YNU-HPCC at SemEval-2026 Task 4: Narrative Similarity via Multi-Perspective E5-Mistral and Embedding Routing

Feiyang Song, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
fysong@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This paper presents the system developed by the YNU-HPCC team for SemEval-2026 Task 4: Narrative Story Similarity and Narrative Representation Learning. The task challenges computational systems to identify narrative similarity across three orthogonal dimensions: abstract theme, course of action, and outcomes. The primary scientific difficulty lies in distinguishing the underlying structural fabula from surface-level lexical overlaps, particularly when facing long-context narratives with subtle plot twists. To address this, our approach employs a hybrid architecture that strategically decouples retrieval and ranking tasks. For Track A, we introduce a dynamic routing mechanism where an instruction-tuned E5-Mistral-7B model handles clear cases, while ambiguous hard samples are routed to a Gemini-3-Flash reasoner. For Track B, we leverage the global semantic modeling capabilities of Gemini-Embedding-001 via a structure-preserving chunking strategy, enhanced by All-But-The-Top (ABTT) during inference. Extensive experiments on the official test set show that this divide-and-conquer strategy effectively balances local instruction following with global open-domain generalization. Our system performs competitively, ranking 5th in Track A and 2nd in Track B among all participating teams.

1 Introduction

Narrative similarity refers to the resemblance between stories in terms of their underlying events and causal relations, distinct from surface-level textual similarity (Waight et al., 2025). As defined in the task guidelines, two texts can be considered to tell the same story even when they differ drastically in wording, characters, or setting, provided they share the same abstract themes, course of action, and outcomes (Hatzel et al., 2026). SemEval-2026 Task 4 formalizes this challenge by asking models to identify narrative similarity across these three orthogonal dimensions while explicitly ignoring concrete settings and character names.

The primary scientific difficulty is that standard dense retrieval models often conflate semantic textual similarity (STS) with narrative structural alignment. Research on semantic textual relatedness (STR) has highlighted that models typically struggle with hard samples, which are pairs characterized by high lexical overlap but divergent semantic logic (Li et al., 2024). For instance, two stories may share an identical course of action, such as a war, but diverge completely in their outcomes, ranging from success to failure. Conventional bi-encoders are efficient but often miss these fine-grained structural divergences.

To address these challenges, this paper presents a system, *Multi-Perspective E5-Mistral with Embedding Routing* (Wang et al., 2024). Drawing on the efficacy of ensemble strategies in recent cross-lingual retrieval tasks (Mao et al., 2025), we employ the instruction-tuned E5-Mistral-7B model implemented via the *sentence-transformers* framework (Reimers and Gurevych, 2019). The approach proceeds in two main steps:

- **Multi-perspective representation:** Instead of encoding a single summary, the system constructs distinct narrative perspectives. These include a raw view of themes, an action view with extracted verbs for plot dynamics, and an outcome view of resolutions. These views are explicitly encoded to model the task's orthogonal dimensions.
- **Embedding routing mechanism:** We implement a dynamic routing strategy that directs ambiguous triplets to a generative LLM for refined reasoning. This approach leverages chain-of-thought prompting to resolve conceptually distinct pairs with high overlap, addressing the limitations of pure embedding-based retrieval.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 440–445

