TeleAI at SemEval-2026 Task 4: Few-Shot Narrative Similarity Modeling for Classification and Ranking

Weiwei Fu[1]*, Shiquan Wang[1], Ruiyu Fang[1], Shuangyong Song[1]
[1]Xingchen AGI Lab, China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd
fuweiwei@chinatelecom.cn, wangsq23@chinatelecom.cn, fangry@chinatelecom.cn, songshy@chinatelecom.cn

Abstract

This paper presents a unified, task-adaptive modeling framework for the two tracks of SemEval-2026 Task 4: Narrative Similarity. For Track A, we build a three-stage pipeline of three-dimensional narrative-anchored chain-of-thought (CoT) reasoning, multi-view data augmentation, and Low-Rank Adaptation (LoRA) fine-tuning. For Track B, we design an architecture fully aligned with the ranking inference pipeline and task objective, along with corresponding data augmentation and expansion methods, and propose Smooth Cosine Contrastive Loss (SCCL) to stabilize training in low-resource settings. Systematic experiments verify the effectiveness of each core module, and our final systems rank 4th in both tracks, providing a reproducible technical solution for few-shot similarity modeling.

1 Introduction

1.1 Research Background and Task Overview

Narrative text similarity calculation is a core fundamental task in natural language understanding. It has important application value in scenarios such as long-text retrieval, content recommendation, and copyright detection (Chaturvedi et al., 2018; Hatzel and Biemann, 2024). This paper conducts research on the two parallel tracks of SemEval-2026 Task 4 (Hatzel et al., 2026). Both share the core goal of narrative similarity evaluation, but there are significant differences in task form and constraints:

- **Track A** is an extreme few-shot relative similarity binary classification task: Given an anchor text and two candidate texts, the model needs to judge which candidate has higher narrative similarity with the anchor text. Only 249 annotated triplets are provided as training data.
- **Track B** is a pairwise ranking and embedding generation task: The model needs to generate narrative-aware vector representations for input texts and complete ranking based on embedding cosine similarity; the supervision signal can only use the 249 annotated triplets from Track A.

1.2 Core Challenges

This task faces multiple technical challenges: The extreme few-shot constraint composed of only 249 annotated triplets is prone to severe overfitting (Hu et al., 2022; Pletenev et al., 2025); Track A is prone to fall into surface-level semantic fitting and is difficult to capture the deep narrative alignment relationship between texts; in Track B, the end-to-end classification architecture does not match the ranking objective, requiring a redesign of the model structure and task flow.

1.3 Main Contributions of This Paper

1. For Track A: We propose a multi-view data augmentation scheme with dimension-guided chain-of-thought reasoning, which alleviates model overfitting in extreme few-shot scenarios.
2. For Track B: We design a ranking-aligned independent encoding architecture and data expansion method, and propose Smooth Cosine Contrastive Loss (SCCL) to solve unstable training and overfitting in few-shot ranking tasks.

2 Related Work

2.1 Narrative Similarity and Text Representation Learning

Existing research on narrative similarity has explored story-level semantic alignment (Chaturvedi

** Corresponding Author.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1228–1236

et al., 2018) and narrative-focused text representation learning (Hatzel and Biemann, 2024; Subbiah et al., 2024), laying the foundation for our dimension-anchored evaluation criteria. For extreme few-shot scenarios, Chain-of-Thought prompting (Wei et al., 2022) enables interpretable stepwise reasoning for LLMs (Jin et al., 2024), while Low-Rank Adaptation (Hu et al., 2022) achieves parameter-efficient fine-tuning with reduced overfitting in low-resource settings (Pletenev et al., 2025). Text embedding optimization works (Wang et al., 2024a) also provide core reference for our ranking-oriented embedding modeling.

2.2 Large Language Model Pre-training and Alignment

Our work is also built on rich practices in bilingual large language model (LLM) pre-training and alignment. The TeleChat series models (He et al., 2024; Wang et al., 2024b, 2025), developed by our team, have achieved leading performance in Chinese-English bilingual understanding, long-context modeling, and step-by-step reasoning through full-process pre-training, supervised fine-tuning, and reinforcement learning from human feedback. Meanwhile, our team has accumulated extensive experience in large-scale LLM training, including the 52B-1T parameter TeleFLM series (Li et al., 2024b,a) and the domestically trained TeleChat3-MoE model (Liu et al., 2025), which provides solid technical support for our LLM-based data augmentation and reasoning pipeline in this work.

