Overview of CCL25-Eval Task6: Chinese Essay Rhetoric Recognition Evaluation (CERRE)

Yujiang Lu[1], Nuowei Liu[1], Yupei Ren[1,2], Yicheng Zhu[3], Man Lan[1,2]†, Xiaopeng Bai[2,3], Mofan Xu[3], Qingyu Liao[4]

[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of AI for Education, East China Normal University
[3]Department of Chinese Language and Literature, East China Normal University
[4]CamScanner (BeeSchool), Shanghai Linguan Data Technology, China
{yujianglu, nwliu, ypren, yichengzhu}@stu.ecnu.edu.cn
mlan@cs.ecnu.edu.cn, {xpbai, mfxu}@zhwx.ecnu.edu.cn
qingyu_liao@intsig.net

Abstract

Literary grace in Chinese composition writing is a hallmark of linguistic sophistication, often realized through various rhetorical devices. The automatic identification and analysis of rhetorical devices in essays play a crucial role in educational NLP applications, particularly for assessing writing proficiency and facilitating pedagogical interventions. Although prior research has predominantly focused on coarse-grained recognition of limited rhetorical devices at sentence level, these approaches prove inadequate for handling complex rhetorical structures and emerging educational demands. In this paper, we present the CCL25-Eval Task6: Chinese Essay Rhetoric Recognition Evaluation (CERRE), a novel framework comprising three distinct evaluation tracks at the document level: (1) Fine-grained Form-level Categories Recognition, (2) Fine-grained Content-level Categories Recognition, and (3) Rhetorical Component Extraction. The evaluation has attracted 29 registered participating teams, with 8 teams submitting valid system outputs. In particular, two participating systems demonstrated superior performance by exceeding the baseline metrics in complete evaluation criteria.

Keywords: Rhetoric Recognition , Rhetorical Component Extraction , Essay Evaluation

1 Introduction

In Chinese composition writing, literary grace, as a formal characteristic of linguistic expression, is often manifested through the application of various rhetorical devices (Guo et al., 2018). Consequently, the identification and interpretation of rhetorical figures in essays not only reflect the level of literary sophistication and language proficiency, but also hold significant value in assisting educators in evaluating essay quality and guiding students to enhance their expressive capabilities. In recent years, research on rhetorical identification in compositions has predominantly employed feature matching and alignment methods, focusing on coarse-grained recognition of `parallelism` and `metaphor` from linguistic features such as sentence structure and semantic information (Niculae, 2013; Song et al., 2016). Other studies have designed model architectures to identify specific rhetorical techniques, such as `simile` (Liu et al., 2018; Zeng et al., 2020). To address multilayered definitions of rhetorical types, a limited number of studies have begun exploring fine-grained type recognition and component extraction for four rhetorical categories: `metaphor`, `personification`, `hyperbole`, and `parallelism` (Song et al., 2024; Wang et al., 2022; Nuowei et al., 2024; Liu et al., 2024b).

Based on the first CCL 2024 evaluation of Chinese essay rhetoric recognition and understanding (Nuowei et al., 2024), the data set for this evaluation similarly originates from authentic pedagogical contexts, comprising essays written by native Mandarin-speaking students, OCR processed by CamScanner

†Corresponding author

Proceedings of the 24th China National Conference on Computational Linguistics, pages 252-260, Jinan, China, August 11-14, 2025.

and BeeSchool. The corpus spans, but is not limited to, narrative and argumentative genres. Compared with the previous evaluation, this iteration introduces the following improvements. Firstly, besides the metaphor, personification, hyperbole, and parallelism, we expand four new devices in terms of rhetorical categories, i.e., repetition, hypophora, rhetorical questions, depiction, which are incorporated to encompass a broader spectrum of expressive forms. Secondly, unlike the prior evaluation on the sentence level identification, this evaluation operates at the document level in order to capture the complicate rhetorical structures across sentences.

As shown in Figure 1, CERRE [1] targets three distinct tracks for a given document:

- **Fine-grained Form-level Categories Recognition (Track1):** Recognition of eight coarse-grained rhetorical devices (metaphor, personification, hyperbole, parallelism, repetition, hypophora, rhetorical questions, depiction).
- **Fine-grained Content-level Categories Recognition (Track2):** Recognition restricted to the four high-frequency coarse-grained rhetorical devices (metaphor, personification, hyperbole, parallelism).
- **Rhetorical Component Extraction (Track3):** Structural decomposition of the four rhetorical devices mentioned above (metaphor, personification, hyperbole, parallelism).