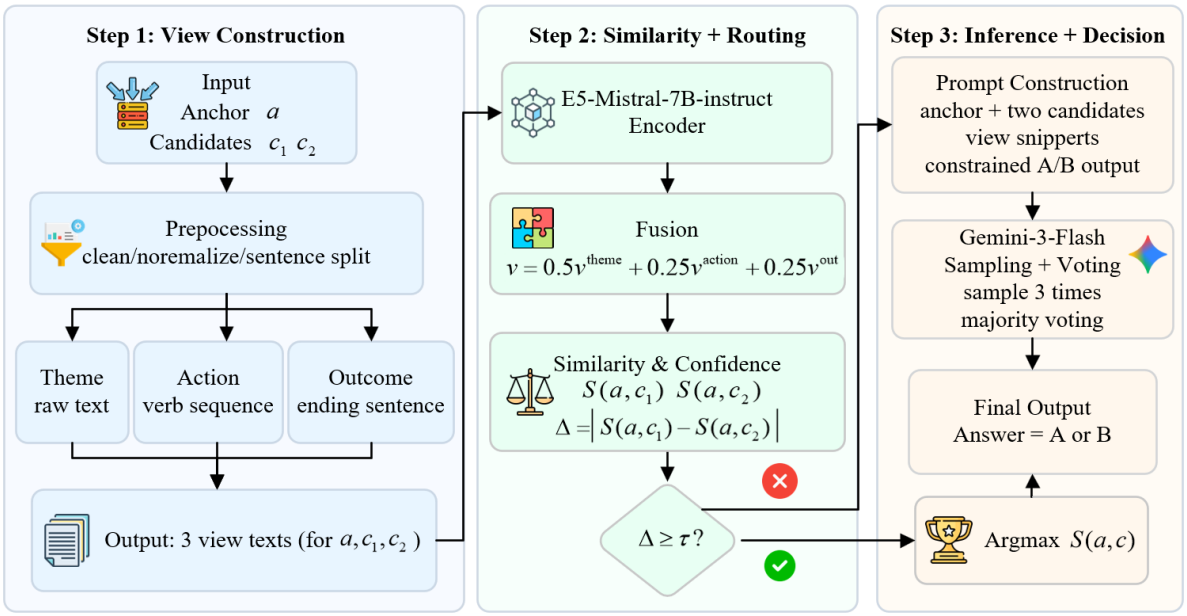

Figure 1: The overview of the system architecture

Extensive experiments on the official datasets demonstrate that our system effectively captures long-range narrative dependencies. By integrating the All-But-The-Top (ABTT) method (Mu et al., 2018) and instruction tuning, our method achieves 0.7425 in Track A (ranking 5th) and 0.7125 in Track B (ranking 2nd).

2 System Overview

Figure 1 illustrates the overall architecture of our proposed system. The pipeline consists of deterministic view construction, similarity routing, and generative decision-making.

2.1 Narrative Encoding

E5 multi-perspective stream. For each story x, we construct three deterministic views: theme (raw text), action (verb sequence), and outcome (ending sentences). E5-Mistral-7B-Instruct encodes each view with an instruction prefix I_k:

$$\mathbf{v}^k = \text{Enc}_{E5}\left(I_k \oplus x^k\right), \quad k \in \{\text{theme}, \text{action}, \text{out}\}. \tag{1}$$

We then fuse the three view embeddings into a single routing representation.

$$\mathbf{v} = 0.5\,\mathbf{v}^{\text{theme}} + 0.25\,\mathbf{v}^{\text{action}} + 0.25\,\mathbf{v}^{\text{out}}. \tag{2}$$

The encoder is fine-tuned with Multiple Negatives Ranking Loss (MNRL) to ensure that anchor stories are closer to their positive counterparts. Meanwhile, other instances in the same batch serve as negatives.

Gemini long-context stream. To capture global dependencies in long narratives, we encode stories with Gemini-Embedding-001 using a head-body-tail strategy that concatenates the first three sentences, three uniformly sampled middle sentences, and the last three sentences. Gemini embeddings are used for Track B retrieval.

2.2 Similarity Scoring and Confidence Margin

Given an anchor story a and a candidate story c, we compute cosine similarity in the routing space:

$$S(a,c) = \frac{\mathbf{v}_a \cdot \mathbf{v}_c}{\|\mathbf{v}_a\|\,\|\mathbf{v}_c\|}. \tag{3}$$

In Track A, each instance contains two candidates c_1 and c_2. We measure confidence by the absolute margin between the two similarity scores:

$$\Delta = |S(a,c_1) - S(a,c_2)|\,. \tag{4}$$

A large margin indicates that the embedding model strongly prefers one candidate and is likely correct. A small margin indicates that the two candidates are close in the embedding space, and the decision is uncertain.

2.3 Hybrid Inference and Routing in Track A

We route each triplet based on a threshold τ. If the margin is at least τ, we output the candidate with higher cosine similarity. Otherwise, we invoke a generative reasoner to disambiguate the pair. The

decision rule is

$$\widehat{y} = \begin{cases} \arg\max_{c \in \{c_1, c_2\}} S(a, c), & \Delta \geq \tau, \\ \text{LLM\_Reasoner}(a, c_1, c_2), & \Delta < \tau. \end{cases} \quad (5)$$

The threshold τ is calibrated on the development set to balance accuracy and cost. This routing strategy allows the system to handle clear cases efficiently while reserving reasoning for hard cases.

