similarity. Best results in **bold**, second best underlined.

Model	Baseline	org. syn	TFB-syn			
			Easy	Medium	Hard	Hard-fs
ALL-MINILM-L6-V2	55.0	56.6	56.2	57.4	57.00	**58.1**
MSMARCO-DISTILBERT	62	62.5	64.1	63.5	63.6	**64.8**
MULTILINGUAL-E5	60.0	61.7	59.3	62.9	61.8	**63.1**

Table 6: Fine-tuning results on dev set per synthetic data tier (incremental training). Best per model in **bold**, second best underlined. Org. syn. = synthetic dataset provided by the organizers. TFB. syn. = synthetic dataset created by our team.

4.4.2 Max-Similarity Results

Table 5 presents results for sentence-level max-similarity across all configurations. The late interaction approach yields improvements for some model-configuration pairs, but the optimal representation and fusion strategy are not generalizable across models. MULTILINGUAL-E5 shows the largest gain: $+8.5$ percentage points with mean fusion and mean sentence representations. Notably, MSMARCO-DISTILBERT achieve their best max-similarity results (62.5%) with sum fusion and first-token representation, while for ALL-MINILM , sum fusion on last token consistently outperforms mean and max fusion.

4.4.3 Fine-Tuning Results

The fine-tuning results across the different synthetic data tiers are presented in Table 6 and model's performances on the final test set are presented in Appendix C in Table 7. TFB-syn hard-fewshot yield the best performance for all the models. The overall highest performance is achieved by the MSMARCO-DISTILBERT model fine-tuned on the TFB-hard fewshot dataset. No model surpasses the 66.5% obtained by sentence-level max-similarity without any fine-tuning (Table 5). For all models, TFB synthetic datasets yield better performance than the organizers' synthetic dataset. This result is consistent with the distribution shift identified in §4.2. Overall, performance differences among the medium, hard, and hard-fewshot tiers are inconsistent across all models, suggesting that differences in difficulty levels are not perfectly reflected in the datasets.

5 Discussion

Task difficulty and the human ceiling. The low agreement (TFB : $\alpha = 0.32$; organizers : $\alpha = 0.33$) indicates substantial ambiguity in narrative similarity judgments. On the annotated subset, several LLMs predominantly predict one class while masking their inability to perform genuine narrative comparison. Although the organizers specify criteria such as theme, course of action, and outcomes, these remain largely intuitive both for LLMs and human annotators. We argue that grounding the task in clearer theoretical assumptions could improve both annotation consistency and model behavior.

Synthetic data quality matters more than quantity. The distribution shift (Figure 2) and sentence-order preservation (Figure 3) present in the organizers' synthetic dataset are partially mitigated by our TFB-syn dataset, which generates more diverse sentence arrangements. This results in improved model performance on Track B for both TFB-hard and hard-fewshot compared to fine-tuning with the organizers' synthetic data, even though TFB-hard and hard-fewshot are smaller in size. Finally, these improvements still do not surpass the best unfine-tuned max-similarity result. This suggests that, for narrative similarity tasks, **representation architecture may matter more than training data volume**.

6 Conclusion

We described the participation of team TFB in SemEval-2026 Task 4, submitting SENTENCE-T5-XXL for Track B and QWEN2.5-7B for Track A. Three findings emerge. First, sentence-level max-similarity can substantially improve embedding-based narrative comparison, but the optimal configuration is model-dependent. Second, synthetic data distribution proximity to the target data matters more than volume. Third, narrative similarity is genuinely hard, as confirmed by low human agreement. We argue that focusing on modeling the task—by theoretically clarifying what constitutes themes, course of action, and outcomes—would

help enhance not only annotators' agreement but also the creation of more realistic synthetic data, thereby improving model performance.
Future work could explore entity pseudonymization (Hatzel and Biemann, 2024) to prevent models from using surface cues, cross-encoder architectures for better conditional narrative understanding, and targeted hard negatives, especially outcome-flipped pairs, to address challenging cases for both humans and models.

Limitations The annotation study is limited by a small sample of 50 triples annotated by three non-native speakers, reducing inter-annotator agreement reliability. In addition, Track B experiments only test four embedding models, so including more (e.g., instruction-tuned large embedders) could improve generalizability.

References

M. Bal and C. van Boheemen. 2009. *Narratology: Introduction to the Theory of Narrative*. Narratology: Introduction to the Theory of Narrative. University of Toronto Press.

Louis Castricato, Spencer Frazier, Jonathan Balloch, and Mark Riedl. 2021. Fabula entropy indexing: Objective measures of story coherence. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 84–94, Virtual. Association for Computational Linguistics.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

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

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

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Susana Onega and José Angel García Landa. 2014. *Narratology: an introduction*. Routledge.

