Narrative Nexus at SemEval-2026 Task 4: Modeling Narrative Similarity via Instruction-Based Fine-Tuning and Synthetic Data Augmentation

Haotan Guo[1], **Hongbin Na**[2], **Zimu Wang**[3], **Wei Wang**[3]
[1]School of Computer Science, The University of Sydney
[2]Australian AI Institute, University of Technology Sydney
[3]School of Advanced Technology, Xi'an Jiaotong-Liverpool University
hguo0293@uni.sydney.edu.au, hongbin.na@student.uts.edu.au
zimu.wang19@student.xjtlu.edu.cn, wei.wang03@xjtlu.edu.cn

Abstract

Narrative similarity assessment requires models to reason beyond surface-level lexical overlap and capture higher-level plot structures and thematic relationships. In this paper, we address SemEval-2026 Task 4 Track A: Narrative Story Similarity by reformulating it as an instruction-following generation problem. We employ parameter-efficient fine-tuning via LoRA to adapt pretrained large language models for triplet-based narrative comparison. To overcome the limitations imposed by the scarcity of human-annotated data, we further incorporate organizer-provided synthetic triplet samples generated by a large language model for data augmentation. Experimental results demonstrate that our fine-tuned Qwen2.5-7B model achieves slightly better performance than the zero-shot GPT-4o-mini baseline. These findings underscore the effectiveness of task-specific adaptation combined with synthetic data augmentation for narrative similarity modeling.

1 Introduction

Narrative texts constitute a fundamental form of human communication, encoding rich semantic information including event evolution, character development, and thematic progression over time. Understanding narrative similarity demands not only surface-level lexical matching but also deep reasoning over plot structure and event dynamics (Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016). In recent years, narrative modeling has attracted considerable attention in computational linguistics, spanning tasks such as event chain learning and story comprehension.

Conventional text similarity methods, such as embedding-based cosine similarity and sentence-level classification models (Reimers and Gurevych, 2019; Devlin et al., 2019), often fall short in capturing higher-level narrative structures. While large language models (LLMs) have exhibited strong reasoning capabilities through instruction tuning and chain-of-thought prompting (Wei et al., 2022; Peng et al., 2023; Kang et al., 2025; Na et al., 2025), zero-shot prompting remains limited in performance on fine-grained narrative comparison tasks (Bucher and Martini, 2024). Moreover, the scarcity of human-annotated triplet data poses a significant bottleneck for model fine-tuning, highlighting the need for efficient model adaptation strategies.

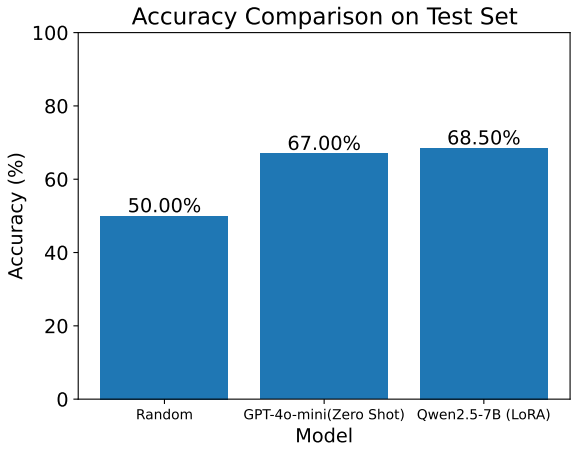

Figure 1: Accuracy comparison between Random, GPT-4o-mini and Qwen2.5-7B (LoRA) on the narrative similarity test set.

To this end, we propose an instruction-based fine-tuning framework for narrative similarity modeling to address the aforementioned challenges. We reformulate the narrative similarity task as an *instruction-following classification* problem, where a generative language model is trained to produce a discrete label ("Text A" or "Text B") as its output and adapt pretrained LLMs through parameter-efficient fine-tuning with LoRA (Hu et al., 2022), as depicted in Figure 2. To mitigate data scarcity, we further leverage synthetic triplet samples constructed by a frontier LLM, Claude-Sonnet-4, enabling the model to learn from a broader range of