3 Experimental Results

3.1 Datasets and Data Augmentation

The experiments were conducted on the official SemEval-2026 Task 4 datasets. As shown in the provided data, the training set for Track A is structured as triplets comprising *anchor_text*, *text_a*, and *text_b*, where a boolean label *text_a_is_closer* indicates the ground truth. This format explicitly requires models to distinguish narrative proximity based on the dimensions of abstract theme, course of action, and outcomes. The Track B dataset consists of raw narrative texts for representation learning.

Data augmentation strategy: To address data scarcity, we explored multiple external narrative corpora. Initially, we experimented with ROCStories (Mostafazadeh et al., 2016) and the CMU Book Summary dataset (Bamman and Smith, 2013). However, preliminary experiments indicated that these general-domain datasets introduced significant domain shifts and failed to align with the task's specific fabula-centric definitions. Consequently, we adopted an LLM-based synthesis approach. Utilizing Gemini-2.5-Pro, we generated a large-scale synthetic corpus consisting of 10,000 narrative triplets. The generation process was strictly guided to adhere to the official annotation guidelines (Hatzel et al., 2026), resulting in hard negatives that share surface-level lexical overlap but differ in deep narrative structure. This synthetic dataset demonstrated superior domain alignment and was used for fine-tuning the final model.

3.2 Evaluation Metrics

According to the official evaluation scripts, both tracks are evaluated using Accuracy, though derived differently to suit their respective output formats:

Track A: For the triplet ranking task, we report the ranking accuracy. This metric measures the percentage of triplets in which the system's binary decision matches the ground truth, i.e., which text is closer to the anchor.

Track B: For the narrative representation task, the system uses embedding consistency accuracy rather than standard retrieval metrics such as MAP. For each triplet, we calculate the cosine similarity between the generated embeddings, denoted as $sim_A = \cos(\mathbf{v}_{anchor}, \mathbf{v}_A)$ and $sim_B = \cos(\mathbf{v}_{anchor}, \mathbf{v}_B)$. A prediction is considered correct if the embedding space preserves the narrative proximity relative to the ground truth, such that $sim_A > sim_B$ when text A is the closer match.

Model	Score
e5-mistral-7b-instruct	**0.670**
e5-large-v2	0.655
Qwen3-Embedding-4B	0.645
bge-large-en-v1.5	0.635
gte-large-en-v1.5	0.615
Qwen3-Embedding-8B	0.610
KaLM-Embedding-Gemma3-12B-2511	0.605
bge-m3	0.590
QZhou-Embedding	0.520

Table 1: Zero-shot performance of candidate embedding models on the Track A development set.

3.3 Model Selection

Our model selection process involved a multi-stage evaluation to identify the optimal components for both local routing and global retrieval.

Phase 1: Local backbone selection for Track A routing. We shortlisted four candidates following the MTEB leaderboard (Muennighoff et al., 2023): E5-Mistral-7B-Instruct (Wang et al., 2024), E5-Large-V2 (Wang et al., 2022), Qwen3-Embedding-4B (Team, 2025), and BGE-Large-EN-V1.5 (Xiao et al., 2024). We first conducted a comprehensive zero-shot evaluation on the development set, with results reported in Table 1. E5-Mistral-7B-Instruct achieved the highest initial score of 0.670 and exceeded representative baselines such as GTE-Large-EN-V1.5 (Zhang et al., 2024) with 0.615 and BGE-M3 (Chen et al., 2024) with 0.590. We then fine-tuned the top candidate on our augmented dataset and observed consistent gains on the development set, as shown in Table 2. Therefore, we selected E5-Mistral-7B-Instruct as the Track A gatekeeper.

Phase 2: Generative reasoner selection for Track A inference. For the generative reasoning module, we compared Gemini-3-Flash against Qwen3-Max and GPT-5.2. We evaluated their performance on the hard sample subset of the devel-

Model	Before	After
e5-mistral-7b-instruct	0.670	**0.705**
e5-large-v2	0.655	0.695
Qwen3-Embedding-4B	0.645	0.650
bge-large-en-v1.5	0.635	0.660

Table 2: Performance comparison of the selected backbone before and after fine-tuning on the development set.

