System Report for CCL25-Eval Task 6: Enhancing Chinese Essay Rhetoric Recognition through Targeted Data Augmentation and Model Ensemble Voting

Jingjun Tang, Zhiwen Tang[†]
School of Information Science & Engineering, Yunnan University, Kunming 650091
12024215193@stu.ynu.edu.cn, zhiwen.tang@ynu.edu.cn

Abstract

This paper presents our approach to the Second Chinese Essay Rhetoric Identification and Understanding Competition, which focuses on analyzing rhetorical features in essays written by primary and secondary school students. The competition includes three tasks: multi-label classification of rhetorical forms, divided into 9 coarse-grained and 19 fine-grained categories; multi-label classification of rhetorical content, comprising 5 coarse-grained and 11 fine-grained categories specific to certain rhetorical types; and extraction of rhetorical components, including connectives, descriptive objects, and specific rhetorical content. To address the challenge of limited training data, we applied targeted data augmentation and manual corrections to build a high-quality dataset. We then fine-tuned large language models using one-shot and in-context learning. Finally, we employed an ensemble strategy that integrates model predictions through a voting mechanism. Our system achieved a score of 52.78 and ranked third in the competition.

Keywords: Rhetorical Analysis, One-Shot Learning, Data Augmentation, Ensemble Decision, Voting Mechanism

1 Introduction

Recent progress in Natural Language Processing (NLP) has significantly advanced the automation of educational assessment tasks, particularly in evaluating student writing (Liu et al., 2024). Among various indicators of writing proficiency, the use of rhetorical devices plays a crucial role in reflecting students' expressive and argumentative capabilities (Burstein, 2003; Lai et al., 2023; Liu et al., 2018; Xiaoxi et al., 2018). However, large-scale manual annotation of rhetorical phenomena remains infeasible, motivating the design of the CCL 2025 Shared Task on Rhetoric Recognition and Understanding. This shared task presents a suite of three technically challenging multi-label subtasks: identifying rhetorical types and their formal linguistic features, understanding the semantic content conveyed by rhetorical expressions, and extracting fine-grained rhetorical components from texts.

These subtasks involve a high degree of structural complexity and linguistic ambiguity. The identification of rhetorical types and formal features requires models to generalize across diverse textual styles, rhetorical strategies, and discourse structures. Semantic understanding further complicates the problem, as rhetorical language often conveys implicit meanings and pragmatic intentions that are highly context-dependent. Additionally, extracting rhetorical components demands sensitivity to subtle syntactic and stylistic variations, which can differ significantly across writing samples, genres, and proficiency levels. These intertwined challenges make the task particularly demanding, especially in low-resource educational settings where annotated data is scarce.

To address these challenges, we design a comprehensive system leveraging the capabilities of large language models (LLMs). For rhetorical type classification, we construct a prompt-based data augmentation pipeline followed by expert-guided filtering to address data sparsity. For rhetorical semantic understanding, we employ instruction-tuned LLMs with one-shot and in-context learning paradigms (e.g.,

[†] Corresponding author

Proceedings of the 24th China National Conference on Computational Linguistics, pages 244-251, Jinan, China, August 11-14, 2025.

Tongyi Qwen 2.5), enabling effective generalization under limited supervision. For rhetorical component extraction, we introduce an ensemble framework that combines multiple model predictions via majority voting, enhancing robustness and prediction accuracy.

Our system achieved a final score of 52.78, placing third in the official competition ranking. These results underscore the potential of synthetic data generation, parameter-efficient tuning, and ensemble-based reasoning in addressing complex, low-resource NLP tasks in educational contexts (Lee et al., 2024).

2 Related Work

Early approaches to rhetorical device recognition primarily relied on rule-based systems and traditional machine learning models, such as Support Vector Machines and Conditional Random Fields, particularly for tasks like metaphor detection (Sun et al., 2019; Lai et al., 2023). The introduction of pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), brought significant improvements in rhetorical form classification tasks (Liu et al., 2018). More recently, large language models (LLMs), including GPT-3 (Brown et al., 2020), have enabled zero-shot and few-shot learning through instruction tuning and in-context learning paradigms (Wang et al., 2023; Hu et al., 2022).