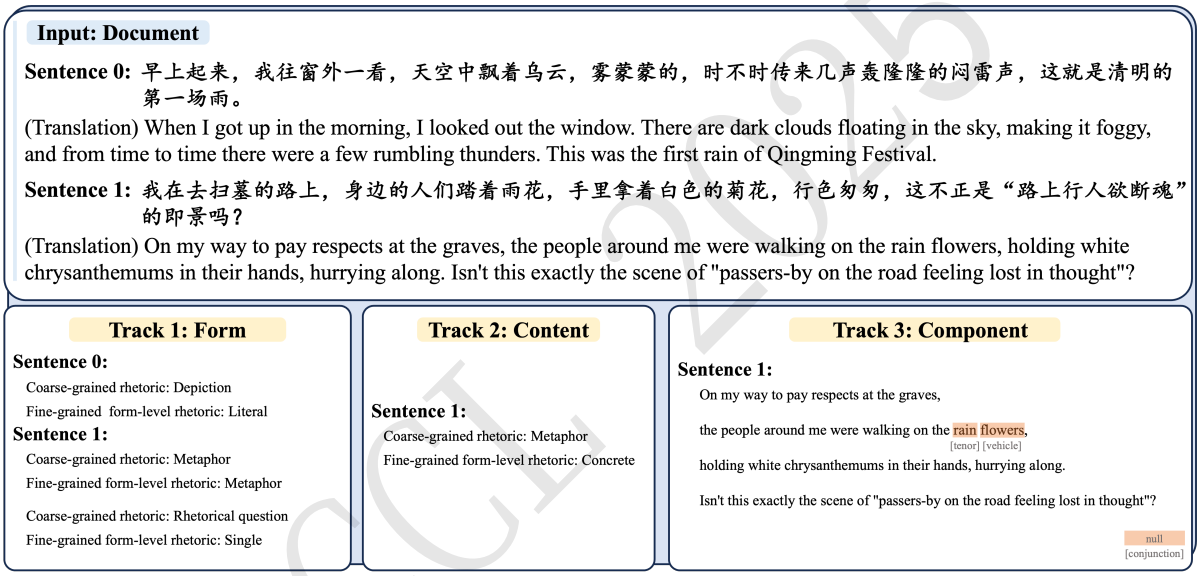

Figure 1: An example of CERRE.

2 Task Descriptions

2.1 Track1: Fine-grained Form-level Categories Recognition

Track1 uses documents as basic units and categorizes the rhetorical devices into eight coarse-grained categories: metaphor, personification, hyperbole, parallelism, repetition, hypophora, rhetorical questions, and depiction. As shown in Table 1, each category is further subdivided into fine-grained form-level categories.

- For metaphor, it is subdivided into simile, metaphor and metonymy.
- For personification, it is subdivided into noun, verb, and attributive.
- For hyperbole, it is subdivided into direct hyperbole and indirect hyperbole.

[1] https://github.com/cubenlp/CERRE-2025CCL/tree/main

Coarse-grained	Metaphor			Personification			Hyperbole		Parallelism	
Fine-grained	Simile	Metaphor	Metonymy	Noun	Verb	Attributive	Direct Hyperbole	Indirect Hyperbole	Constituent Parallelism	Sentence Parallelism
Coarse-grained	Repetition			Hypophora			Rhetorical Questions		Depiction	
Fine-grained	Immediate Repetition	Intermittent Repetition		Cohesive Hypophora	Disjointed Hypophora		Single Questions	Complex Questions	Synesthetic Depiction	Literal Depiction

Table 1: The relationship between coarse-grained categories and fine-grained form-level categories.

- For `parallelism`, it is subdivided into constituent `parallelism` and sentence `parallelism`.
- For `repetition`, it is subdivided into immediate `repetition` and intermittent `repetition`.
- For `hypophora`, it is subdivided into cohesive `hypophora` and disjointed `hypophora`.
- For `rhetorical questions`, it is subdivided into single `questions` and complex `questions`.
- For `depiction`, it is subdivided into synesthetic `depiction` and literal `depiction`.

Track1 is a multi-label classification problem, involving predicting the coarse-grained rhetorical category and fine-grained form-level category used in a given document.

