CUNI at SemEval-2026 Task 4: Multi-Head Narrative Aspect Disentanglement via Entangled Synthetic Dataset

Jan Mitka[1] **and Jindřich Helcl**[2]
[1]Charles University, Faculty of Mathematics and Physics
[2]University of Oslo, Language Technology Group
Correspondence: mitka.honza@gmail.com

Abstract

We participate in Track B of the SemEval 2026 Task 4 on narrative similarity, focusing on narrative representation learning. We introduce a synthetic dataset designed to disentangle core narrative aspects-abstract theme, course of action, and outcome-and propose a multi-head multi-positive extension of the InfoNCE objective to train aspect-specific embeddings. Our best model achieves 64.25% accuracy on the test set. A nearest-centroid analysis indicates partial aspect-specific structure in the submitted checkpoint, while the training dynamics reveal a partial misalignment between the contrastive objective and the triplet-based evaluation protocol.

1 Introduction

The narrative similarity task focuses on identifying similar stories. The similarity of stories is determined by underlying structural aspects rather than surface-level elements such as character names, environmental settings, or specific objects. The core narrative aspects-abstract theme, course of action, and outcome-are often entangled within a story, making the similarity comparison challenging. To address this, we adopt a disentangled representation learning approach to separate these features in the latent space (Higgins et al., 2017). Our team participated in Track B, which focuses on narrative representation learning. In this setting, stories are first encoded into embedding representations, and similarity is evaluated through a triplet-based comparison. The dataset consists of English Wikipedia story summaries.

We create a model based on the `Qwen3-Embedding-4B` (Zhang et al., 2025) with three embedding heads, each for a single core narrative aspect. We construct a synthetic dataset designed to simulate the entanglement of the aspects. The model is then fine-tuned using an InfoNCE-based loss function on this dataset.

We observe that optimizing the loss function does not directly translate into improved performance. Indeed, the peak development accuracy occurred early in the training. Our best model ranks 14th out of 28 teams, achieving 64.25% accuracy on the test set. The synthetic disentanglement training improves development performance, but the improvement only partially generalizes to the blind test set.

Our contributions are: (i) a synthetic dataset for core-aspect disentanglement, (ii) a multi-head multi-positive InfoNCE objective for aspect-specific representation learning, and (iii) an ablation and representation analysis of the learned aspect structure.

2 Task Overview

Stories can be similar in many different semantic aspects. Quoting the task organizers (Hatzel et al., 2026), they define three categories for narrative similarity comparison as follows:

- **Course of Action**: The sequence of happenings in the story.
- **Outcomes**: The results of the happenings in the story, excluding intermediate results that change later on.
- **Abstract Theme**: The motifs and themes explored in the story. This aspect does not cover the concrete setting of a story.

The task comprises two tracks, A and B. Both of them follow a triplet-based evaluation protocol: given an anchor story and two candidates, the system must determine which candidate shares a higher degree of narrative similarity with the anchor. The system performance is measured using accuracy. In Track A, the systems perform the comparison end to end, receiving the whole triplet on the input. In Track B, the systems produce story

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2163–2178

embeddings, which are then used in an external decision process that uses cosine similarity to resolve which of the two stories is more similar to the anchor.

The task organizers provided the dataset to participants, containing a development set with 479 stories forming 200 triplets. The final evaluation was a blind test hosted on Codabench[1], consisting of 849 stories forming 400 triplets.

3 Synthetic Dataset

The core narrative aspects of a story are usually intertwined. We aim to separate (disentangle) them in the latent embedding space (Higgins et al., 2017). We create a model with three embedding heads, each designed to disentangle one core narrative aspect from the text.

To facilitate the disentanglement while retaining control over the three aspects, we generate training data using LLMs. The generation process follows a two-step procedure: first, we construct the configuration of the story; second, conditioned on this configuration, the LLM generates the corresponding story.

The resulting synthetic dataset for this task consists of 50,245 training samples generated by `gpt-oss-120b` (OpenAI, 2025). Every training sample is a five-tuple of stories, which translates into 251,225 stories in total. Each story was then evaluated by `Qwen3-30B-A3B-Thinking-2507-FP8` (Qwen Team, 2025).

To ensure highly controlled story generation, we defined discrete option lists for the core narrative aspects (See Appendix B). These lists were curated with the assistance of ChatGPT 5.2 and Gemini 3 to ensure a diverse yet consistent range of narrative scenarios.

For each story a, we construct its configuration $\text{conf}(a) = (T_a, C_a, O_a, S_a)$ where T_a is the abstract theme, C_a is the course of action, O_a is the outcome, and S_a is the stylistic metadata (setting, narrative style, syntax profile, and length bucket).

3.1 Training Example

Each training example is structured as a five-tuple of generated stories $(a, a_S^+, a_T^-, a_C^-, a_O^-)$ consisting of an anchor and four contrastive examples, where:

- $\text{conf}(a) = (T_a, C_a, O_a, S_a)$
- $\text{conf}(a_S^+) = (T_a, C_a, O_a, S \neq S_a)$
- $\text{conf}(a_T^-) = (T \neq T_a, C_a, O_a, S_a)$
- $\text{conf}(a_C^-) = (T_a, C \neq C_a, O_a, S_a)$
- $\text{conf}(a_O^-) = (T_a, C_a, O \neq O_a, S_a)$

This setup provides the anchor story, one hard positive (a_S^+) representing stylistic variation which we consider to result in a similar story, and three hard negatives (a_T^-, a_C^-, a_O^-) that differ from the anchor story in exactly one core aspect.

Training sample creation. First, we generate an anchor story by sampling uniformly from the discrete option lists, with the exception of the length bucket stylistic attribute, where we consistently select the shortest bucket.

Second, for every anchor that passes our quality threshold (see Section 3.2), we generate a hard positive sample a_S^+. This story maintains the original core narrative aspects but utilizes different stylistic options. Regarding length, we apply a probabilistic approach: with $p = 0.6$, the story retains the anchor's length, while the remaining instances are assigned to one of the two longer buckets. To maintain prompt simplicity and model focus, the hard positive is generated using the same prompt template as the anchor, without including the anchor text itself in the context.

Finally, for each validated hard positive, we construct three hard negatives corresponding to the core narrative aspects (a_T^-, a_C^-, and a_O^-). For each negative, we sample uniformly from the remaining options of the target aspect while keeping all other parameters identical to the anchor. Unlike the positive samples, the prompt for hard negatives explicitly includes the anchor story; the model is instructed to mirror the anchor as closely as possible, modifying only the single designated narrative aspect.

We provide additional statistics on the aspect option distributions and negative pairing patterns in Appendix C.