Despite these advancements, the application of LLMs to complex multi-label rhetorical analysis—especially in the context of Chinese student essays—remains largely underexplored. Our work builds upon recent progress in low-resource adaptation (Lee et al., 2024) by integrating synthetic data generation, parameter-efficient fine-tuning via LoRA (Hu et al., 2022), and ensemble-based inference strategies. This combination aims to enhance model performance and generalizability in educational NLP tasks involving nuanced rhetorical understanding.

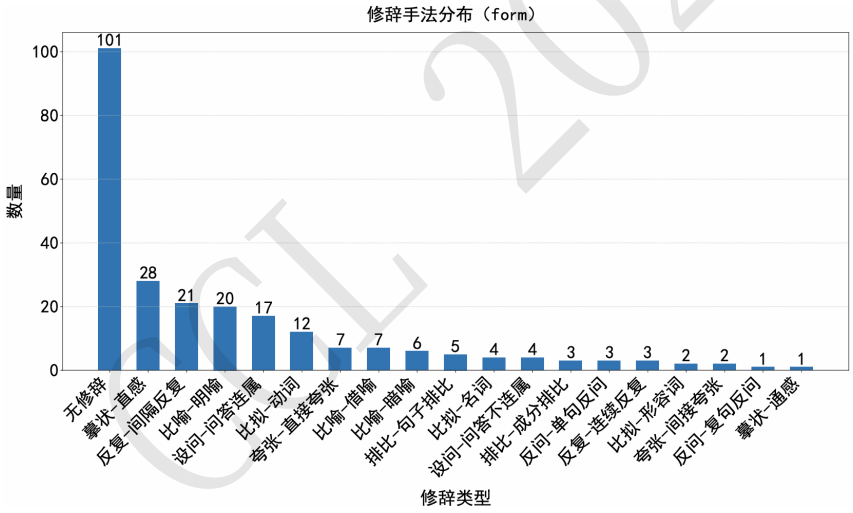

Figure 1: Distribution of Rhetorical Devices in Original Training Dataset (Track 1)

3 Methodology

3.1 Framework Overview

Our proposed framework consists of a comprehensive pipeline designed to enhance model performance in rhetorical analysis tasks. As illustrated in Figure 2, the approach integrates data augmentation, manual refinement, instruction-based fine-tuning, and ensemble inference to address the challenges of limited training data and complex multi-label classification.

First, we construct a high-quality training dataset through a semi-automatic annotation process. Specifically, we design tailored prompts for DeepSeek-V3 to automatically generate rhetorical annotations for unlabeled data. These initial annotations are then manually reviewed and corrected to eliminate errors, resulting in a reliable and diverse dataset.

Next, we perform supervised fine-tuning (SFT) on multiple configurations of the Qwen 2.5 model using the refined dataset. This stage adopts an instruction-based learning strategy using one-shot to improve model generalization across subtasks.

Finally, we integrate the fine-tuned models into an ensemble system. During inference, each model independently processes the same input, and their outputs are aggregated using a majority voting mechanism. This ensemble strategy not only consolidates model predictions but also significantly enhances the overall prediction accuracy.

In summary, our methodology effectively combines automatic and manual data curation with tailored fine-tuning and ensemble decision-making, resulting in robust performance improvements in low-resource rhetorical analysis tasks.