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.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Wolf Schmid. 2010. *Narratology: an introduction*. Walter de Gruyter.

Karen Sparck Jones. 1988. *A statistical interpretation of term specificity and its application in retrieval*, page 132–142. Taylor Graham Publishing, GBR.

Danna Zheng, Mirella Lapata, and Jeff Z. Pan. 2025. Long-form information alignment evaluation beyond atomic facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11007–11027, Suzhou, China. Association for Computational Linguistics.

A Track A Prompts

We use two system prompts with a shared user template. Models must answer with a single letter (*A* or *B*).

Prompt-1 (Expert Persona)

```
You are a narrative analysis expert.
Which story (A or B) is more similar to the anchor?
Answer only: A or B.
```

Prompt-2 (Minimal)

```
Which text (A or B) is closer to the anchor?
Answer only: A or B
```

User Template

```
Anchor: {anchor}

A: {text_a}

B: {text_b}
```

B Synthetic Data Generation

Triplets follow strict narrative requirements: (i) theme, (ii) structure (goal, conflict, stakes), (iii) style (tone, dynamics, genre), and (iv) logic (causality, setting). A is a strong sibling narrative; B is weaker but valid.

Easy

```
Generate (anchor, A, B).
A: strong alignment
B: clearly weaker

Return JSON:
{"difficulty":"easy","anchor":"...",
"text_a":"...","text_b":"..."}
```

Medium

```
Generate:
- Anchor (120–200 words)
- A (80–160): strong
- B (80–160): weaker, mild inconsistencies

Return JSON:
{"difficulty":"medium","anchor":"...",
"text_a":"...","text_b":"..."}
```

Hard (Few-Shot)

```
Given example (format only).

A & B: both strong
A slightly better (subtle factors)

Generate:
Anchor (120–200), A/B (80–160)

Return JSON:
{"difficulty":"hard-fewshot","anchor":
"...","text_a":"...","text_b":"..."}
```

C Performance on the final test set

Model	Baseline	org. syn	TFB-syn			
			Easy	Medium	Hard	Hard-fs
all-MiniLM-L6-v2	59.75	51.1	51.0	49.4	49.95	**51.35**
msmarco-distilbert	59.75	53.3	53.2	53.6	**54.3**	54.1
multilingual-e5	61.75	53.0	52.4	52.1	**52.4**	**53.6**

Table 7: Fine-tuning results on test set per synthetic data tier (incremental training). Best per model in **bold**, second best underlined. Org. syn. = synthetic dataset provided by the organizers. TFB. syn. = synthetic dataset created by our team.

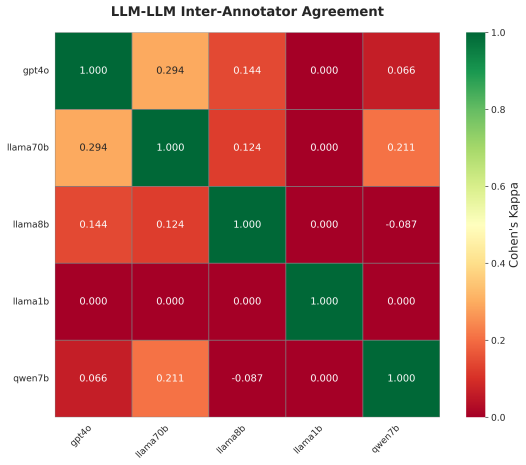


Figure 4: Inter-LLM agreement (Cohen's κ).

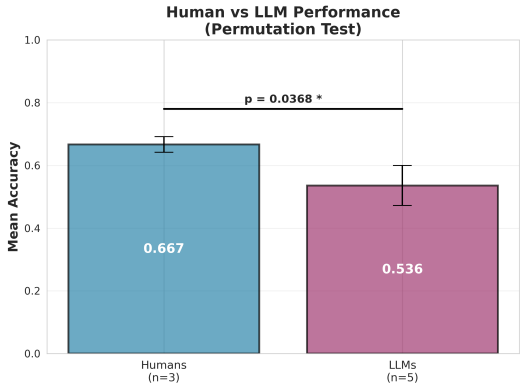


Figure 5: Mean accuracy of human annotators vs. LLMs on the 50-triple subset. The difference is statistically significant (permutation test, $p = 0.037$).

Dataset	n_{texts}	Avg. tok	Ent./100 tok
TFB-syn easy	3,600	78	0.62
TFB-syn medium	3,600	142	3.17
TFB-syn hard	3,600	134	2.85
TFB-syn hard-fewshot	1,800	146	4.46
Org. synthetic	3,794	180	2.68
Dev (Track A)	400	141	7.09
Test (Track A)	800	142	7.08

Table 8: Entity density (named entities per 100 tokens, spaCy `en_core_web_sm`) by dataset. Dev and test have 2.5–11× higher density than any synthetic tier.