3.2 Judging of Stories

Not all combinations from the discrete option lists yield coherent stories. Additionally, we aimed to ensure the high quality of the synthetic dataset. We employed an LLM-based judge to evaluate the generated stories on their overall coherence and how well the story aligns with the provided core narrative options. Each dimension was scored on

[1] https://www.codabench.org/competitions/10273/

a 1–5 scale, where 1 indicates no alignment, and 5 indicates a clear and strong match. Starting from 100,000 generated anchor stories, the thresholding process resulted in 50,245 training samples.

The resulting dataset is publicly available.[2]

3.3 Sampling Based on the Development Distribution

We also experimented with sampling the training data according to the estimated aspect distribution of the development set. We used the `gpt-oss-20b` (OpenAI, 2025) model to annotate each development story with a discrete option for each core narrative aspect $A \in \{T, C, O\}$ (Theme, Course, Outcome). For each aspect A, let Ω_A denote the finite set of allowed options, augmented with a NONE option.

Based on the annotations, we estimated the aspects' options distribution $P_A(o)$ for $o \in \Omega_A$. To account for the annotation error and the NONE option, we use a smoothed distribution:

$$P'_A(o) = (1 - \alpha)P_A(o) + \alpha \cdot \frac{1}{|\Omega_A|},$$

where $\alpha \in [0, 1]$ is a smoothing parameter.

Each training sample $x \in \mathcal{D}$ is associated with an aspect option $a_A(x) \in \Omega_A$ for each $A \in \{T, C, O\}$ (given by our data generation metadata). Assuming independence across aspects, we define a target weight

$$w(x) = \prod_{A \in \{T,C,O\}} P'_A\big(a_A(x)\big).$$

We then sample from the buffer proportionally to these weights:

$$\Pr(x \mid \mathcal{B}_t) = \frac{w(x)}{\sum_{x' \in \mathcal{B}_t} w(x')} \quad \text{for } x \in \mathcal{B}_t.$$

4 Model

Our model consists of a backbone and three embedding heads, each for one aspect. The backbone used is the `Qwen3-Embedding-4B` (Zhang et al., 2025). The text representation from this model is obtained from the EOS[3] token and has 2560 dimensions. The embedding heads are three separate linear projections applied to the EOS token representation. Each head has a normalized 2560-dimensional output, creating the aspect embedding. See Figure 9 for the model's architecture.

The final story embedding is obtained by concatenating the three narrative aspect embeddings produced by the heads. Before concatenation, the aspect embeddings can be assigned equal or varying weights. The resulting concatenated vector is then normalized to produce the final story embedding. Narrative similarity is then computed by the dot product of the final story embeddings.

Stories can also be compared using only a subset of the three aspect embeddings, following the same procedure as described above. This approach enables story comparison based only on selected narrative aspects.

4.1 Multi-Head Multi-Positive InfoNCE Loss

To remain consistent with the contrastive objective used during backbone pre-training, we used the contrastive loss function InfoNCE (van den Oord et al., 2018). We adapted this loss to the three embedding heads and the entangled dataset, which we call the multi-head multi-positive InfoNCE loss function.

For each training sample, we consider an anchor story and a set of candidate stories within the batch. For head $h \in \{\text{theme, course, outcome}\}$, let $\mathbf{e}_i^{(h)}$ denote the anchor embedding and $\mathbf{e}_j^{(h)}$ the embeddings of all other stories in the batch.

For each anchor embedding $\mathbf{e}_i^{(h)}$, positives S_i^+ are stories in the batch sharing the same aspect label, while all others act as negatives S_i^-.

We define the multi-positive InfoNCE loss as

$$\mathcal{L}^{(h)} = -\log \frac{\sum_{j \in S_i^+} \exp(s_{ij}/\tau)}{\sum_{j \in S_i^+ \cup S_i^-} \exp(s_{ij}/\tau)},$$

where $s_{ij} = \mathbf{e}_i^{(h)} \cdot \mathbf{e}_j^{(h)}$ denotes cosine similarity (dot product on normalized embeddings) and τ is a temperature parameter. The loss is averaged across anchors in the batch.

The multi-head aggregation is defined as:

$$\mathcal{L}_{headNCE} = \sum_{h=1}^{3} \alpha_h \mathcal{L}^{(h)},$$

where the α_h is the weight of the particular head.

4.2 Disentangled Representation Learning

Our disentangled representation learning setup is based on three components: the multi-head architecture, the controlled synthetic dataset, and the

[2] https://huggingface.co/datasets/ufal/NarrativeAspect-EntangledSynthetic

[3] End of Sequence token. This token is appended at the end of the input sequence.

multi-head contrastive objective. The architecture assigns one embedding head to each core narrative aspect: abstract theme, course of action, and outcome. The synthetic dataset provides supervision by generating story tuples in which only one narrative aspect is changed at a time, allowing each head to receive an aspect-specific training signal.

Following the standard view of disentangled representations, we consider three desirable properties: modularity, compactness, and explicitness (Carbonneau et al., 2022). In our setting, modularity means that changing one narrative aspect should primarily affect the corresponding head. Compactness means that information about one aspect should be concentrated in its assigned representation rather than spread across all heads. Explicitness means that the aspect should be recoverable from its corresponding embedding space.

In addition to this supervised signal, we use Centered Kernel Alignment (CKA) regularization (Kornblith et al., 2019) to penalize similarity between the representational geometries of different heads. This acts as a proxy objective for reducing information overlap across aspect representations, but it is not the only source of disentanglement. Specifically, the CKA loss is computed as:

$$\mathcal{L}_{CKA} = \frac{Tr(K_X K_Y)}{\sqrt{Tr(K_X K_X) \cdot Tr(K_Y K_Y)}}$$

where K_X and K_Y are the kernel matrices of the embeddings for the different heads, and Tr denotes the trace of the matrix, which measures the alignment between the embeddings of different heads.

We define the total loss as a weighted sum of the multi-head InfoNCE and the CKA loss:

$$\mathcal{L} = \mathcal{L}_{headNCE} + \lambda \mathcal{L}_{CKA}$$

where λ is the weight of the CKA loss.

5 Experimental Setup

5.1 Dataset Usage

We only use a subset of the created dataset. We limited the length of training samples to 512 tokens. We also restricted the quality of the training samples and removed samples that scored ≤ 3. With these restrictions, we got 36 840 training samples (184 200 stories in total). To reduce the RAM overhead, we created a moving window of size 16,384 from which we do weighted sampling (see Section 3.3).

5.2 Training Configuration

We used the AdamW optimizer (Loshchilov and Hutter, 2019), and set the learning rate to 1e-5 for the heads and 8e-6 for the backbone. We use cosine scheduler (Loshchilov and Hutter, 2017) with linear warmup (Devlin et al., 2019) with 920 steps. Set the weight decay to 0.01, and dropout set to 0.1. We used the bf16 precision.

All head weight losses are set to 1, and $\tau = 0.2$. We compute the CKA loss on non-anchor vectors, with weight 0.05. We start applying the CKA regularization at step 920 with a 450-step ramp-up.