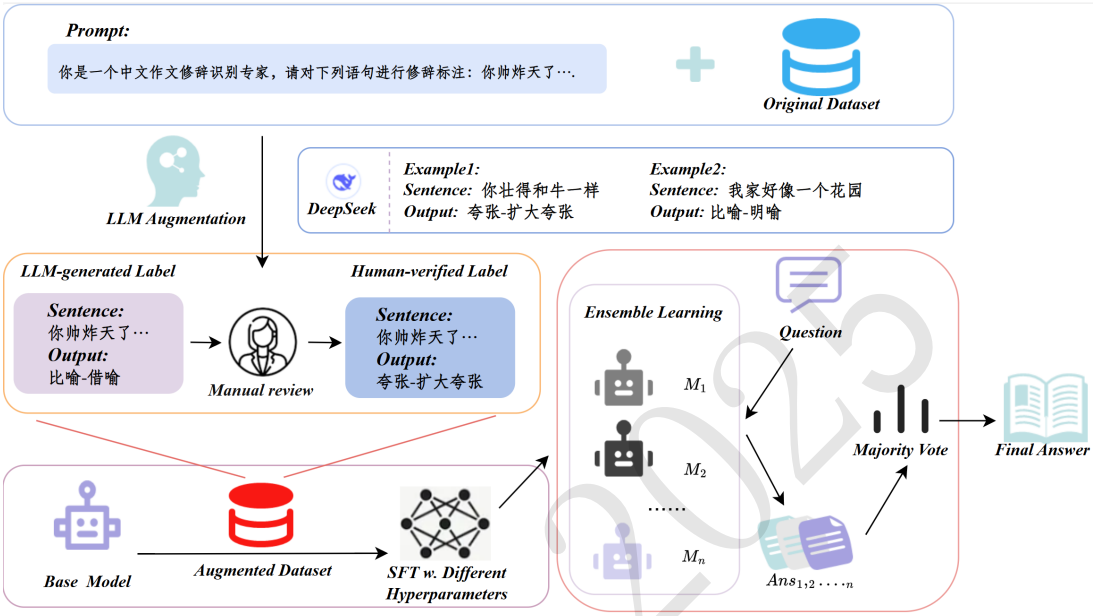

Figure 2: Overall Framework Architecture

3.2 Data Augmentation

An analysis of the original training datasets—each containing only 50 samples per track—revealed a pronounced imbalance in both the quantity and distribution of sentences containing rhetorical devices versus those without. For instance, the distribution in Track 1, as shown in Figure 1, clearly illustrates this disparity. Similar analyses for Track 2 and Track 3 are provided in the Appendix.

Moreover, the public datasets exhibited similar issues: not only was the overall sample size extremely limited, but the rhetorical categories were also unevenly distributed and in some cases not fully covered. To address these challenges, we employed DeepSeek-V3 for data augmentation, applying the same strategy to both our own training data and the publicly available data.

Specifically, we identified rhetorical categories with the lowest representation in the training set—such as metaphor, irony, and parallelism—and generated additional samples for these categories using instruction-tuned prompts (see Appendix for examples). For each under-represented category, we crafted prompts to guide the model in producing stylistically and semantically coherent examples that exhibited the target rhetorical technique. All generated samples were manually reviewed to ensure label correctness and rhetorical consistency.

As a result of this augmentation and filtering process—integrated into our data preparation pipeline (see Figure 2)—each track's dataset was expanded to approximately 400 instances, with roughly 20 examples per rhetorical category. This significantly reduced the label imbalance and improved the model's generalization performance across diverse rhetorical structures.

3.3 Model Ensemble Voting

To leverage complementary strengths and increase diversity, our ensemble consists of multiple models fine-tuned with different hyperparameter configurations, such as varying learning rates, batch sizes, and prompt designs. This diversity helps to reduce individual model biases and improve overall system robustness.

Building on this, we adopted an ensemble learning strategy that aggregates predictions from these fine-tuned models, effectively mitigating errors from any single model and enhancing generalization. The workflow is illustrated in Figure 2.

Let $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$ denote a set of N distinct models obtained from the supervised fine-tuning phase, where each M_i represents a fine-tuned variant of the base large language model. For a given input question Q, each model M_i independently performs inference and produces its prediction $A_i = M_i(Q)$. These individual predictions $\{A_1, A_2, \ldots, A_N\}$ serve as the basis for aggregation.

At the core of our ensemble approach lies a majority voting mechanism. For each unique candidate answer c among the predictions, we count the number of votes it receives across all models:

$$\text{votes}(c) = \sum_{i=1}^{N} \mathbb{I}(A_i = c) \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if $A_i = c$, and 0 otherwise.

The final ensemble prediction, A_{final}, is determined by selecting the candidate answer with the highest vote count:

$$A_{\text{final}} = \arg\max_{c} \big(\text{votes}(c)\big) \tag{2}$$

This ensemble technique leverages the diversity among models, resulting in more stable and reliable predictions compared to any single model alone. The complementary strengths captured through different training configurations significantly enhance the system's overall performance.