2.2 Track2: Fine-grained Content-level Categories Recognition

Similar to track1, track2 uses documents as basic units and categorizes the rhetorical devices into four coarse-grained categories: `metaphor`, `personification`, `hyperbole`, and `parallelism`. As shown in Table 2, each category is further subdivided into fine-grained content-level categories.

Coarse-grained	Metaphor			Personification		Hyperbole			Parallelism		
Fine-grained	Concrete	Action	Abstract	Personification	Objectification	Amplification	Understatement	Prolepsis	Parallel	Sequential	Gradation

Table 2: The relationship between coarse-grained categories and fine-grained content-level categories.

- For `metaphor`, it is subdivided into `concrete`, `action`, and `abstract`.
- For `personification`, it is subdivided into `personification` and `objectification`.
- For `hyperbole`, it is subdivided into `amplification`, `understatement`, and `prolepsis`.
- For `parallelism`, it is subdivided into `parallel`, `sequential`, and `gradation`.

Track2 is a multi-label classification problem, involving predicting the coarse-grained rhetorical category and fine-grained content-level category used in a given document.

2.3 Track3: Rhetorical Component Extraction

Rhetorical components include the described object in the given document and the specific content of the description. Extracting these components helps understanding students' use of rhetoric, reflecting their language expression skills. As shown in Table 3, track3 uses documents as basic units and categorizes the rhetorical components in the documents into conjunction, tenor, and vehicle.

- For `metaphor-simile`, the rhetorical components include simile comparator, tenor, and vehicle. For metaphor-metaphor, the rhetorical components include metaphor comparator(may not exist), tenor, and vehicle. For metaphor-metonymy, the rhetorical components include vehicle only.

- For `personification`, regardless of form-level category, the rhetorical components include `personification` object and `personification` content.
- For `hyperbole`, regardless of form-level category, the rhetorical components include `hyperbole` object and `hyperbole` content.
- For `parallelism`, regardless of form-level category, the rhetorical components include `parallelism` marker and `parallelism` content.

Rhetorical Component	Metaphor			Personification	Hyperbole	Parallelism
	Simile	Metaphor	Metonymy			
Conjunction	Simile Comparator	Metaphor Comparator / -	-	-	-	Parallelism Marker
Tenor	Tenor	Tenor	-	Personification Object	Hyperbole Object	-
Vehicle	Vehicle	Vehicle	Vehicle	Personification Content	Hyperbole Content	Parallelism Content

Table 3: Rhetorical components of different fine-grained form-level categories. "-" indicates the absence of explicit rhetorical component for the given rhetorical device. Note: "Metaphor Comparator / -" in the table denotes that the conjunction of metaphor (fine-grained rhetorical device) may be either absent or present as metaphor comparator.

3 Dataset Statistics

3.1 Dataset Annotation

CERRE's dataset consists of authentic student-authored compositions systematically curated from standardized writing assessments administered across primary and secondary education levels. This collection captures developmental trajectories in writing acquisition through multi-genre textual artifacts spanning narrative discourse, argumentative essays, and expository writing. The dataset's ecological validity stems from its foundation in assessment-driven environments, providing stratified writing samples that reflect progressive competence levels from elementary syntactic mastery to advanced rhetorical organization.

The annotation process involved ten domain annotators specializing in computer science, education, and Chinese linguistics. We first established preliminary annotation guidelines, then conducted a pilot annotation phase where five annotators jointly labeled 150 essays (60 per annotator). Following this, we evaluated inter-annotator agreement (IAA) and refined the guidelines accordingly. In the final annotation stage, each of the five annotators independently labeled 200 essays. The complete corpus comprises 650 dual-annotated essays ($1,300$ total annotations), enabling systematic reliability verification through pairwise IAA analysis across all textual features. This dual-annotation design ensures measurement consistency while maintaining annotation throughput efficiency.

3.2 Dataset Statistics

Tracks 1, 2, and 3 utilize the same training, validation, and test sets, but each track features unique annotation schemes. Specifically, track1 targets fine-grained form-level categorization, whereas track2 focuses on content-level classification. In contrast, track3 is designed for rhetorical component analysis. As detailed in Table 4, the dataset statistics are provided, with the evaluation subset comprising approximately 5% of the total test set.