2.3 Text Classification and Semantic Matching

The core tasks of this paper, classification and ranking of narrative similarity, are essentially rooted in fine-grained text semantic matching and classification. Our team has long been committed to related research: we have proposed a series of advanced frameworks for hierarchical text classification (Zhao et al., 2022; Ning et al., 2023; Xiong et al., 2024), dialogue intent classification (Song et al., 2017; Pang et al., 2022; Xu et al., 2023), user satisfaction prediction (Yao et al., 2020; Sun et al., 2022), and retrieval-based question answering (Song et al., 2020). These works have explored multi-view feature fusion, prompt tuning, contrastive learning, and graph-based semantic modeling, which lay a solid foundation for our dimension-anchored reasoning framework and contrastive loss design in this paper.

Different from prior works focusing on single classification or ranking task settings, our framework addresses both tracks of the task under extreme few-shot constraints, with targeted designs for core pain points including narrative alignment, training-inference consistency and overfitting mitigation.

3 Methodology for Track A: Narrative Similarity Classification

The overall framework is based on explicit narrative dimension criteria, expands training samples through multi-view data augmentation (DA) without data leakage, and finally conducts lightweight Low-Rank Adaptation (LoRA) parameter-efficient fine-tuning based on the Qwen3-30B-A3B-Instruct model. Hyperparameter optimization and model effectiveness verification are completed via 5-fold stratified cross-validation, effectively mitigating overfitting in the extreme few-shot setting.

3.1 Narrative Dimension Criteria and Instruction Template Construction

We anchor the core evaluation dimensions of narrative similarity to three key levels in accordance with the official task description, serving as the fundamental basis for model inference (Li et al., 2025; Hatzel and Biemann, 2024):

1. Abstract Theme: The core conflict and central thesis of a story.

2. Plot Development: The sequence of events, character behaviors, conflict evolution, and temporal logic of a story.

3. Story Ending: The plot trajectory at the end of the text, including the resolution of core conflicts and the final fate of characters.

Meanwhile, we explicitly define irrelevant factors that must be completely ignored in similarity judgment, including writing style, specific background setting, names of characters and locations, text length, and richness of narrative details, to prevent the model from being disturbed by non-narrative noise. Based on the above criteria, we optimized a structured instruction template through multiple rounds of iterative ablation on the training split, strictly following the three narrative dimensions

and excluding non-narrative noise. The design of the dimension-anchored template also draws on our team's previous experience in prompt tuning for fine-grained text classification (Xiong et al., 2024).

3.2 Multi-View Data Augmentation Scheme

We designed a multi-view DA scheme with zero data leakage. All augmentation operations are performed exclusively within the training subset of 5-fold stratified cross-validation, and the validation subset is not involved in any augmentation, hyperparameter selection, or model training. In each fold of cross-validation, we fixedly split the original 249 labeled triplets into 220 training samples and 29 validation samples. All augmentation operations are generated based only on the 220 training samples. Data augmentation is divided into two core stages:

1. Judgment Logic Reasoning Mining: Based on the optimized instruction template, we use Qwen3-235B-A22B-Instruct to complete inference on the 220 training triplets, outputting the reasoning analysis for the relative similarity judgment of each sample, clarifying the core consideration dimensions of narrative similarity without including the final judgment result. This stage generates Chinese-English bilingual reasoning analysis samples, totaling 220×2.

2. Multi-Task Sample Generation: We generate 4 types of complementary training tasks based on the original triplets and reasoning results, all strictly preserving the original annotation labels:

   (a) Similarity Judgment Task 1: Input is original English triplets and their Chinese literal translations, label is the relative similarity judgment result, generating 220×2 samples;

   (b) Similarity Judgment Task 2: Input is Chinese literal translations of triplets, label is the relative similarity judgment result and corresponding CoT reasoning process, generating 220×2 samples;

   (c) Summarization Task 1: Input is English original or Chinese literal translation of each single text in the triplet, label is the narrative summary generated based on core narrative dimensions, generating Chinese-English bilingual versions totaling $220 \times 3 \times 2$ samples;

   (d) Summarization Task 2: Input is English original or Chinese literal translation of each single text in the triplet, label is the focused narrative summary generated in combination with the relative similarity result, generating Chinese-English bilingual versions totaling $220 \times 3 \times 2$ samples.

Rationale for Augmentation Design: The multi-view multi-task scheme consolidates the model's decision logic via similarity judgment tasks, and anchors the core narrative dimensions via summarization tasks; the cross-lingual augmentation further reduces surface overfitting and leverages the bilingual capability of the base model. All designs strictly follow the zero data leakage principle, and jointly enhance the model's few-shot generalization performance.