4 Experiments

4.1 Experimental Setup

We adopted the Qwen2.5-14B-Instruction model as our base, utilizing bitsandbytes for 4-bit quantization combined with LoRA for efficient adaptive fine-tuning. The AdamW optimizer was employed with a weight decay of 1×10^{-2}. The initial learning rate was set to 5×10^{-5} and followed a cosine annealing schedule throughout the training epochs, without any warm-up steps. Training was performed with a batch size of 2 and 8 gradient accumulation steps, while the maximum gradient norm was clipped at 1.0. Additionally, the maximum sequence length was restricted to 2048 tokens, and all computations were carried out using bf16 precision.

To improve robustness and explore the effect of different hyperparameters, we integrated multiple configurations of the same base model in our experiments. Specifically, we trained models with learning rates of 5×10^{-5}, 4×10^{-4}, and 3×10^{-3}, corresponding to 15, 10, and 5 epochs, respectively. The batch sizes used were 2, 4, and 6, while all other settings remained unchanged. Each configuration was trained using both the original and the augmented versions of the training dataset.

4.2 Results and Analysis

This section presents the experimental results of our proposed framework on the competition test sets. We compare our system's performance with other participating teams across the three subtasks: Rhetoric Form Recognition, Rhetoric Content Recognition, and Rhetoric Component Extraction. According to the competition rules, the overall score was not computed directly by us, but rather transformed by the organizers based on the performance of the official baseline system.

As shown in Table 1, our team achieved a combined score of 52.78, securing third place in the competition. We performed particularly well in Content Recognition (60.97) and Form Recognition

Team	Form Conversion	Content Conversion	Component Conversion	Combined Score
Team1	64.81	63.07	63.94	63.94
Team2	59.71	60.36	61.26	60.45
Ours	55.36	60.97	42.02	52.78
Team3	51.90	51.85	29.49	44.41
Team4	37.42	41.07	52.89	43.79
Team5	54.37	49.00	0.00	34.46
Team6	43.43	55.40	0.00	32.94
Team7	35.78	41.45	0.00	25.74

Table 1: Competition Results of Participating Teams

(55.36). Although there is room for improvement in Component Extraction (42.02), our overall strategy—combining data augmentation, fine-tuning, and ensemble learning—proved effective. This approach enabled us to achieve competitive results despite the limited training data, demonstrating its potential for tackling complex educational NLP tasks.

4.3 Ablation Study

To verify the effectiveness of our method, we conducted an ablation study. However, due to the large number of blind test sets and equipment limitations, it was not possible to perform the ablation through the official evaluation platform within the specified time. Therefore, the results presented here are for reference only. We randomly selected 10 samples from the official training data as the validation set, using the remaining 50 official training samples for training. The experimental results are shown in Table 2.

Module	Form Conversion	Content Conversion	Component Conversion	Combined Score
Full Setting	55.36	60.97	42.02	52.78
- w/o Ensemble Learning	50.46	54.89	37.67	47.67
- w/o Data Augmentation	44.78	47.56	33.45	41.93
- w/o Data Augmentation & Ensemble Learning	39.96	41.23	20.31	33.83

Table 2: Ablation Study Results for Qwen2.5-14B-Instruction

5 Conclusion and Future Work

Our multi-stage pipeline utilized large language models (LLMs) with prompt-based data augmentation, manual verification, and supervised fine-tuning (SFT) for multi-label classification and component extraction. An ensemble majority voting strategy was employed to ensure robust and consistent predictions. Our system achieved a commendable combined score of 52.78, securing third place overall and demonstrating the effectiveness of our approach. However, the performance on Track 3 was relatively limited, which may be attributed to the increased complexity of the task, the model's weaker ability in component extraction for this specific track, or suboptimal parameter configurations. Future work will focus on enhancing component extraction through specialized model architectures, exploring larger or domain-specific LLMs, and investigating more advanced ensemble strategies.

Acknowledgements

This work is supported by the Open Research Project of the Yunnan University Resilience and Excellence Children's Character Development Platform (K207003250007), and Yunnan Fundamental Research Project (202501AT070231).

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.

Jill Burstein. 2003. The e-rater scoring engine: Automated essay scoring with natural language processing. *Automated essay scoring: A cross-disciplinary perspective*, 113121.

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.