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2793–2799

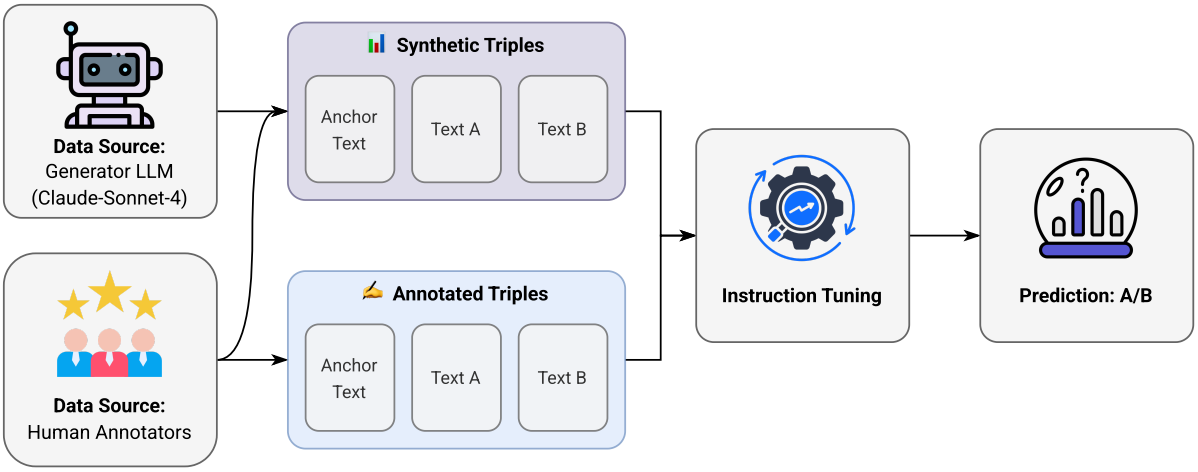

Figure 2: Synthetic data augmentation framework. Claude-Sonnet-4 is used to generate additional narrative triplets, which are combined with human-annotated data to enrich supervision for fine-tuning Qwen2.5-7B.

narrative structures and semantic relations during instruction fine-tuning.

We conduct comprehensive experiments on the SemEval-2026 Task 4 dataset (Hatzel et al., 2026). Results demonstrate that our approach effectively captures narrative-level similarity patterns. In particular, synthetic data augmentation plays a critical role in enhancing generalization, while LoRA-based adaptation enables efficient task-specific learning without sacrificing pretrained knowledge. As illustrated in Figure 1, our fine-tuned model with smaller parameters (Qwen2.5-7B, Yang et al., 2024) achieves slightly better accuracy than a substantially larger proprietary model (GPT-4o-mini) under zero-shot prompting on the narrative similarity task. This finding highlights that task-aligned adaptation can effectively compensate for differences in model scale, while offering notable advantages in computational efficiency, training cost, and deployment flexibility.

2 Methodology

2.1 Task Formulation

The task is formulated as a triplet comparison problem. Each instance consists of an anchor story s_a and two candidate stories s_1, s_2. The objective is to determine which candidate is narratively closer to the anchor, i.e.,

$$f(s_a, s_1, s_2) \in \{TextA, TextB\}. \quad (1)$$

The dataset provides human-annotated labels indicating the preferred candidate for each triplet. Importantly, the task requires only a relative similarity judgment between the two candidates with respect to the anchor, rather than estimating an absolute similarity score for each pair. This makes the task inherently comparative, as the model must jointly consider both candidates in the context of the anchor to make a decision.

2.2 Synthetic Data Augmentation

The manually annotated training set is limited in size, which poses a challenge for fine-tuning large language models. To address this, the shared task organizers have provided an additional set of synthetic triplet samples generated by Claude-Sonnet-4, following prior work demonstrating that model-generated data can effectively supplement human annotations in low-resource settings (Wang et al., 2023). As illustrated in Figure 2, the synthetic triplets are combined with the human-annotated data to form the final training set. All synthetic samples are converted into the same instruction-based format used during fine-tuning to ensure consistency across the two data sources.