Through the above scheme, each fold of cross-validation can obtain 3960 compliant training samples.

4 Methodology for Track B: Narrative Similarity Ranking and Embedding

To eliminate the misalignment between training and inference objectives, we adopt an independent encoding architecture, stabilize low-resource pairwise ranking training via our custom loss function, and mitigate few-shot overfitting via multi-view data augmentation.Our design also draws on our team's early practice in text classification and ranking tasks (Song and Meng, 2015).

4.1 Ranking Objective-Aligned Modeling Architecture

We adopt an Independent Encoding Architecture (IEA) with cosine similarity ranking, fully aligned with the task core:

1. Independent Input: Anchor text A, positive candidate P, and negative candidate N are fed into the model separately, with no concatenation, matching the single-text input mode during inference (Wang et al., 2024a).

2. Embedding Generation: The model outputs sentence embeddings $f(A), f(P), f(N)$ and applies L2 normalization.

3. Similarity Calculation: Compute cosine similarities $\cos_{AP} = \cos(f(A), f(P))$ and $\cos_{AN} = \cos(f(A), f(N))$, then derive the similarity difference $\Delta = \cos_{AP} - \cos_{AN}$.

4. Loss Optimization: Maximize Δ to ensure ranking accuracy.

To address gradient vanishing, hard sample bias, and overfitting of vanilla triplet loss in extreme few-shot settings, we propose Smooth Cosine Contrastive Loss (SCCL) with two core components. The design of this contrastive loss draws on our team's previous research on contrastive learning for fine-grained text classification and semantic matching (Song et al., 2020; Xiong et al., 2024):

- Label Smoothing: Define binary label $y \in \{0,1\}$ ($y = 1$ means P is more narratively similar to A than N), and map hard labels to soft labels $\hat{y} = y \cdot (1 - 2\beta) + \beta$, where $\beta \in (0, 0.5)$ is the label smoothing coefficient.

- Soft Boundary Loss: Use a log soft boundary loss with no gradient truncation. The final SCCL loss is defined as:

$$L_{SCCL} = \log\Big(1 + \exp\big(-(2\hat{y} - 1) \cdot (\Delta - m)\big)\Big) \quad (1)$$

  where m is the margin hyperparameter. Batch-level loss is the average of all sample losses in the batch.

4.2 Training Data Construction

Supervised training signals are exclusively from the 249 labeled triplets provided by Track A. To address limited labeled samples, we perform multi-view Narrative-Aware Preprocessing (NAP) and data augmentation on original triplets, strictly preserving the original positive-negative relative ranking. Four complementary views are applied, generating 4 augmented triplets per original triplet, expanding the training data size to 5 times the original:

1. Full Chinese-English translation: Translate all original English texts into Chinese.

2. Three-element narrative-aware summary: Generate summaries around abstract theme, plot development, and story ending.

3. Plot-focused summary: Generate summaries only around plot development and story ending.

4. Theme-focused summary: Generate summaries only around abstract theme and core conflict.

LoRA fine-tuning on gte-qwen2-7b and final model training is conducted on the full augmented training set.

5 Experiments and Results Analysis for Track A

This section conducts systematic comparative experiments and ablation studies for Track A to verify the effectiveness of each core module in our proposed framework, with results summarized in Table 1 and Table 2.

5.1 Experimental Setup

5.1.1 Dataset and Evaluation Metric

We use the official 249 labeled triplets as the training set. All offline validation results are the average accuracy of 5-fold stratified cross-validation (CV) to ensure statistical reliability and reduce variance from small sample size. We set temperature = 0.2, and other inference hyperparameters are set differentially to adapt to the ensemble voting of models from 5-fold CV. All LoRA fine-tuning experiments uniformly adopt the AdamW optimizer.

5.1.2 Base Model Setting

All experiments and baselines for Track A are based on the unified base model Qwen3-30B-A3B-Instruct

5.2 Effectiveness of Core Modules

We design controlled experiments to verify the performance gain of each core module. The results are summarized in Table 1. From the results (Table 1), we can see CoT reasoning based on dimension criteria significantly improves the model's judgment accuracy. Multi-view data augmentation further alleviates overfitting in the few-shot scenario.

Setup	Val Acc
BL-Vanilla	0.552
BL+CoT	0.618
BL+CoT+DA	0.672
Full-PL	0.718

Table 1: Val Acc of Core Modules for Track A

LoRA fine-tuning achieves validation accuracy improved by 0.166 compared to the original baseline.