We set the batch size to 4 (a total of 20 stories). Gradient accumulation did not yield any improvements. The moving-window has size 16,384 sampling based on the development distribution, and we set $\alpha = 0.4$. We limited training to 4,000 steps based on development performance. We evaluate the model every 5 steps. The model is saved whenever it achieves a new peak accuracy across any possible subset of its three heads.

We used an NVIDIA H100 GPU for dataset generation and all experiments.

5.3 Ensemble

We ensemble heads by aggregating their per-triplet preference signals $\Delta_h = \mathrm{sim}_h(a) - \mathrm{sim}_h(b)$ into a single score $S = \sum_h w_h \Delta_h$, predicting a when $S > 0$. We use a non-negative weighted ensemble, where w_h are tuned on the development set using constrained L-BFGS-B.

6 Results

Table 1 reports the results for Track B. Our best single model achieves 75.5% accuracy on the development set and 63.75% on the test set. The ensemble improves development performance to 76.5%, with 64.25% on the test set. Although all heads are trained jointly, the best performance uses only the *cr+ou* combination, consistent with Section 6.2. The ensemble assigns weights of 0.56 and 0.44 to course and outcome.

Compared to the backbone model (`Qwen3-Embedding-4B`), our approach improves performance by +10.5% on the dev, by +3.5% on the test.

We compare against the official baselines: random choice, `all-MiniLM-L6-v2`, and story-emb (Hatzel and Biemann, 2024), as well as an additional `all-mpnet-base-v2` model. While some

System	Dev (%)	Test (%)
Random	50.00	50.00
all-MiniLM-L6-v2	55.00	58.50
all-mpnet-base-v2	61.50	55.50
story-emb	57.00	63.25
Qwen3-Embedding-4B	65.00	60.75
Ours (*cr+ou* combination)	75.50	63.75
Ours (ensemble)	76.50	64.25
COGNAC	-	**72.00**

Table 1: Task B results.

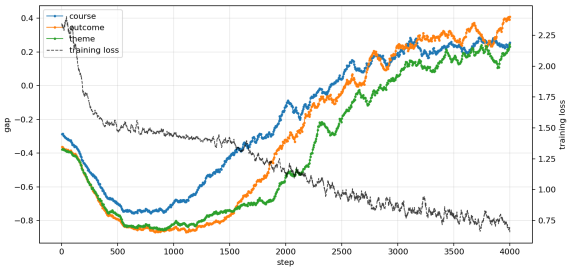

Figure 1: Gap graph between the similarities of (anchor, hard positive) and (anchor, hard negative) for every head. We also included the training loss.

baselines exhibit noticeable dev–test variance, our model remains competitive across both splits.

In the official Track B ranking, we placed 14th out of 28 teams. The top-performing system (COGNAC) achieved 72% accuracy on the test set.

6.1 Analysis

The objective of the training encourages attracting embeddings of positive samples to anchors while drawing away negative samples. To monitor this behavior, we track the cosine gap between (anchor, hard positive) and (anchor, hard negative) similarities, see Figure 1. Except for the warm-up phase, the gap increases, indicating that the model follows the intended contrastive objective. (The individual similarity trajectories are shown in Figure 10.)

Figure 2 shows the learning curves for two head combinations. The highest accuracy is achieved at around 2,000 training steps, then the accuracy declines. Interestingly, the cosine gap at this point remains negative (approximately -0.2), indicating that, on average, hard negatives are still closer to the anchor than positives. However, the gap is computed as a batch-level average and measures separation between synthetically constructed hard positives and negatives, which are extreme contrastive cases. In Task B, the evaluation triplets

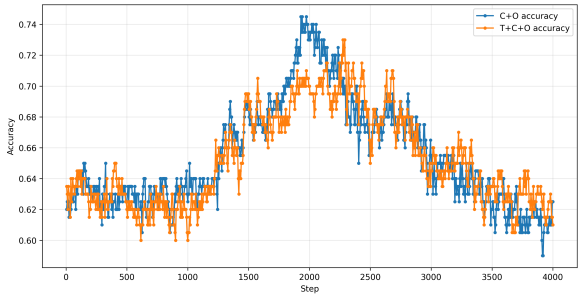

Figure 2: Graph of accuracies of two combinations: all heads combined, and course-outcome.

are not explicitly designed as such extreme oppositions. Therefore, the negative cosine gap does not necessarily imply incorrect ranking inside the evaluation triplets.

Further training increases the cosine gap, yet accuracy drops. This indicates partial misalignment between the training objective and the downstream evaluation metric. While the multi-head multi-positive InfoNCE objective pushes away all negatives and increases global margins, Task B evaluates only the relative ranking within individual triplets.

Although early stopping was applied on the development set, performance drops on the test set. Given the limited size of the development split (200 triplets), checkpoint selection may be sensitive to statistical noise. Interestingly, the official baseline model (story-emb) exhibits the opposite trend, performing better on the test set than on development, which further indicates variability between the two splits.

6.2 Ablation Analysis

For each run, we recorded the best development accuracy achieved across all head combinations. Table 2 reports, for each model variant, only the best-performing head combination, averaged over n independent runs. The standard deviation is shown in parentheses. The highest mean development accuracy is achieved by the disentangled baseline with the *cr+ou* head combination, reaching 72.40%. Across all variants of the proposed architecture, the best-performing head combination is consistently *cr+ou*, even though the models are trained with all three narrative heads.

The frozen-backbone ablation performs substantially worse than the fully trainable variants, suggesting that adapting the backbone is necessary for the synthetic supervision to affect the representa-

System	n	Dev (%)
Qwen3-Embedding-4B	–	65.00
Single-head	3	67.33 (1.04)
Frozen-backbone	3	66.50 (1.32)
Course-outcome	3	64.83 (2.08)
Disent. baseline	5	**72.40** (1.39)
Distr. sampling	5	71.80 (0.84)
CKA	5	71.20 (1.30)
XCov	5	71.90 (0.65)
Orthogonality	5	71.80 (1.99)

Table 2: Task B ablation results on the development set. For each model variant, we report the best development accuracy across all available head combinations, averaged over n independent runs, with standard deviation in parentheses. The dashed line separates ablation baselines from variants of the proposed multi-head architecture, including sampling and regularization variants (CKA, XCov, Orthogonality). Disent. is an abbreviation for disentangled, and distr. is an abbreviation for distribution.

tion space.

The Course-outcome ablation removes the theme head and trains only the course and outcome heads, testing whether the best-performing *cr+ou* decision can be learned without the theme representation. This variant underperforms the full three-head models, indicating that the theme head remains useful during training even when the best evaluation performance is obtained from the *cr+ou* head combination.