4 Evaluation Metircs

In CERRE, we adopt three key evaluation metrics. For all tracks, we employ the micro-F1 score (F_1) to assess classification performance. Additionally, the Intersection over Union (IoU) metric is utilized to

#Training set (documents)	#Test set (documents)
50	37459

Table 4: Statistics of dataset used in CERRE.

evaluate the accuracy of rhetorical sentence group localization. The final CERRE score is computed as the arithmetic mean of the results across track1, track2, and track3.

4.1 Track1: Fine-grained Form-level Categories Recognition

As shown in Equation 1, the overall F1 score of track1 consists of two components: the F1 scores for both coarse-grained categories and fine-grained form-level categories. Additionally, the final score S for track1 is computed by combining this overall F1 score with the IoU score.

$$\begin{aligned} F_1 &= 0.3 \times F_1^{\text{rhetorical}} + 0.7 \times F_1^{\text{form}} \\ S &= 0.3 \times IoU + 0.7 \times F_1 \end{aligned} \tag{1}$$

Where $F_1^{\text{rhetorical}}$ denotes the F1 score of coarse-grained categories, F_1^{form} denotes the F1 score of fine-grained form-level categories, and IoU denotes the IoU score of rhetorical sentence group localization.

4.2 Track2: Fine-grained Content-level Categories Recognition

As shown in Equation 2, the overall F1 score of track2 consists of two components: the F1 scores for both coarse-grained categories and fine-grained form-level categories. Additionally, the final score S for track2 is computed by combining this overall F1 score with the IoU score.

$$\begin{aligned} F_1 &= 0.3 \times F_1^{\text{rhetorical}} + 0.7 \times F_1^{\text{content}} \\ S &= 0.3 \times IoU + 0.7 \times F_1 \end{aligned} \tag{2}$$

Where $F_1^{\text{rhetorical}}$ denotes the F1 score of coarse-grained categories, F_1^{content} denotes the F1 score of fine-grained content-level categories, and IoU denotes the IoU score of rhetorical sentence group localization.

4.3 Track3: Rhetorical Component Extraction

As shown in Equation 3, the overall F1 score of track3 consists of two components: the F1 scores for both coarse-grained categories and fine-grained form-level categories. Additionally, the final score S for track3 is computed by combining this overall F1 score with the IoU score.

$$\begin{aligned} F_1 &= \frac{1}{3} \times F_1^{\text{conjunction}} + \frac{1}{3} \times F_1^{\text{tenor}} + \frac{1}{3} \times F_1^{\text{vehicle}} \\ S &= 0.3 \times IoU + 0.7 \times F_1 \end{aligned} \tag{3}$$

Where $F_1^{\text{conjunction}}$, F_1^{tenor} and F_1^{vehicle} denotes the F1 score of conjunctions, tenors and vehicles respectively, and IoU denotes the IoU score of rhetorical sentence group localization.

5 Baselines

In this section, we introduce the baseline approaches used in CERRE and the scores on track1, track2 and track3.

For Track1, track2, and track3, we formulate all tasks as sequence-to-sequence problems and conduct inference on the test set using Qwen3-14B [2] (Yang et al., 2025) in a few-shot setting. The few-shot demonstrations are exclusively sampled from the training set. Specifically, we construct an initial

[2] https://huggingface.co/Qwen/Qwen3-14B

prompt using 30 in-context examples to guide the model across all rhetorical tasks, followed by task-specific inference with 15 examples per task. To facilitate automated evaluation, we enforce structured JSON-formatted outputs by explicitly specifying this requirement in the task prompts. Our baseline fully follows the official recommendations of Qwen3-14B configuration.

As shown in Table 5, we report the baseline scores on the test set for reference.

Track	S (%)
Track1	43.68
Track2	51.40
Track3	37.48

Table 5: Baseline results on the test set.

6 Results

In this section, we first discuss the overall results, including the statistics of the participating teams and their scores on each track (See Section 6.1). Considering the correlation between different tracks, most of the teams choose to combine the dataset from different tracks for joint training. Therefore, we then discuss the approaches they use respectively (See Section 6.2 - Section 6.4). Finally, an overall analysis will be discussed in Section 6.5.

6.1 Overall Results

For CCL25-Eval Task6, a total of 29 teams registered to participate in CERRE. Utimately, 8 teams submitted evaluation results and obtained valid scores, with 2 of these teams achieving an overall score that surpassed the baseline. Details are listed in Table 6. It is worth noting that the combined score based on each track is obtained by weighted average of the baseline (the default score of baseline is 60%).

Furthermore, the statistics on the usage of LLMs, external data, data augmentation and in-context learning methods by the top 3 teams based on their combined scores are listed in Table 7.