Model	Score
Gemini-3-Flash	**0.730**
GPT-5.2	0.695
Qwen3-Max	0.675

Table 3: Evaluation of LLMs for the generative reasoning module on the development set.

opment set. As shown in Table 3, Gemini-3-Flash demonstrated the highest reasoning accuracy of 0.73 while maintaining a reasonable cost profile. Consequently, it was selected to handle the ambiguous triplets routed by the local model.

Phase 3: Final architecture decision for Track B. For the narrative retrieval task in Track B, we initially considered a fusion of the fine-tuned E5-Mistral and the Gemini-Embedding-001 API. However, comparative experiments on the test set revealed that the standalone Gemini-Embedding-001 achieved better generalization on open-domain narratives than the fine-tuned local model. Therefore, to streamline the architecture, we used Gemini-Embedding-001 exclusively for Track B, while retaining the fine-tuned E5-Mistral for the Track A routing mechanism.

3.4 Implementation Details

3.4.1 Local model fine-tuning for routing in Track A

We fine-tuned it using LoRA (Hu et al., 2022) ($r = 32$, $\alpha = 64$, target modules: *q_proj*, *v_proj*) on the augmented dataset. Training was conducted for 2 epochs with a learning rate of 3×10^{-7} and an effective batch size of 48, optimizing the MNRL (Henderson et al., 2017). We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a linear warmup.

To explicitly model the narrative dimensions during the routing phase, we applied deterministic view construction and fusion:

Action view: We derive an action-centric view by extracting a compact verb sequence from each narrative using a lightweight rule-based heuristic. We retain verb-like tokens that describe events and filter out function words and punctuation, yielding a sequence that emphasizes plot dynamics.

Outcome view: Constructed by truncating the narrative to its final two sentences to represent the resolution.

Fusion: The final routing embedding was derived as a weighted average of the theme, action, and outcome views, with weights of 0.5, 0.25, and 0.25, respectively.

3.4.2 Generative inference configuration in Track A

For hard samples routed by the local model, where $\Delta < 0.072$, we deployed Gemini-3-Flash. The temperature was set to 1.0 to facilitate diverse reasoning within the chain-of-thought process, with a maximum output token count of 1024. To mitigate the variance introduced by the high temperature and ensure decision robustness, we employed a triple sampling strategy. Specifically, we generated three independent reasoning paths for each query and used majority voting to determine the final classification of A or B. This approach effectively aggregates diverse reasoning patterns while filtering out stochastic anomalies.

3.4.3 Retrieval configuration in Track B

We utilized the Gemini-Embedding-001 model via the Google GenAI API. To align with the task objective, we applied the head-body-tail chunking strategy and prepended the instruction *represent this narrative story to each input for similarity comparison*. The API was configured with the task type set to *SEMANTIC_SIMILARITY* and an output dimension of 768. Post-inference, we implemented the ABTT method, removing the first principal component from the batch to mitigate common frequency biases in the embedding space.

3.5 Parameters Fine-tuning

A critical hyperparameter in our Track A system is the routing threshold τ, which dictates the trade-off between the efficient dense retriever and the accurate but costly generative reasoner. We performed a grid search on the development set, varying τ from 0.0 to 0.1 with a step size of 0.002.

As illustrated in Figure 2, the optimal performance on the development set was achieved at $\tau = 0.072$. We observed that setting the threshold too low ($\tau < 0.05$) forced the system to rely excessively on the local dense retriever. Since the

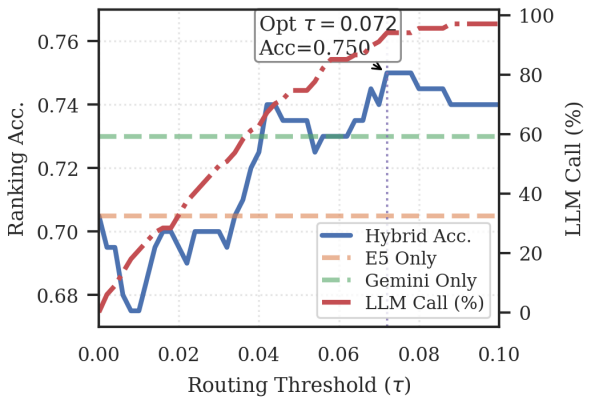

Figure 2: Impact of the routing threshold τ on ranking accuracy and LLM inference rate.