The single-head ablation replaces the three aspect-specific heads with a single embedding head trained on a global similarity target, testing whether the synthetic dataset is useful without disentangled representation learning. Although this model improves over the original backbone, it remains below the multi-head variants, suggesting that the synthetic dataset is most effective when combined with the proposed aspect-specific representation learning.

Since the official submission was selected from a single development-optimized training run, we report it separately from the averaged ablation results.

6.3 Representation Analysis

The nearest-centroid recovery matrix evaluates how well each embedding head can recover each narrative aspect by assigning stories to the closest aspect-option centroid in the corresponding embedding space. For an ideally explicit representation, each head should recover its assigned aspect best, so the highest values should appear on the diagonal.

Head	Theme	Course	Outcome
Theme	**59.7**	46.4	12.3
Course	39.8	**70.3**	12.9
Outcome	30.8	**31.8**	31.1
Random	3.3	2.5	2.9

Table 3: Nearest-centroid recovery matrix for the submitted model. All values are reported in percent. Rows denote the embedding head used for retrieval, while columns denote the recovered narrative aspect. Higher values indicate more recoverable aspect information. The Random row shows the chance-level baseline.

Table 3 shows partial explicitness in the submitted checkpoint: the theme and course heads recover their assigned aspects well above chance, with the course head showing the clearest separation. However, the theme head also recovers course information strongly, and the outcome head does not recover its assigned aspect more clearly than the other aspects, indicating substantial information leakage. These results suggest that the submitted model learns aspect-specific structure, but the separation remains imperfect. Together with the gap analysis, this supports the interpretation that stronger optimization of the synthetic contrastive objective and better aspect separation are only partially aligned with the downstream triplet-based evaluation.

7 Conclusion

We demonstrate that disentangling narrative components using a multi-head multi-positive InfoNCE objective improves triplet evaluation compared to using an off-the-shelf model directly. Our best model ranks 14th out of 28 teams, achieving 64.25% accuracy on the test set. We observe a partial misalignment between the downstream evaluation metric and the training objective.

References

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. 2022. Measuring disentanglement: A review of metrics. *Preprint*, arXiv:2012.09276.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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 62st Annual Meeting of the Association for Computational Linguistics*, Miami, Florida. Association for Computational Linguistics.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. *Preprint*, arXiv:1608.03983.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A LLM-based Judging Procedure

We employed Qwen3-30B-A3B-Thinking-2507-FP8 as an automated judge to evaluate generated stories. The model assessed each story along four dimensions:

- Overall coherence
- Alignment with abstract theme
- Alignment with course of action
- Alignment with outcome

Each dimension was scored on a 1–5 scale, where 1 indicates no alignment and 5 indicates a clear and strong match.

After every generation cycle, the stories evaluated by the judge. Any story receiving a score ≤ 2 on any of the three main narrative aspects was discarded.

For the hard-negative stories, we also created an additional judging procedure evaluating:

- Pivot-change strength (how well the designated aspect was modified)
- Non-pivot preservation (how well the remaining aspects were preserved)
- Story similarity to the anchor
- Overall pivot quality

Stories scoring ≤ 2 on any of these criteria were excluded from the final dataset.

B Synthetic Dataset Vocabularies

This appendix lists the controlled vocabularies used for generating the synthetic disentanglement dataset.

Table 4: Abstract Theme Types

#	Theme	Description
1	The Cost of Ambition	Success demands sacrifice of happiness or morality.
2	Fate vs. Free Will	Conflict between destiny and individual agency.
3	The Loss of Innocence	Transition from naive to realistic worldview.
4	The Corrupting Influence of Power	Authority alters human psychology.
5	The Cyclical Nature of Violence	Revenge breeds further revenge.
6	Individual vs. The Collective	Conformity versus personal identity.
7	The Subjectivity of Truth	Reality changes based on perspective.
8	Redemption through Sacrifice	Past sins are repaid through sacrifice.
9	The Illusion of Safety	Security is fragile and socially constructed.
10	Nature vs. Technology	Conflict between natural and artificial systems.
11	The Burden of Legacy	Struggle with ancestral expectations.
12	Isolation in a Crowd	Alienation despite social proximity.
13	The Fragility of Memory	The past is reconstructed and unreliable.
14	Love as a Destructive Force	Affection leads to ruin rather than salvation.
15	Duty vs. Desire	Conflict between obligation and personal wish.
16	The Fear of the Unknown	Anxiety toward the unseen or incomprehensible.
17	Appearance vs. Reality	Surface impressions conceal deeper truth.
18	The Absurdity of Existence	Life lacks inherent meaning.
19	Hubris	Extreme pride leads to downfall.
20	Tradition vs. Progress	Established customs versus change.
21	The Monster Within	Evil resides within the protagonist.
22	Found Family	Emotional bonds outweigh biological ties.
23	The Inevitability of Death	Mortality as natural inevitability.
24	Class Struggle	Wealth disparity and social stratification.
25	Identity Crisis	Struggle to define self beyond labels.
26	The Duality of Human Nature	Capacity for both good and evil.
27	The Unbreakable Human Spirit	Hope persists in desolation.
28	The Commodification of Life	Treating people as products or resources.
29	Obsession	Single focus consumes life.
30	The Passing of Time	Decay and the fleeting nature of moments.

Table 5: Course of Action Types

#	Pattern	Description
1	The Linear Quest	Journey to obtain or deliver a goal.
2	The Siege Defense	Protagonist confined, responding to attacks.
3	The Investigation	Clue discovery and deduction.
4	The Heist Sequence	Plan, execute, withdraw risky operation.
5	The Chase	Prolonged pursuit.
6	The Escape	Plan and attempt to flee captivity.
7	The Tournament	Structured contests with eliminations.
8	The Metamorphosis	Gradual physical or mental transformation.
9	The Cat and Mouse	Strategic back-and-forth conflict.
10	The Descent	Progressive worsening circumstances.
11	The Seduction	Gradual manipulation of trust.
12	The Road Trip	Episodic journey encounters.
13	The Time Loop	Repeated events with variation.
14	The Negotiation	Tense bargaining process.
15	The Survival	Securing resources under threat.
16	The Infiltration	Enter hostile group under disguise.
17	The Rebellion	Organized resistance against power.
18	The Reconstruction	Reverse reconstruction of past events.
19	The Discovery	Hidden fact revealed; consequences unfold.
20	The Creation	Step-by-step invention with consequences.
21	The Fall from Grace	Progressive loss of status.
22	The Rashomon	Same event from multiple perspectives.
23	The Parallel Narrative	Alternating storylines converge.
24	The Bet	Escalating challenge after wager.
25	The Waiting Game	Confined anticipation.
26	The Hunt	Systematic pursuit.
27	The Body Swap	Characters exchange roles or bodies.
28	The Trial	Legal or formal proceedings.
29	The Exchange	Tense trade attempt.
30	The Odyssey	Interrupted journey home.
31	The Test	Sequence of trials to gain access.
32	The Mistake	Small error escalates.
33	The Guardian	Protect vulnerable entity through danger.
34	The Double Cross	Betrayal disrupts plan.
35	The Addiction	Cyclical compulsion and relapse.
36	The Experiment	Controlled conditions alter behavior.
37	The Invasion	External force enters environment.
38	The Gathering	Diverse characters converge; tensions surface.
39	The Relay	Responsibility passed sequentially.
40	The Duel	One-on-one confrontation.