2.3 Instruction Tuning

Each training instance is formatted as an instruction–input–output example following the Alpaca template (Taori et al., 2023). The instruction I_i describes the task, the input $X_i = (x_{a,i}, x_{1,i}, x_{2,i})$ contains the concatenated anchor and two candidate stories, and the output y_i is the target label ("Text A" or "Text B"). Figure 3 shows the prompt template used for formatting each instance. The model is trained autoregressively to maximize the conditional likelihood:

Narrative Similarity Prompt Template

```
{
  "instruction": "Analyze the
  following stories. Which one (Text
  A or Text B) is semantically closer
   and more similar in narrative to
  the Anchor Text? Respond with 'Text
   A' or 'Text B'.",

  "input": "Anchor Text:
  <anchor_text>

  Text A:
  <text_a>

  Text B:
  <text_b>",

  "output": "Text A or Text B"
}
```

Figure 3: JSON instruction prompt template used for narrative similarity fine-tuning in Track A.

$$\mathcal{L} = -\sum_{i=1}^{N} \log P_\theta(y_i \mid I_i, X_i), \tag{2}$$

where θ denotes the model parameters and N is the number of training instances.

We adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient fine-tuning. LoRA injects trainable low-rank matrices into selected linear projection layers while keeping the pretrained weights frozen. For a pretrained weight matrix $W \in \mathbb{R}^{d\times k}$, the update is parameterized as:

$$\Delta W = BA, \tag{3}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable with rank $r \ll \min(d, k)$. The adapted weight is computed as:

$$W' = W + \Delta W. \tag{4}$$

Only A and B are updated during training. We apply LoRA to Qwen2.5-7B (Yang et al., 2024) and train on the combined human-annotated and synthetic dataset. At inference time, the model generates a label token via greedy decoding.

3 Experiments

3.1 Dataset

This study uses the dataset provided in Track A of the SemEval-2026 Task 4: Narrative Story Similarity shared task (Hatzel et al., 2026). The dataset is organized as narrative triplets, where each sample contains one Anchor Text and two candidate stories. The dataset consists of four subsets: sample, development, synthetic, and test. The development set contains 200 human-annotated narrative similarity triplets and the sample set contains 39 triplets. The synthetic subset contains 1,900 organizer-provided triplets generated by LLMs. In our experiments, we use the organizer-provided synthetic split directly.

In our experiments, we use the sample set (39 triplets), the development set (200 triplets), and an additional synthetic dataset generated using Claude Sonnet 4 (200 triplets) as training data for model fine-tuning.

3.2 Baseline and Evaluation Metrics

The baselines for Track A (Narrative Story Similarity) include a random baseline and a prompting-based GPT-4o-mini baseline provided by the task organizers (Hatzel et al., 2026). The random baseline predicts the closer story by randomly selecting between Text A and Text B, serving as a lower-bound reference for task performance. The GPT-4o-mini baseline adopts a zero-shot prompting approach using GPT-4o mini (OpenAI, 2024). The input to the model consists of an Anchor Text and two candidate stories (Text A and Text B), which are concatenated into a single prompt. The model is instructed to determine which candidate story is narratively closer to the Anchor Text and produces a structured output including an explanation and a final decision. Performance is evaluated using classification accuracy.

3.3 Experimental Setup

In the training data preparation stage, we first converted the original training data into the Alpaca format required by the LLaMA-Factory fine-tuning framework (Zheng et al., 2024). During the conversion process, each original sample was mapped into an instruction–input–output triplet structure. The *instruction* describes the task objective, requiring the model to analyze the given stories and determine whether Text A or Text B is semantically and narratively closer to the Anchor Text. The *input* concatenates the three stories in a fixed structure, including the Anchor Text, Text A, and Text B. The *output* converts the original Boolean label into a textual answer: when `text_a_is_closer` is True, the output is "Text A", otherwise the output is "Text B". This format enables the construction of super-

vised training data and allows the model to learn narrative similarity through text generation.

In the model fine-tuning stage, we adopted the parameter-efficient fine-tuning method LoRA within the LLaMA-Factory framework to fine-tune pretrained language models (Zheng et al., 2024). The LoRA configuration is as follows: rank $r = 8$, scaling factor $\alpha = 16$ (effective scaling ratio $\alpha/r = 2.0$), and dropout rate of 0.1. Adapters are applied to all linear projection layers, including query, key, value, output, and feed-forward projections. We additionally employ the LoRA+ training strategy (Hayou et al., 2024), which sets an asymmetric learning rate ratio of $\lambda = 16.0$ between adapter matrices B and A, promoting more stable and efficient convergence. We apply this configuration to fine-tune Qwen2.5-7B (Yang et al., 2024) on the combined human-annotated and Claude-Sonnet-4 synthetic training data. During fine-tuning, the model was trained for 5 epochs with a batch size of 4, and an effective batch size of 16 was achieved through 4-step gradient accumulation. The optimizer learning rate was set to 3×10^{-5} and dynamically adjusted using a cosine learning rate decay schedule. A warmup strategy was applied at the beginning of training with a warmup ratio of 0.1. Gradient norm clipping was applied with a maximum norm of 1.0. All experiments were conducted on a single NVIDIA A100 Tensor Core GPU.