Team Name	Track1 (%)	Track2 (%)	Track3 (%)	Combined Score (%)
The Open University of China (OUC)	**47.18**	**54.03**	**39.94**	**63.94**
Beijing Language and Culture University (BLCU)	43.47	51.71	38.27	60.45
Baseline	43.68	51.40	37.48	60.00
Yunnan University (YNU)	40.30	52.23	26.25	52.78
Beijing Institute of Technology (BIT)	37.78	44.42	18.42	44.41
Shenyang Aerospace University (SAU)	27.24	35.18	33.04	43.79
Zhengzhou University (ZZU)	39.58	41.98	-	34.46
Individual Team	40.00	52.66	-	32.94
China Fire and Rescue Institute (CFRI)	26.05	35.51	-	25.74

Table 6: Scores of the participating teams. "-" indicates that the team did not submit evaluation results on the track, and the combined score is calculated based on the baseline.

6.2 Team OUC

OUC presents a framework for Chinese essay rhetoric recognition leveraging LLMs through three key methods: (1) Structured knowledge integration via JSON-formatted outputs (Shorten et al., 2024), where rhetorical elements (e.g., components, devices) are formalized into a Chinese-translated JSON schema to align with LLMs' generation capabilities while preserving structural coherence. This approach bridges the gap between free-form generation and task-specific structured outputs by encoding domain knowledge into the system prompt and utilizing natural language-aligned keys. (2) Hybrid training paradigms combining LoRA and in-context learning (Brown et al., 2020), where low-rank adaptation (LoRA) fine-tuning (Hu et al., 2022) is applied to Qwen2.5-72B (Yang et al., 2024) for parameter-efficient knowl-

Team Name	LLMs	External Data	Data Augmentation	In-Context
OUC	✓	✗	✗	✓
BLCU	✓	✓	✓	✓
YNU	✓	✗	✓	✓

Table 7: Statistics on the usage of LLMs, external data and data augmentation methods. "LLMs" indicates whether to use Large Language Models (LLMs). "External Data" indicates whether data outside the provided dataset for CERRE is used. "Data Augmentation" indicates whether any augmentation is performed on the provided dataset for CERRE. "In-Context" indicates whether to use in-context learning.

edge injection, while in-context learning dynamically incorporates training examples for closed-source models. A novel "SEPA" strategy further decouples shared coarse-grained labels from track-specific fine-grained classifications to optimize context utilization. (3) Ensemble strategies with task-specific mechanisms, including linear-weighted ensembles for rhetoric type classification and fallback protocols for component extraction to handle JSON parsing failures or content safety constraints. These methods synergistically enhance robustness, particularly for complex tasks requiring precise localization of rhetorical components. OUC's post hoc analysis reveals that even with zero training loss, augmenting LoRA with in-context examples improves extraction accuracy significantly.

6.3 Team BLCU

BLCU proposes a multi-strategy fusion framework centered on data augmentation and model collaboration for document-level rhetoric recognition and component extraction in Chinese essays. For rhetorical categorization (Tracks 1-2), they introduce a recursive document generation method to address data scarcity: leveraging sentence-level annotations from the CERD dataset (Liu et al., 2024b), they iteratively generate context-aware sentences with controlled rhetorical patterns using a teacher model (DeepSeek-v3 (Liu et al., 2024a)), forming coherent, diverse document-level training data. This is combined with efficient supervised fine-tuning via LoRA adaptation of Qwen2.5-32B-instruct (Yang et al., 2024), decomposing documents into sentence-level units for intra-sentence pattern learning. To capture inter-sentence rhetorical relationships, a few-shot relation modeling module is integrated, enhancing comprehension through syntactic-semantic interaction. For component extraction (Track3), a two-stage pipeline is designed: a lightweight model first identifies rhetorical categories, followed by DeepSeek-v3-driven entity recognition conditioned on prior classification, achieving hierarchical precision improvement. BLCU's ablation studies reveal that sentence-level decomposition improves performance on track1 by $+1.67$ points, while few-shot relation modeling contributes $+0.51$ gains on track2.