Table 6: Outcome Types

#	Outcome	Description
1	Total Victory	Goal achieved; antagonist neutralized; protagonist survives.
2	Total Tragedy	Goal fails and protagonist is destroyed.
3	Pyrrhic Victory	Goal achieved at devastating cost.
4	Moral Victory, Physical Defeat	Practical loss but values/message endure.
5	Physical Victory, Moral Defeat	Outward win via violated principles.
6	Ambiguous Ending	Ends decisively without revealing outcome.
7	Circular Ending	Ends essentially where it began.
8	The Cycle Continues	Conflict resolved but pattern persists.
9	The Twist Reveal	Late revelation reframes the story.
10	Deus Ex Machina	External force suddenly resolves conflict.
11	Mutual Destruction	Protagonist and antagonist destroy each other.
12	The Compromise	Uneasy truce; neither side wins.
13	Institutionalization	Protagonist ends under institutional control.
14	Exile	Protagonist survives but is permanently expelled.
15	Ascension	Protagonist departs ordinary world permanently.
16	The Villain Wins	Antagonist succeeds and imposes new status quo.
17	The Open Door	Resolution hints at larger future conflict.
18	Peaceful Acceptance	Protagonist accepts unchangeable outcome.
19	The Sacrifice	Protagonist gives up life or something irreplaceable.
20	Integration	Opposing force reconciled or incorporated.
21	The Dream	Events revealed as dream/simulation/hallucination.
22	Passing the Torch	Responsibility passes to successor figure.
23	The Hollow Victory	Victory achieved but purpose becomes empty.
24	The Great Escape	Protagonist escapes without solving root issue.
25	Assimilation	Protagonist joins/conforms to opposing system.
26	Succumbing to Madness	Protagonist loses grasp on reality.
27	The Unresolved Mystery	Central mystery remains unanswered.
28	Divine Judgment	Higher authority intervenes to decide fates.
29	Inversion	Hero/villain roles reverse by conclusion.
30	The Fade Out	Tensions dissolve gradually as life continues.
31	Self-Destruction	Protagonist deliberately destroys self/work/cause.
32	The Lucky Mistake	Accident/misstep produces successful resolution.
33	Matriculation	Protagonist abandons goal after deep change.
34	The Silent Departure	Protagonist resolves problem then leaves quietly.

Table 7: Settings

#	ID	Description
1	Modern City	Streets, office buildings, apartments, cafés, shops, public transport.
2	Small town	Main square, shops, town hall, close-knit residential areas.
3	Countryside	Fields, forests, scattered farms/houses, small village, long distances.
4	Transport hub	Train station/airport/bus terminal with platforms and waiting areas.
5	Large complex	Large campus/building with corridors, offices, multi-purpose rooms.
6	Industrial area	Factories, warehouses, loading docks, logistics halls, truck parking.
7	Preindustrial realm	Villages, markets, roads, fortifications in low-technology world.
8	Space habitat	Corridors and common areas inside a space station or starship.
9	Tourist area	Hotels, beaches/mountains, promenades, restaurants, leisure spots.
10	Wilderness	Remote forests/mountains/desert, trails, camps, improvised shelters.

Table 8: Narrative Styles

#	ID	Description
1	3P neutral past	Third-person, past tense, neutral tone, moderate access to thoughts.
2	1P colloquial past	First-person, past tense, informal voice with commentary and emotions.
3	News report	Impersonal report-like narration emphasizing chronology and facts.
4	Fairytale storyteller	Oral storyteller tone, sometimes addressing the reader directly.
5	Casefile documentary	Formal document style with entries/timestamps and evidence focus.

Table 9: Syntax Profiles

#	ID	Description
1	Simple sparse	Short sentences, active voice, few subordinate clauses.
2	complex embedded	Mix of medium/long sentences with embedded clauses; natural style.
3	dialogue heavy	Story primarily told via dialogue with minimal narration.

Table 10: Length Bucket

#	ID	Specification
1	Short	Target $\approx$ 125 words; min 75, max 175.
2	Medium	Target $\approx$ 250 words; min 200, max 300.
3	Long	Target $\approx$ 400 words; min 300, max 500.

C Additional Synthetic Dataset Statistics

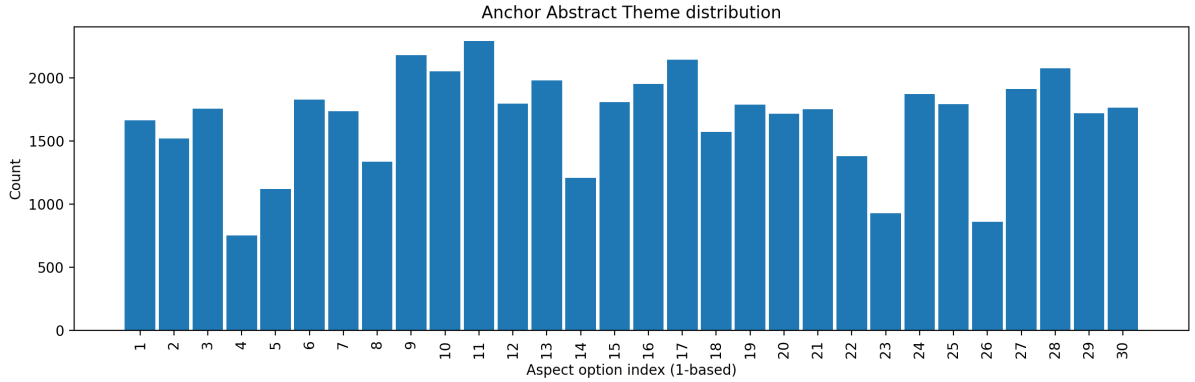

Figure 3: Anchor abstract theme distribution (30 options). Counts correspond to the selected anchor aspect option indices in the synthetic dataset.

