DISCERN: Chain-of-Thought-Augmented Syntactic-based In-Context Learning for Chinese Semantic Error Detection

Bowen Ruan[1], Hongyan Wu[2], Yitong Han[1], Shengyi Jiang[1], Lianxi Wang[1], Nankai Lin[1,†]

[1] *School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China*

[2] *College of Computer, National University of Defense Technology, Changsha, China*

Abstract

Semantic errors in Chinese text can significantly impact text comprehension and information accuracy. Although semantic error detection is crucial for improving the quality and reliability of texts, existing models still face major challenges in detection performance due to the diversity and complexity of semantic error types. To address this issue, we propose a chain-of-thought-augmente**D** syntact**I**c-ba**S**ed in-**C**ont**E**xt Lea**RN**ing framework (**DISCERN**), which aims to enhance the performance of Large Language Models' (LLMs) in Chinese semantic error detection tasks. DISCERN consists of three core modules: Semantic Error Mechanism Mining (SEMM), Demonstration Sample Selection (DSS), and Template Filling Output (TFO). It integrates Chain-of-Thought (CoT) prompting technique of LLMs with dependency syntax tree-based similarity calculation to select appropriate demonstration examples. Experiments on the CSED-R dataset demonstrate that compared to existing methods, DISCERN can effectively improve the performance of LLMs in semantic error detection tasks.

Keywords: In-context Learning; Chain-of-Thought; Chinese Semantic Error Detection; Large Language Models; Syntax Tree-based Similarity Calculation

1. Introduction

In today's era of digital transformation and deepening global communication, text, as the primary carrier of information transmission, directly impacts operational efficiency and decision-making quality across various sectors of society. With the proliferation of Artificial Intelligence (AI) writing assistants and automated content generation tools, the scale and speed of Chinese text production have grown exponentially. Efficiently analyzing and detecting semantic errors in these textual data has become a crucial research direction in Natural Language Processing (NLP). Semantic errors, as a special category of textual errors, not only involve fundamental theoretical issues in linguistics, such as lexical semantics and syntactic semantics, but are also closely related to language comprehension and information processing mechanisms in cognitive science. From an information theory perspective, semantic errors increase information entropy and reduce communication efficiency, potentially leading to severe economic losses, legal disputes, or life-threatening situations in professional domains such as medical diagnostic reports, legal

[†]Corresponding author: Nankai Lin (Email: neakail@outlook.com; ORCID: 0000-0003-2838-8273)

documents, technical specifications, and financial contracts. Compared to spelling and grammatical errors, semantic errors are more prominently manifested in complex syntactic structures and semantic relationships. Sentences containing semantic errors often appear superficially fluent, making them difficult for even human readers to detect. For example, the sentence "反复阅读这些优美的篇章、节奏、语气和韵味，有助于养成良好的文言语感" (Repeatedly reading these exquisite passages, rhythms, tones, and nuances helps cultivate a strong sense of Classical Chinese) is formally grammatically correct but exhibits a collocational mismatch: "节奏、语气和韵味" (rhythms, tones, and nuances) are improperly coordinated with "篇章" (passages), resulting in inconsistent semantic categories that undermine precision and may lead to misunderstanding. In practice, semantic errors commonly take the form of word order errors, missing constituents, improper collocations, redundancy, confusion, fuzziness, and illogicality. Such errors are not only difficult to detect with conventional text-editing tools but often require specialized knowledge and specific contextual information for accurate identification. Therefore, developing efficient semantic error detection technologies holds significant theoretical value in linguistics while providing technical support for enhancing social information exchange efficiency and reducing communication costs.

However, semantic error detection faces multiple technical challenges. First, the diversity and complexity of semantic errors make it difficult for traditional rule-based methods to comprehensively cover all error types. Second, semantic understanding itself is a complex cognitive process requiring models to possess strong language comprehension and knowledge reasoning capabilities. In recent years, Large Language Models (LLMs) have offered new solutions for semantic error detection through their powerful language understanding abilities and flexible In-Context Learning (ICL) capabilities. LLMs can quickly acquire task-relevant knowledge from few-shot examples without requiring large-scale training data and complex model fine-tuning. Nevertheless, selecting effective ICL demonstration samples to enhance model performance in error detection remains a significant challenge. Furthermore, since semantic error detection fundamentally focuses on semantic-level analysis, effectively integrating more fine-grained semantic information remains a critical issue to be resolved.

To address these challenges, we propose a chain-of-thought-augmente**D** syntact**I**c-ba**S**ed in-Cont**E**xt Lea**RN**ing framework (**DISCERN**). The framework comprises three core modules: Semantic Error Mechanism Mining module (SEMM), Demonstration Sample Selection module (DSS), and Template Filling and Output module (TFO). The SEMM module achieves precise detection, classification, and causal analysis of semantic errors by integrating the LLMs' Chain of Thought (CoT) prompting technique. The DSS module calculates semantic similarity between sentences through dependency tree kernel functions to ensure the selection of the most representative ICL demonstration samples. The TFO module ensures consistency and interpretability of model inputs and outputs through standardized template design. Experimental results demonstrate that DISCERN effectively improves the performance of large models in Chinese semantic error detection tasks. Our main contributions are summarized as follows:

1) We propose a novel Chinese semantic error detection framework DISCERN based on CoT and ICL. By integrating the advantages of these two techniques, the framework effectively enhances the LLMs' understanding and detection capabilities for complex semantic errors.

2) We design a demonstration sample selection mechanism based on dependency syntax trees, which effectively improves the accuracy of demonstration example selection.

3) We conduct extensive experiments on the CSED-R dataset. The results show that our method DISCERN enhances the LLMs' performance for the semantic error detection tasks.