5.3 Ablation Study for Track A

To verify the necessity of each core module, we conduct single-variable ablation studies. The transposed results are shown in Table 2. The ablation study results (Table 2) show that all core modules proposed in this paper bring significant positive performance gains to the model, verifying the necessity and effectiveness of each module.

Module	1	2	3	4
CoT	no	yes	yes	yes
DA	no	no	yes	yes
LoRA	no	no	no	yes
CV Acc	0.552	0.618	0.672	0.718

Table 2: Ablation Study Results for Track A

5.4 Results Analysis

Ablation and controlled experiments consistently show that all core modules proposed for Track A in this paper bring significant positive performance gains: CoT reasoning based on the three-dimensional narrative criteria establishes interpretable quantitative evaluation criteria for narrative similarity, addresses the inherent ambiguity of similarity definition in the task, and serves as the core foundation for model performance improvement; the multi-view data augmentation scheme expands effective training samples under the constraint of zero data leakage, and significantly alleviates model overfitting in the extreme few-shot scenario; LoRA fine-tuning achieves efficient adaptation to task-specific features while preserving the general semantic capability of the pre-trained model, and realizes the final performance breakthrough (Hu et al., 2022; Pletenev et al., 2025). The final system trained with the optimal hyperparameter combination achieves 0.748 classification accuracy on the official Track A test set, ranking 4th in the track.

6 Experiments and Results Analysis for Track B

All experiments strictly follow the official task guidelines, with no external labeled narrative datasets used, and are conducted on the gte-qwen2-7b base model, with results summarized in Table 3, Table 4 and Table 5.

6.1 Experimental Setup

The supervised training signal comes exclusively from the 249 labeled triplets provided by Track A. The final evaluation is completed via online submission on the official held-out test set with 849 unlabeled single texts. The official evaluation metric is ranking accuracy, defined as the proportion of triplets where the model-predicted relative order of cosine similarity matches human annotations. All offline validation results are the average accuracy of 5-fold stratified CV.

6.2 Effectiveness of Core Modules

We verify the performance gain of each core module through controlled experiments. The results are summarized in Table 3.

Setup	Val Acc
BL-Vanilla	0.641
BL+NAP	0.697
BL+NAP+SCCL	0.751
BL+NAP+SCCL+DA	0.773
Full-PL	0.784

Table 3: Effectiveness of Core Modules for Track B

6.3 Comparison of Modeling Architectures and Loss Functions

This group of experiments is conducted as the setting of narrative-aware preprocessing without data augmentation, comparing the performance of different modeling architectures and loss functions. We compare three mainstream setups: 1) End-to-End Classification Architecture with Cross-Entropy Loss (E2E-CA-CE), the conventional pairwise classification baseline; 2) Independent Encoding Architecture with Basic Contrastive Loss (IEA-BCL), the standard triplet contrastive loss for text ranking and embedding tasks (Wang et al., 2024a); 3) our proposed Independent Encoding Architecture with SCCL (IEA-SCCL). The results are shown in Table 4. This group of

Setup	Val Acc
E2E-CA-CE	0.615
IEA-BCL	0.724
IEA-SCCL	0.751

Table 4: Comparison of Modeling Architectures and Loss Functions for Track B

experiments fixes the SCCL loss hyperparameters to $\beta = 0.08$ and $m = 0.1$, and conducts a grid search on the core LoRA hyperparameters, and determines the final optimal LoRA hyperparameter combination as: rank $r = 4$, learning rate 1×10^{-5}.

6.4 Ablation Study

We verify the necessity of each core module through univariate ablation experiments, fixing all other settings and removing one core module at a time. The transposed results are shown in Table 5.

Module	1	2	3	4
NAP	no	yes	yes	yes
SCCL	no	no	yes	yes
DA	no	no	no	yes
CV Acc	0.641	0.697	0.751	0.784

Table 5: Ablation Study Results for Track B

6.5 Results Analysis

Ablation and controlled experiments consistently show that all core modules proposed in this paper bring significant positive performance gains: narrative-aware preprocessing effectively filters non-narrative noise and is the foundation of performance improvement; SCCL loss solves the unstable training problem of ranking tasks in low-resource scenarios through label smoothing and soft boundary optimization; multi-view data augmentation further alleviates few-shot overfitting. The final system trained with optimal hyperparameters achieves 0.695 ranking accuracy on the official Track B test set, ranking 4th in the track.