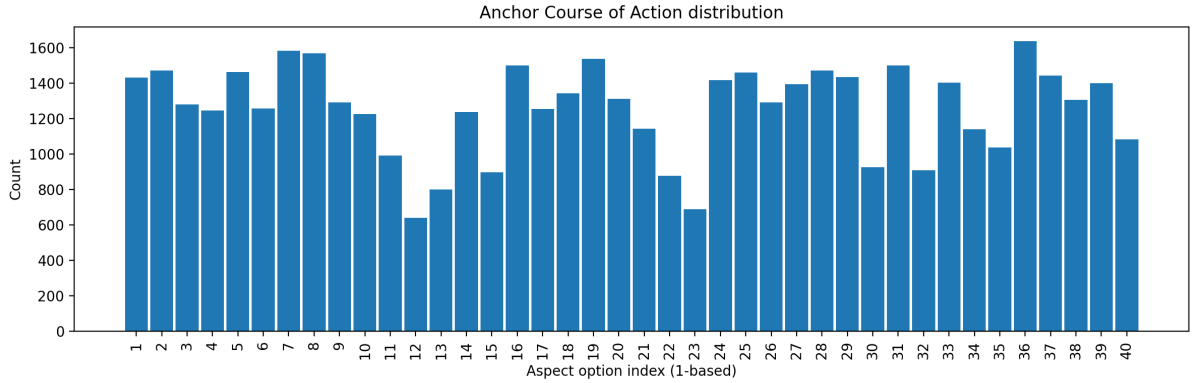

Figure 4: Anchor course-of-action distribution (40 options). The sampling aims to be approximately uniform, with mild deviations due to filtering and generation constraints.

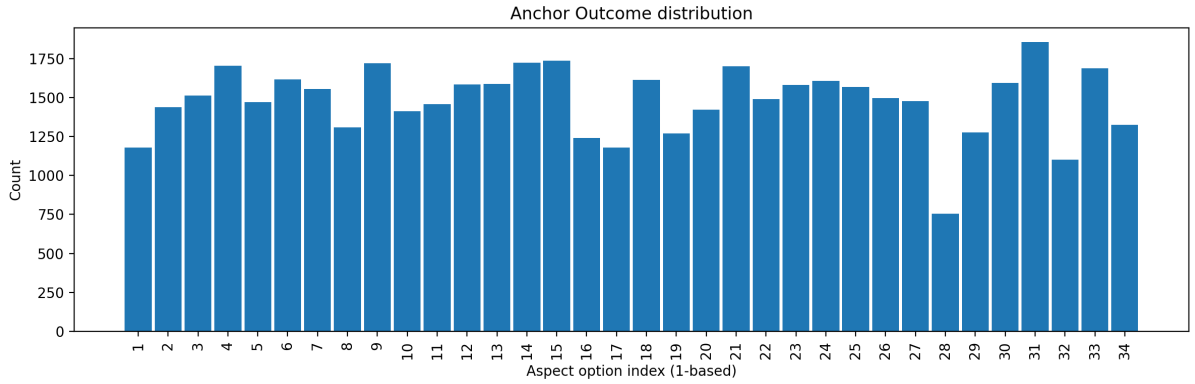

Figure 5: Anchor outcome distribution (34 options). Counts correspond to anchor aspect option indices used during synthetic data generation.

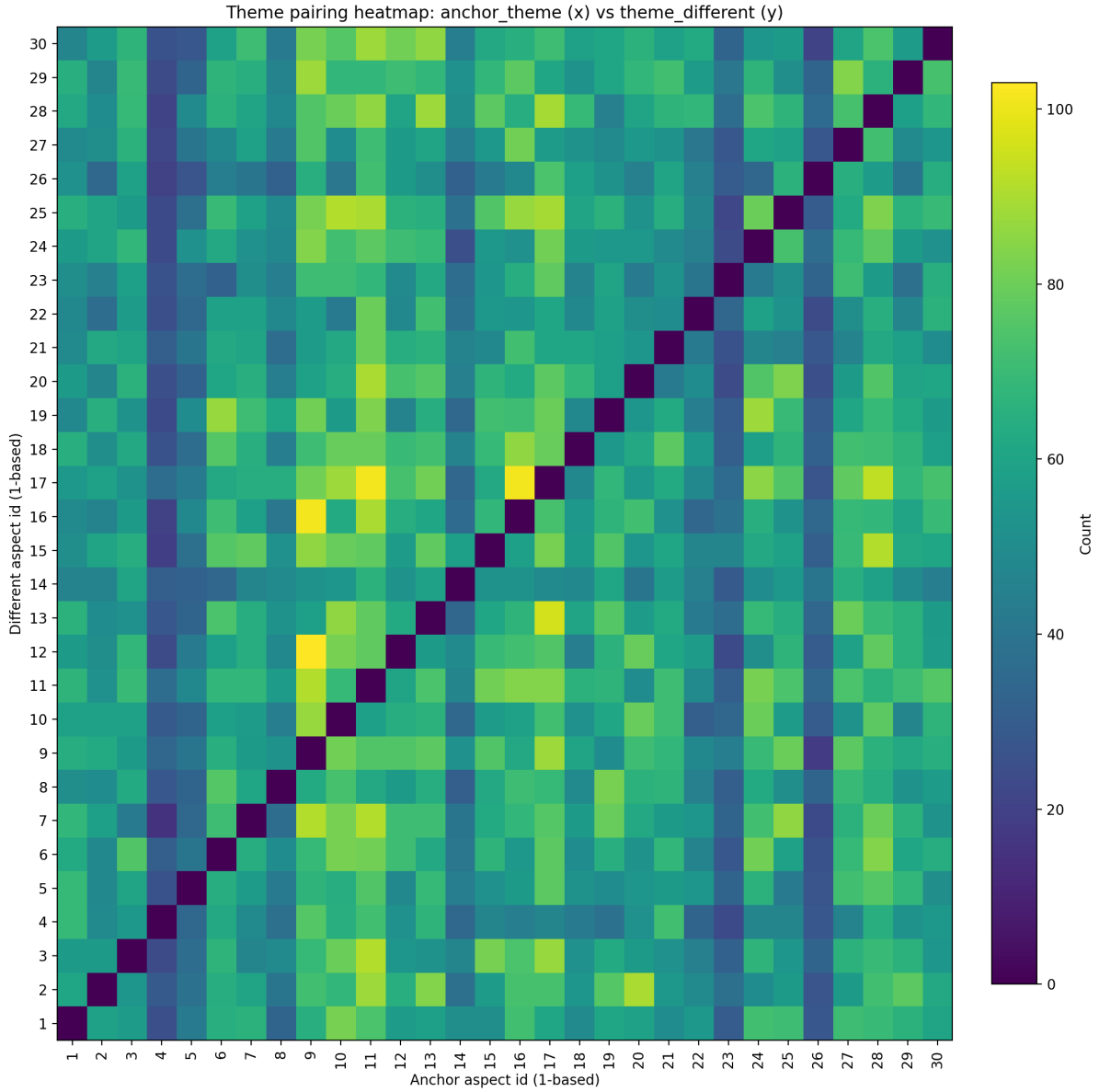

Figure 6: Theme pairing heatmap used for constructing negatives: anchor theme (x-axis) vs. different theme (y-axis). The diagonal is suppressed by design since negatives differ in the selected aspect.