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.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. *arXiv preprint arXiv:2306.00121*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nuowei Liu, Xinhao Chen, Yupei Ren, Man Lan, Xiaopeng Bai, Yuanbin Wu, Shaoguang Mao, and Yan Xia. 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, Taiyuan, China. Chinese Information Processing Society of China.

Weiwei Sun, Yufei Chen, Xiaojun Wan, and Meichun Liu. 2019. Parsing chinese sentences with grammatical relations. *Computational Linguistics*, 45(1):95–136.

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*.

Huang Xiaoxi, Li Hanyu, Wang Rongbo, Wang Xiaohua, and Chen Zhiqun. 2018. Recognizing metaphor with convolution neural network and svm. *Data Analysis and Knowledge Discovery*.

Appendix

A Examples of the fine-tuning data

A.1 Track1:

Prompt:
你是一个中文修辞手法识别专家，请你识别出以下句子中的修辞类别，类别共有8大类和18小类。若你认为有多个结果，请使用换行符隔开。若你认为无修辞则输出无修辞即可
示例：这就是我家的小小“动物园”好玩又温馨。
修辞类别选项如下:比喻:明喻, 暗喻, 借喻。比拟:名词, 动词, 形容词。夸张:直接夸张, 间接夸张。排比:成分排比, 句子排比。反复:间隔反复, 连续反复。设问:问答连属, 问答不连属。反问:单句反问, 复句反问。摹状:通感,直感。
答：比喻-借喻
请识别下列句子:怎么样，厉害吧?
Output:
反问-单句反问

A.2 Track2:

Prompt:
你是一个中文修辞手法识别专家，请你识别出以下句子中的修辞类别，共有4大类和11小类。若你认为有多个结果，请使用换行符隔开。若你认为无修辞则输出无修辞即可
示例：这就是我家的小小“动物园”好玩又温馨。
修辞类别选项如下:比喻：实在物、动作、抽象概念。比拟：拟人、拟物。夸张：扩大夸张、缩小夸张、超前夸张。排比：并列、承接、递进。
答：比喻-抽象概念
请识别下列句子:一刹那，气垫船猛地飞速下滑，让人有强烈的下坠感，我的心砰砰地跳着，好像要随着船体飞出去。
Output:
夸张-扩大夸张

A.3 Track3:

Prompt:
你是一个中文修辞手法识别专家，请你判断出以下句子中的修辞类别并抽取修辞成分，共有4大类和10小类。若你认为有多个结果，请使用换行符隔开。若你认为无修辞则输出无修辞即可。
注意：排比只需要连接词；比喻需要对像、连接词（借喻不用）、内容；比拟需要对象、内容；夸张需要对象、内容。
示例：这就是我家的小小“动物园”好玩又温馨。
修辞类别选项如下:比喻：明喻、暗喻、借喻。比拟：名词、动词、形容词。夸张：直接夸张、间接夸张。排比：成分排比、句子排比。
答：比喻-借喻。对象：我家 - 连接词：无 - 内容：小小“动物园”
请识别下列句子:一刹那，气垫船猛地飞速下滑，让人有强烈的下坠感，我的心砰砰地跳着，好像要随着船体飞出去。
Output:
夸张-直接夸张。对象：心砰砰地跳 - 连接词：无 - 内容：好像要随着船体飞出去

B Prompt Example for DeepSeek-V3

B.1 Track1:

Prompt:
你是一个擅长于中文作文修辞句写作和中文修辞手法识别专家。
第一：请你生成一个带有某种修辞（比喻:明喻, 暗喻, 借喻。比拟:名词, 动词, 形容词。夸张:直接夸张, 间接夸张。排比:成分排比, 句子排比。反复:间隔反复, 连续反复。设问:问答连属, 问答不连属。反问:单句反问, 复句反问。摹状:通感, 直感。）的句子。
第二：请你给出你生成句子的修辞类别。
Output:
句子：你帅炸天了
修辞类别：夸张：直接夸张

B.2 Track2:

Prompt:
你是一个擅长于中文作文修辞句写作和中文修辞手法识别专家。
第一：请你生成一个带有某种修辞（比喻：实在物、动作、抽象概念。比拟：拟人、拟物。夸张：扩大夸张、缩小夸张、超前夸张。排比：并列、承接、递进。）的句子。
第二：请你给出你生成句子的修辞类别。
Output:
句子：你帅炸天了
修辞类别：夸张：扩大夸张