Model	Track A	Track B
Baselines (Official)		
Random Baseline	0.5000	0.5000
Jaccard Similarity	0.5625	–
GPT-4o-mini	0.6700	–
all-MiniLM-L6-v2	–	0.5850
story-emb	–	0.6325
Ours		
e5-large-v2	0.6650	0.6450
bge-large-en-v1.5	0.6300	0.6225
e5-mistral-7b-instruct	0.6825	0.6600
Gemini-3-Flash	0.7300	–
Gemini-Embedding-001	–	**0.7125**
Gemini-3-Flash+e5-mistral-7b	**0.7425**	–

Table 4: Official Test Set Results. Comparison of system performance, measured by Accuracy, on both tracks.

instruction-tuned E5 model achieved a standalone accuracy of 0.705—lower than the Gemini-3-Flash baseline of 0.730—this conservative routing strategy failed to correct errors in subtle hard samples.

By setting the threshold to 0.072, the system adopted a more aggressive routing strategy, delegating a larger proportion of ambiguous pairs (with narrow confidence margins) to the generative reasoner. This configuration successfully leveraged the LLM's superior reasoning capabilities for high-uncertainty cases while still handling clear-cut samples with E5. Consequently, the hybrid system achieved a peak accuracy of 0.750, surpassing both the standalone local model and the pure LLM approach.

3.6 Comparative Results

The performance of our proposed system on the official SemEval-2026 test set is presented in Table 4.

Track A: Our hybrid routing strategy achieved a ranking accuracy of 0.7425, significantly outperforming the standalone E5-Mistral model. This confirms that the generative reasoner effectively resolves ambiguity in hard samples.

Track B: The standalone Gemini-Embedding-001 model attained an embedding accuracy of 0.7125. This performance indicates a stronger ability to abstract narrative structure compared to local fine-tuned models, confirming that large-scale model capacity is a decisive factor for the representation task.

3.7 Discussion

Analysis of Track A results reveals that dynamic routing is critical for resolving hard samples characterized by high lexical overlap but divergent outcomes, where Gemini-3-Flash's causal reasoning effectively mitigates the dense retriever's bag-of-words bias. In Track B, the superior performance of Gemini-Embedding-001 over the fine-tuned E5-Mistral suggests that for open-ended narrative retrieval, model capacity and world knowledge play a more decisive role than task-specific fine-tuning alone. Additionally, ablation studies confirm that fine-tuning on synthetic triplets improved zero-shot performance by approximately 4%, validating our dimension-aligned training strategy.

4 Conclusion

This paper presents a unified framework for SemEval-2026 Task 4 that bridges surface-level textual similarity and deep narrative alignment. Our system decouples retrieval and ranking to address the distinct challenges of the two tracks. For Track B, we adopt Gemini-Embedding-001 as a standalone encoder and observe stronger open-domain generalization than fine-tuned local models. For Track A, a confidence-based routing strategy is used to distinguish hard negatives with high lexical overlap by delegating ambiguous cases to a Gemini-3-Flash reasoner when the similarity margin falls below a calibrated threshold. This hybrid design combines the efficiency of dense retrieval with the robustness of LLM reasoning. Future work will explore knowledge distillation to transfer the reasoning capability of Gemini-3-Flash into smaller local bi-encoders, reducing reliance on external APIs while preserving fine-grained narrative discrimination.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos.61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

David Bamman and Noah A. Smith. 2013. New alignment methods for discriminative book summarization. *CoRR*, abs/1305.1319.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Natalia Fedorova, Evelyn Gius, and Chris Biemann. 2026. Narrative similarity-annotation guidelines. Technical report.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Weijie Li, Jin Wang, and Xuejie Zhang. 2024. YNU-HPCC at SemEval-2024 task 1: Self-instruction learning with black-box optimization for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 792–799, Mexico City, Mexico. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Yuheng Mao, Jin Wang, and Xuejie Zhang. 2025. YNU-HPCC at SemEval-2025 task 7: Multilingual and cross-lingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 234–240, Vienna, Austria. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Qwen Team. 2025. Qwen3 technical report. Technical report.

Hannah Waight, Sol Messing, Anton Shirikov, Margaret E Roberts, Jonathan Nagler, Jason Greenfield, Megan A Brown, Kevin Aslett, and Joshua A Tucker. 2025. Quantifying narrative similarity across languages. Technical report.

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. *CoRR*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649. Association for Computing Machinery, Inc.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.