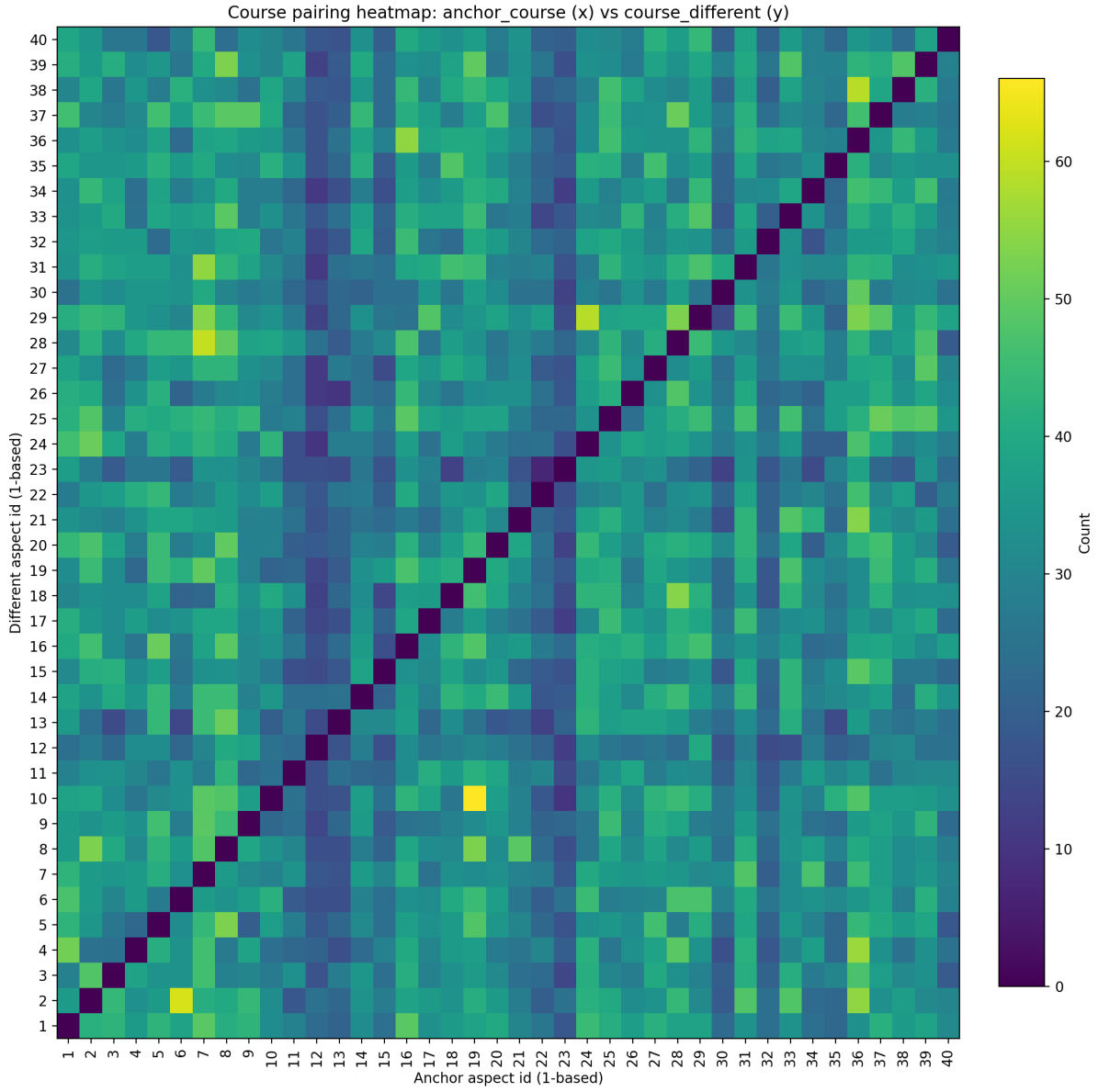

Figure 7: Course-of-action pairing heatmap for negatives: anchor course (x-axis) vs. different course (y-axis). Off-diagonal counts indicate the frequency of specific negative pairings.