B.3 Track3:

Prompt:
你是一个擅长于中文作文修辞句写作和中文修辞手法识别专家。
第一：请你生成一个带有某种修辞（比喻：明喻、暗喻、借喻。比拟：名词、动词、形容词。夸张：直接夸张、间接夸张。排比：成分排比、句子排比。）的句子。（注意：排比只需要连接词；比喻需要对像、连接词（借喻不用）、内容；比拟需要对象、内容；夸张需要对象、内容。）
第二：请你给出你生成句子的修辞类别以及提取相关的对象、连接词、内容。
Output:
句子：我和妈妈回到房间我把卷子拿了出来，当我妈看到成绩的那一刻，时间仿佛都静止了。
修辞类别：比喻-明喻。对象：时间 - 连接词：仿佛 - 内容：静止了

C Statistics of rhetorical techniques in the original training set

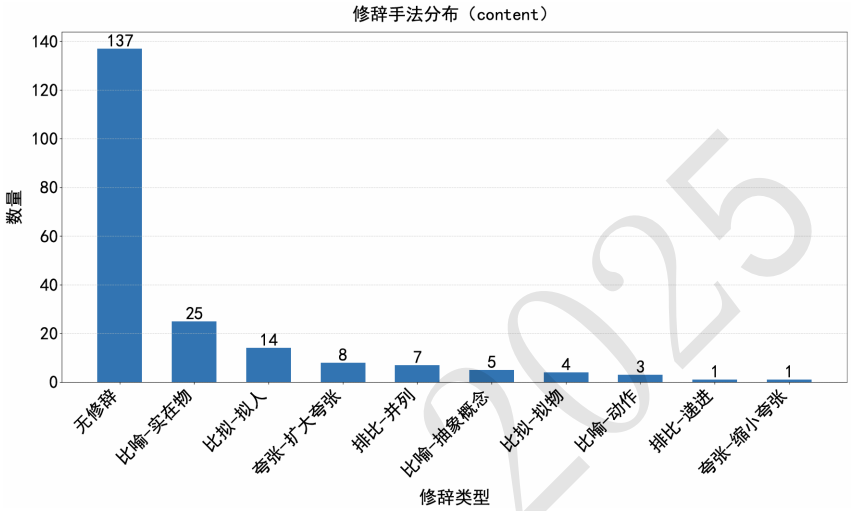

Figure 3: Distribution of Rhetorical Devices in Original Training Dataset (Track 2)

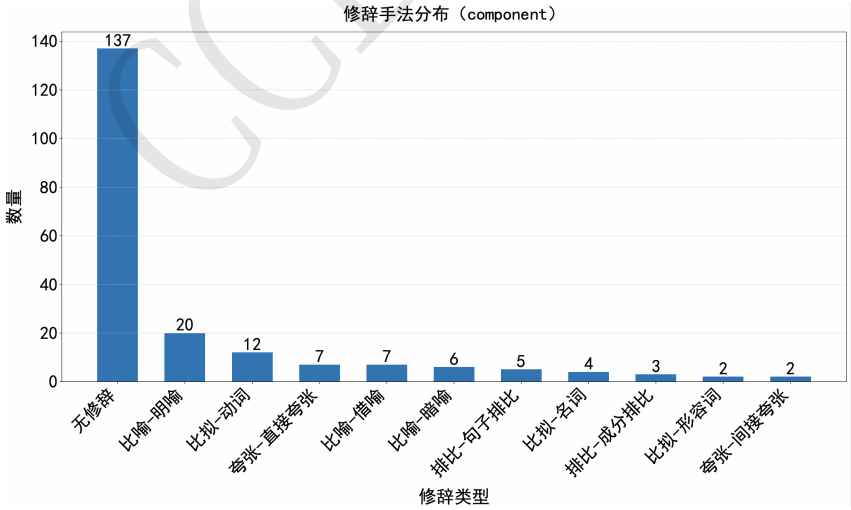

Figure 4: Distribution of Rhetorical Devices in Original Training Dataset (Track 3)