7 Conclusion

This paper presents a unified, task-adaptive framework for both tracks of SemEval-2026 Task 4: Narrative Similarity. For Track A, we build a three-stage pipeline of dimension-aware reasoning, multi-view data augmentation, and lightweight LoRA fine-tuning to address extreme few-shot classification challenges. For Track B, we design a ranking-aligned independent encoding architecture with a novel SCCL, ensuring full consistency between training and inference pipelines. Systematic experiments verify the effectiveness of our core modules, with our final systems ranking 4th in both tracks, providing a reproducible, task-adaptive technical reference for low-resource narrative similarity modeling and representation learning.

8 Limitations

Our work has inherent limitations rooted in the task's extreme few-shot constraint:

1. **Generalization Deficiency**: Limited training data causes insufficient out-of-distribution generalization, with test set performance degradation from narrative distribution shift.

2. **Augmented Data Bias**: LLM-generated data has inherent distribution homogenization, amplifying the model's sensitivity to out-of-distribution samples.

3. **Lightweight Deployment Boundary**: Our framework adapts stably to lightweight models for the ranking track, but suffers significant performance loss on the classification track.

References

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have i heard this story before? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 673–678, New Orleans, Louisiana, USA. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Zhongjiang He, Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, and 1 others. 2024. Telechat technical report. *arXiv preprint arXiv:2401.03804*.

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 *International Conference on Learning Representations*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Baixuan Li, Yunlong Fan, and Zhiqiang Gao. 2025. Seaver: Attention reallocation for mitigating distractions in language models for conditional semantic textual similarity measurement. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 6789–6800. Association for Computational Linguistics.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, and 1 others. 2024a. 52b to 1t: Lessons learned via tele-flm series. *arXiv preprint arXiv:2407.02783*.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, and 1 others. 2024b. Tele-flm technical report. *arXiv preprint arXiv:2404.16645*.

Xinzhang Liu, Chao Wang, Zhihao Yang, Zhuo Jiang, Xuncheng Zhao, Haoran Wang, Lei Li, Dongdong He, Luobin Liu, Kaizhe Yuan, and 1 others. 2025. Training report of telechat3-moe. *arXiv preprint arXiv:2512.24157*.

Bo Ning, Deji Zhao, Xinjian Zhang, Chao Wang, and Shuangyong Song. 2023. Ump-mg: A uni-directed message-passing multi-label generation model for hierarchical text classification. *Data Science and Engineering*, 8(2):112–123.

Jinhui Pang, Huinan Xu, Shuangyong Song, Bo Zou, and Xiaodong He. 2022. Mfdg: A multi-factor dialogue graph model for dialogue intent classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 691–706. Springer.

Sergey Pletenev, Maria Marina, Daniil Moskovskiy, Vasily Konovalov, Pavel Braslavski, Alexander Panchenko, and Mikhail Salnikov. 2025. How much knowledge can you pack into a lora adapter without harming llm? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3210–3222. Association for Computational Linguistics.

Shuangyong Song, Haiqing Chen, and Zhiwei Shi. 2017. Intension classification of user queries in intelligent customer service system. In *2017 International Conference on Asian Language Processing (IALP)*, pages 83–86. IEEE.

Shuangyong Song and Yao Meng. 2015. Classifying and ranking microblogging hashtags with news categories. In *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*, pages 540–541. IEEE.

Shuangyong Song, Chao Wang, Haiqing Chen, and Huan Chen. 2020. Tcnn: Triple convolutional neural network models for retrieval-based question answering system in e-commerce. In *Companion Proceedings of the Web Conference 2020*, pages 844–845.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310.

Yang Sun, Liangqing Wu, Shuangyong Song, Xiaoguang Yu, Xiaodong He, and Guohong Fu. 2022. Tracking satisfaction states for customer satisfaction prediction in e-commerce service chatbots. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 616–625.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Zihan Wang, Xinzhang Liu, Yitong Yao, Chao Wang, Yu Zhao, Zhihao Yang, Wenmin Deng, Kaipeng Jia, Jiaxin Peng, Yuyao Huang, and 1 others. 2025. Technical report of telechat2, telechat2. 5 and t1. *arXiv preprint arXiv:2507.18013*.

Zihan Wang, Yitong Yao, Li Mengxiang, Zhongjiang He, Chao Wang, Shuangyong Song, and 1 others. 2024b. Telechat: An open-source billingual large language model. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 10–20.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Sishi Xiong, Yu Zhao, Jie Zhang, Li Mengxiang, Zhongjiang He, Xuelong Li, and Shuangyong Song. 2024. Dual prompt tuning based contrastive learning for hierarchical text classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12146–12158.