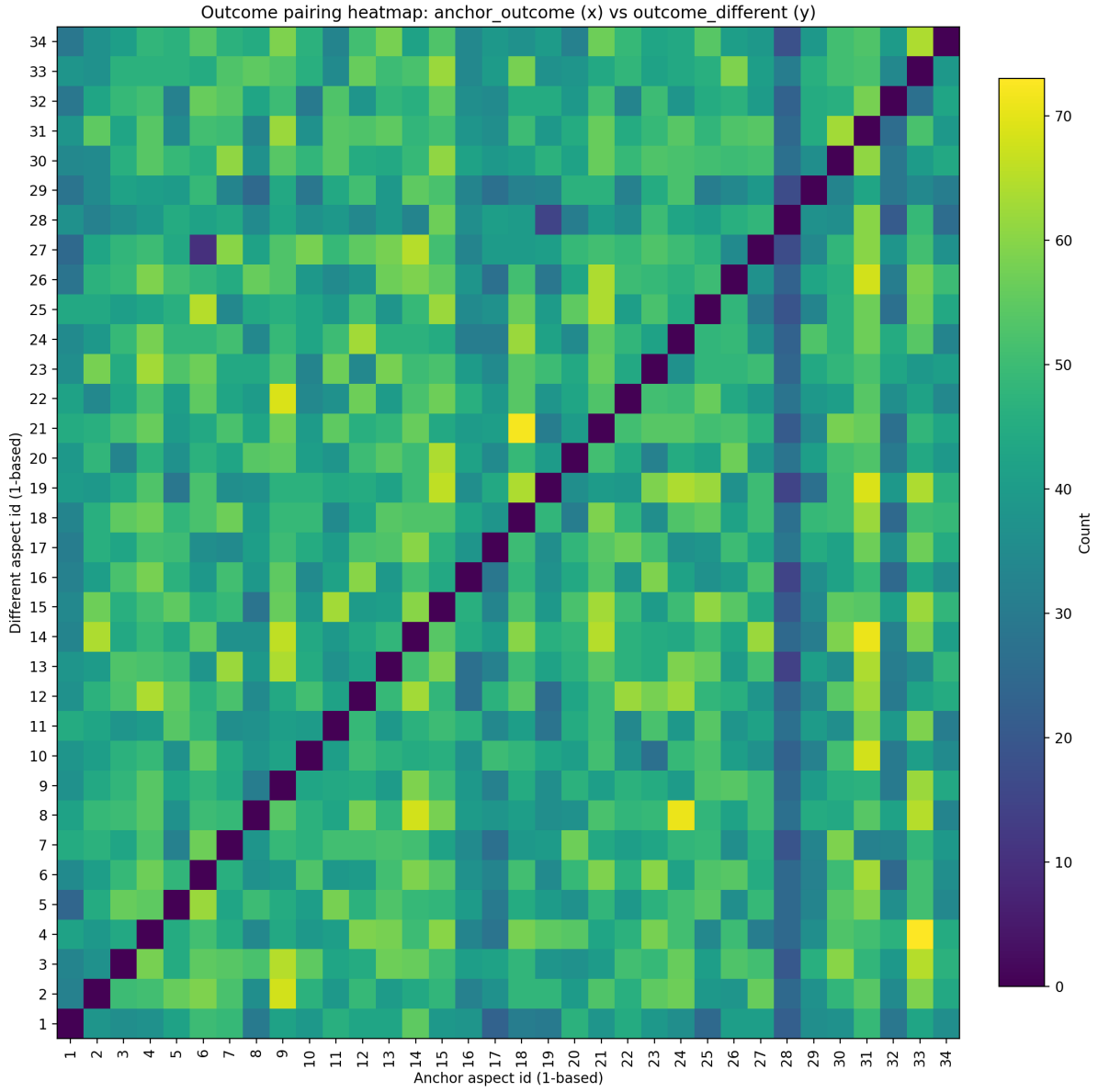

Figure 8: Outcome pairing heatmap for negatives: anchor outcome (x-axis) vs. different outcome (y-axis). The diagonal is suppressed by construction.

D Model's Architecture

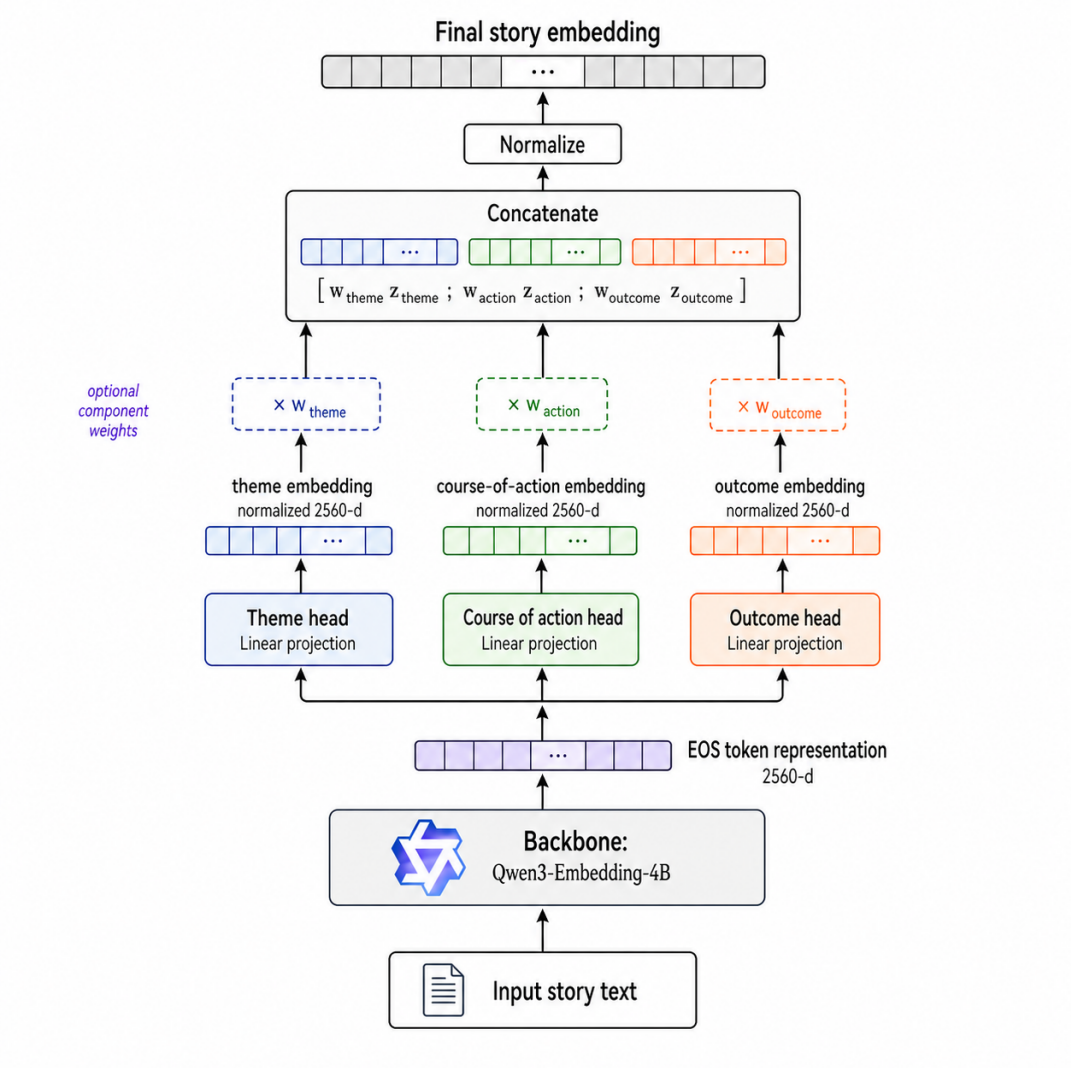

Figure 9: Proposed model's architecture. Backbone gets the story, then the EOS token is used as an input into the aspect heads. The heads produce aspect representations, which are optionally weighted. Concatenated normalization is the final story embedding.

E Retrieval Dynamics During Training

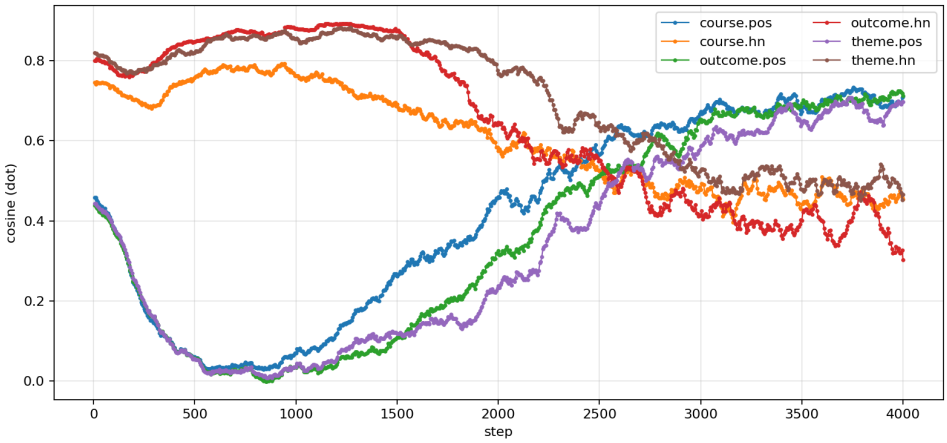

Figure 10: Graph plotting similarities of (anchor, hard positive) and (anchor, hard negative).