After fine-tuning, the model is evaluated on the official test set. During inference, inputs are constructed using the same instruction and prompt format adopted in the training stage to maintain consistency between fine-tuning and evaluation. We set `temperature = 0.1` and `top_p = 0.9` to approximate near-deterministic decoding while retaining a small degree of flexibility in token selection. The `max_new_tokens` is set to 1024 as an upper bound to avoid truncation.

3.4 Experimental results

Model	Accuracy (%)
Random	50.00
GPT-4o-mini	67.00
Qwen2.5-7B	**68.50**

Table 1: Experimental results of our proposed method against baselines on Track A.

Table 1 presents the performance of our proposed method compared with the baselines on Track A. The random baseline achieves 50.00% accuracy, which corresponds to chance-level performance for the binary decision task. The GPT-4o-mini zero-shot prompting baseline achieves 67.00% accuracy, demonstrating the strong reasoning capability of large language models even without task-specific fine-tuning. Our fine-tuned Qwen2.5-7B model achieves the best performance, reaching 68.50% accuracy on the test set. This result indicates that instruction-based fine-tuning combined with synthetic data augmentation effectively improves narrative similarity modeling beyond zero-shot prompting. Although the margin is modest (1.5 percentage points), this result suggests that task-specific fine-tuning can be competitive with zero-shot prompting from much larger proprietary models.

4 Related Work

4.1 Narrative Modeling and Story Understanding

Narrative modeling has long been studied in computational linguistics as a means of capturing event progression, character interactions, and plot-level coherence beyond surface lexical similarity. Pioneering work by Chambers and Jurafsky introduced narrative event chains to capture commonsense event sequences from large-scale corpora (Chambers and Jurafsky, 2008). Building on this foundation, subsequent research has transitioned toward more sophisticated story-understanding frameworks that integrate complex temporal and causal reasoning. narrative benchmarks further advanced this line of research. ROCStories and the Story Cloze Test, for instance, provide standard testbeds for evaluating commonsense reasoning over short narratives (Mostafazadeh et al., 2016).

Beyond sentence-level encoding, recent work has explored story-level embeddings and structured narrative representations that explicitly model global coherence and discourse dynamics (Hatzel and Biemann, 2024). Recent work proposes hybrid LLM-based architectures that combine classification-based and generative components to model longitudinal narrative progression in social media timelines (Qian et al., 2024; Wang et al., 2025). By integrating structured evidence extraction with generative summarization, such approaches capture evolving narrative states over extended contexts.

4.2 Text Similarity and Narrative Structure

Text similarity has traditionally been addressed through embedding-based approaches. Pretrained language models such as BERT (Devlin et al., 2019) and Sentence-BERT (Reimers and Gurevych, 2019) produce contextual representations that enable semantic similarity computation via cosine distance or classification heads. These methods have achieved strong performance in sentence-level and paragraph-level semantic matching tasks.

However, semantic similarity based solely on embedding representations or surface features primarily captures local semantic correspondence between textual segments, and often fails to model global narrative organization such as overarching plot arcs, pacing, and evolving character relationships. Recent benchmarks reveal systematic deficiencies in frontier models' ability to jointly represent narrative structure and orchestration across extended contexts, indicating that even advanced reasoning capabilities in large language models do not substantially improve the modeling of global narrative structure (Lu et al., 2026). Likewise, methods such as MLD-EA explicitly address gaps in narrative logic and emotional cohesion, demonstrating that complex story coherence often requires mechanisms beyond pairwise embedding similarity (Zhang and Long, 2025). Other frameworks focusing on long-term narrative consistency further highlight that capturing temporal progression and coherent story states remains a challenge for standard embedding models (Yi et al., 2025).