Huinan Xu, Jinhui Pang, Shuangyong Song, and Bo Zou. 2023. Improving dialogue intent classification with a knowledge-enhanced multifactor graph model (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16366–16367.

Riheng Yao, Shuangyong Song, Qiudan Li, Chao Wang, Huan Chen, Haiqing Chen, and Daniel Dajun Zeng. 2020. Session-level user satisfaction prediction for customer service chatbot in e-commerce (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13973–13974.

Deji Zhao, Bo Ning, Shuangyong Song, Chao Wang, Xiangyan Chen, Xiaoguang Yu, and Bo Zou. 2022. Tosa: A top-down tree structure awareness model for hierarchical text classification. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 23–37. Springer.

9 Appendix: Full Definition of Abbreviations

This section systematically lists the full definition and functional interpretation of all abbreviations used in this paper, categorized by general usage and track-specific content, to ensure clarity and avoid layout overflow.

9.1 General Abbreviations

CoT Chain-of-Thought, a step-by-step reasoning paradigm to enhance the interpretability and accuracy of large language model inference.

LoRA Low-Rank Adaptation, a parameter-efficient fine-tuning method for large pre-trained language models, which effectively mitigates overfitting in low-resource scenarios.

SCCL Smooth Cosine Contrastive Loss, the novel loss function proposed in this paper to stabilize training and alleviate overfitting for few-shot pairwise ranking tasks.

LLM Large Language Model, the pre-trained generative language model used for data augmentation and reasoning analysis in this work.

CV Cross-Validation, specifically 5-fold stratified cross-validation adopted in all offline experiments for hyperparameter tuning and module effectiveness verification.

Acc Accuracy, the core official evaluation metric for both tracks in this task.

LR Learning Rate, the core hyperparameter for model optimization during fine-tuning.

DA Data Augmentation, the multi-view in-distribution sample expansion scheme designed for the extreme few-shot setting in this task.

NAP Narrative-Aware Preprocessing, the preprocessing method to filter non-narrative noise and retain core narrative information of input texts.

OOD Out-of-Distribution, referring to scenarios where test data is outside the distribution of the training dataset.

IEA Independent Encoding Architecture, the model architecture adopted for Track B to ensure full consistency between training and inference pipelines.

9.2 Track A Specific Abbreviations

Setup Experiment Setup, indicating the specific configuration of each controlled experiment.

Train Acc Training Accuracy, the classification accuracy of the model on the training dataset.

Val Acc 5-Fold Validation Accuracy, the average classification accuracy on the validation subset of 5-fold stratified cross-validation.

BL-Vanilla Baseline model without CoT reasoning, data augmentation, or fine-tuning.

BL+CoT Baseline model equipped with dimension-guided CoT reasoning.

BL+CoT+DA Baseline model equipped with dimension-guided CoT reasoning and multi-view data augmentation.

Full-PL Full pipeline of our proposed framework, including all core modules and LoRA fine-tuning.

R LoRA Rank, the rank of the low-rank matrix in LoRA fine-tuning.

Scaling Factor, the hyperparameter to scale the output of the LoRA module during fine-tuning.

9.3 Track B Specific Abbreviations

Setup Experiment Setup, indicating the specific configuration of each controlled experiment.

Val Acc 5-Fold Cross-Validation Accuracy, the average ranking accuracy on the validation subset of 5-fold stratified cross-validation.

E2E-CA-CE End-to-End Classification Architecture with Cross-Entropy Loss, the conventional baseline architecture for Track B.

IEA-BCL Independent Encoding Architecture with Basic Contrastive Loss, the intermediate baseline for architecture and loss function comparison.

IEA-SCCL Independent Encoding Architecture with SCCL, our proposed architecture with the novel loss function.

BL-Vanilla Baseline model with raw text input and vanilla pre-trained model without any optimization modules.

BL+NAP Baseline model equipped with Narrative-Aware Preprocessing.

BL+NAP+SCCL Baseline model equipped with Narrative-Aware Preprocessing and SCCL.

BL+NAP+SCCL+DA Baseline model equipped with Narrative-Aware Preprocessing, SCCL, and multi-view data augmentation.

Full-PL Full pipeline of our proposed framework for Track B, including all core modules and optimal hyperparameter fine-tuning.

R LoRA Rank, the rank of the low-rank matrix in LoRA fine-tuning.