6.4 Team YNU

YNU presents a multi-stage framework for enhancing Chinese essay rhetoric recognition in low-resource settings (Lee et al., 2024). The methodology integrates three core components: (1) Targeted data augmentation with semi-automatic annotation. A prompt-based pipeline leverages DeepSeek-V3 to generate synthetic rhetorical annotations for unlabeled data, followed by manual refinement to ensure label accuracy and diversity. This addresses data scarcity and class imbalance (e.g., underrepresented categories like `metaphor` and `parallelism`). (2) Instruction-tuned LLMs adaptation (Wang et al., 2023). The Qwen2.5 model is fine-tuned using one-shot learning and in-context examples, with parameter-efficient tuning via LoRA and 4-bit quantization. Training employs a cosine-annealed AdamW optimizer with gradient clipping to stabilize convergence. (3) Ensemble voting for robust prediction. Multiple models fine-tuned with diverse hyperparameters (learning rates, batch sizes, prompt designs) are combined through majority voting, reducing individual model biases and improving generalization. For component extraction, the ensemble aggregates predictions across models to enhance structural sensitivity to syntactic and stylistic variations. YNU's approach highlights synthetic data generation, parameter-efficient

adaptation, and ensemble reasoning as effective strategies for multi-label rhetorical analysis in resource-constrained NLP tasks.

6.5 Overall Analysis

Overall, the teams leveraging LLMs consistently outperformed those employing alternative approaches across all tracks. While the use of external data and various data augmentation strategies generally enhanced performance, it is noteworthy that the top-performing team (OUC) achieved superior results without employing either technique. The majority of participants adopted parameter-efficient fine-tuning through LoRA and implemented instruction fine-tuning during training, with inference predominantly conducted via in-context learning paradigms. Data augmentation methodologies varied considerably across teams, though their effectiveness remained contingent on task-specific requirements. Due to space constraints and their comparative performance outcomes, detailed analyses of non-LLM-based approaches have been omitted from this section.

7 Conclusion

In this paper, we propose the CCL25-Eval Task6: **C**hinese **E**ssay **R**hetoric **R**ecognition **E**valuation (CERRE), comprising three core subtasks: fine-grained form-level category recognition, content-level category recognition, and rhetorical component extraction. The evaluation attracted 29 registered teams from the NLP community, with 8 submitting valid system outputs for comprehensive assessment. Furthermore, our discussion of the approaches in CERRE reveals that top-performing systems (ranking top-3 in overall scores) predominantly leveraged LLMs enhanced through in-context learning paradigms, achieving substantial performance gains across all subtasks.

Acknowledgements

This work was supported by National Natural Science Foundation of China (72192820, 72192824), Fundamental Research Funds for the Central Universities (2024QKT004), Artificial Intelligence-Powered Research Paradigm Reform and Discipline Leap Plan from Shanghai Municipal Education Commission (2024AI02004), Science and Technology Commission of Shanghai Municipality (22DZ2229004), and Shanghai Trusted Industry Internet Software Collaborative Innovation Center.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jingjin Guo, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. 2018. Attention-based bilstm network for chinese simile recognition. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 144–147. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. 2024. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Nuowei Liu, Xinhao Chen, Hongyi Wu, Changzhi Sun, Man Lan, Yuanbin Wu, Xiaopeng Bai, Shaoguang Mao, and Yan Xia. 2024b. Cerd: A comprehensive chinese rhetoric dataset for rhetorical understanding and generation in essays. *arXiv preprint arXiv:2409.19691*.

Vlad Niculae. 2013. Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 110–114.

Liu Nuowei, Chen Xinhao, Ren Yupei, Man Lan, Bai Xiaopeng, Wu Yuanbin, Mao Shaoguang, and Xia Yan. 2024. Chinese essay rhetoric recognition and understanding (cerru). In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 253–261.

Connor Shorten, Charles Pierse, Thomas Benjamin Smith, Erika Cardenas, Akanksha Sharma, John Trengrove, and Bob van Luijt. 2024. Structuredrag: Json response formatting with large language models. *arXiv preprint arXiv:2408.11061*.

Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Learning to identify sentence parallelism in student essays. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 794–803.

Jinwang Song, Hongying Zan, and Kunli Zhang. 2024. System report for ccl24-eval task 6: Essay rhetoric recognition and understanding using synthetic data and model ensemble enhanced large language models. In *The 23rd China National Conference on Computational Linguistics (Evaluation Workshop)*.

Xiaoyue Wang, Linfeng Song, Xin Liu, Chulun Zhou, and Jinsong Su. 2022. Getting the most out of simile recognition. *arXiv preprint arXiv:2211.05984*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9515–9522.