4.3 Parameter-Efficient Instruction Fine-Tuning

Recent advances in large language models have shown that many NLP tasks can be reformulated as instruction-following problems rather than task-specific classification pipelines. Instruction tuning enables pretrained generative models to learn from natural-language task descriptions and produce structured textual outputs. The Alpaca framework further demonstrates that instruction–input–output templates provide a simple and reproducible format for supervised instruction tuning, making it suitable for converting diverse NLP datasets into a unified generative learning format (Taori et al., 2023).

This line of work is particularly relevant to narrative comparison tasks, where the comparative structure between an anchor and two candidate stories maps naturally onto an instruction-following format. Prior work has also shown that fine-tuned smaller language models can outperform zero-shot generative models on text classification tasks (Bucher and Martini, 2024).

However, full fine-tuning of large language models is computationally expensive, especially in low-resource shared-task settings. Parameter-efficient fine-tuning methods address this issue by updating only a small number of additional parameters while keeping most pretrained weights frozen. LoRA introduces trainable low-rank matrices into selected projection layers, allowing task-specific adaptation with substantially fewer trainable parameters than full fine-tuning (Hu et al., 2022). Recent extensions such as LoRA+ further improve optimization by assigning different learning rates to the two low-rank adapter matrices, leading to more stable and efficient adaptation (Hayou et al., 2024).

5 Conclusions and Future Work

In this work, we addressed the Narrative Story Similarity task in SemEval-2026 Track A by reformulating it as an instruction-following generation problem. We adopted LoRA-based parameter-efficient fine-tuning to adapt pretrained large language models and incorporated synthetic triplet samples generated by Claude-Sonnet-4 to augment the limited human-annotated data. Our experimental results demonstrate the effectiveness of the proposed method, which slightly outperforms the zero-shot GPT-4o-mini baseline. These findings suggest that task-specific fine-tuning combined with synthetic data augmentation provides measurable improvements for narrative similarity modeling.

Limitations

Our study has several limitations. **First**, although synthetic data augmentation improves performance, the overall gain over strong zero-shot baselines remains moderate, indicating that narrative similarity remains a challenging high-level semantic reasoning task. **Second**, our approach primarily relies on generative modeling without explicitly incorporating structured narrative representations such as event graphs, discourse relations, or character dynamics. This may limit the model's ability to capture deeper narrative coherence. **Finally**, we only evaluated one pretrained backbone model (Qwen2.5-7B). Future work could explore larger-scale models or alternative architectures to further

investigate performance scalability.

References

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned "small" llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv preprint arXiv:2406.08660*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

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

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. LoRA: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Xiaoqiang Kang, Zimu Wang, Xiaobo Jin, Wei Wang, Kaizhu Huang, and Qiufeng Wang. 2025. Template-driven llm-paraphrased framework for tabular math word problem generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24303–24311.

Mingzhe Lu, Yiwen Wang, Yanbing Liu, Qi You, Chong Liu, Ruize Qin, Haoyu Dong, Wenyu Zhang, Jiarui Zhang, and Yunpeng Li. 2026. A benchmark for narrative orchestration in literary text. Preprint.

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.

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2025. Thinker-DDM: Modeling deliberation for machine translation with a drift-diffusion process. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 45–63, Sydney, Australia. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini model. `https://platform.openai.com/docs/models/gpt-4o-mini`.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Lu Qian, Yuqi Wang, Zimu Wang, Haiyang Zhang, Wei Wang, Ting Yu, and Anh Nguyen. 2024. Domain-specific guided summarization for mental health posts. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 150–159, Tokyo, Japan. Tokyo University of Foreign Studies.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zimu Wang, Hongbin Na, Rena Gao, Jiayuan Ma, Yining Hua, Ling Chen, and Wei Wang. 2025. From posts to timelines: Modeling mental health dynamics from social media timelines with hybrid LLMs. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 249–255, Albuquerque, New Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, and 1 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Shiyao Qian, Miao Zhang, Li Sun, and Tianyu Shi. 2025. Story coherence and retrieval enhancement for AI narratives. *arXiv preprint arXiv:2503.23512*.

Jinming Zhang and Yunfei Long. 2025. MLD-EA: Check and complete narrative coherence by introducing emotions and actions. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 1892–1907. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.