2. Related Work

2.1. Chinese Semantic Error Detection

In the field of Chinese semantic error detection, while spelling and grammatical error detection have garnered widespread attention and research, studies on semantic errors remain relatively underdeveloped [1]. Existing research mainly revolves around three major technical approaches: rule-based methods, neural network-based methods, and pre-trained model-based methods.

Rule-based methods. In the early studies on semantic error detection, rule-based methods dominated the field. These methods rely on statistical theories and dependency analysis to identify collocation relationships between words by analyzing large-scale corpora [2]. While such methods can yield good results in certain scenarios, their major drawback lies in weak generalization capabilities when faced with complex semantic structures. In particular, rule-based methods often fall short in handling scenarios where semantic dependencies across sentences or even paragraphs are required [3].

Neural network-based methods. With the rapid development of deep learning technology, neural networks have gradually become the main tool for Chinese semantic error detection. Neural network models can automatically learn contextual information from large-scale data, thereby capturing complex semantic relationships [4]. PHMOSpell enhances the detection accuracy of spelling and semantic errors by incorporating phonetic and glyph information [4]. However, neural network-based models are still limited by the training data and model architecture when dealing with long-distance dependent semantic structures, leading to certain limitations in specific scenarios.

Pre-trained model-based methods. In recent years, the development of pre-trained models has greatly improved the performance of Chinese semantic error detection. The Desket model, which combines dependency syntax analysis [2], captures dependency relationships and part-of-speech information in sentences to achieve more accurate semantic error correction. Wu et al. propose a CSER method with the Dependency Syntactic Attention mechanism (CSER-DSA) to explicitly infuse dependency syntactic information only in the fine-tuning stage, achieving robust performance [5]. These pre-trained models can effectively handle long-distance semantic dependencies and demonstrate strong generalization capabilities, making them adaptable to more complex semantic structures and diverse linguistic environments. By leveraging the learning from large-scale corpora, pre-trained models have significantly improved the accuracy and robustness of semantic error detection [6, 7, 8].

2.2. In-context Learning

ICL is a learning paradigm that has attracted widespread attention with the emergence of LLMs [9, 10]. Typically, ICL prompts LLMs with contextual examples, enabling them to learn tasks from only a few examples. Positive impacts of ICL on LLMs have been observed in various tasks such as text classification and answering [11, 12], images generations [13], speech tasks [14], and multi-modal scenarios [15, 16]. Recent works have aimed to enhance ICL by selecting valuable demonstrations [17, 18], optimizing the order of demonstrations [19], etc.

To the best of our knowledge, there is no previous work investigating the potential of ICL for Chinese semantic error detection tasks. Existing ICL demonstration selection methods struggle to incorporate fine-grained semantic information, and thus the ICL capabilities of LLMs have not been fully leveraged in error detection tasks requiring deep semantic understanding. Our

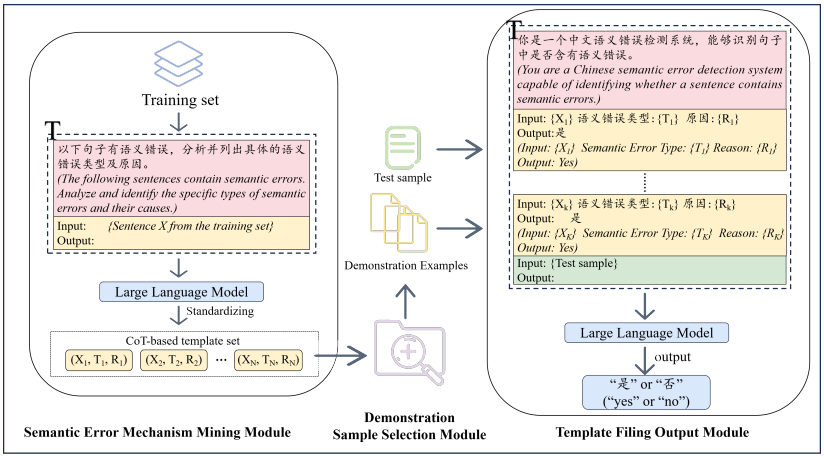

Figure 1: DISCERN framework. DISCERN first identifies and classifies semantic errors in training sentences through the SEMM module using LLM's CoT prompting technique. Next, the DSS module selects appropriate ICL examples for each test sample based on dependency tree similarity computation. Finally, in the TFO module, DISCERN populates a predefined template with selected examples to assist LLMs in Chinese semantic error detection. The text in parentheses provides the English translation of the original prompt.

work succeeds in addressing this limitation to some extent and confirms the potential of ICL for Chinese semantic error detection.

3. Methods

3.1. Overview

We propose a chain-of-thought-augmente**D** syntact**I**c-ba**S**ed in-**C**ont**E**xt Lea**RN**ing framework (**DISCERN**), which comprises three primary modules: Semantic Error Mechanism Mining (SEMM) module, the Demonstration Sample Selection (DSS) module, and the Template Filling and Output (TFO) module. The complete framework flow chart is presented in Fig. 1. In SEMM, we utilize LLM's CoT prompting technique to identify and classify semantic errors in sentences and extract specific reasons for the error. By populating predefined templates with sentences from the training set and feeding them into LLMs, we obtain specific error types and their underlying causes, which are then standardized into structured formats for subsequent analysis. DSS aims to identify appropriate ICL examples based on semantic similarity between sentences. We employ tree kernel functions to compute the similarity between dependency trees extracted from sentences. Finally, in TFO, we design a predefined template to standardize the input format for the LLMs. The constructed template is subsequently fed into the LLMs, which determines whether the test samples contain any semantic errors.

3.2. Semantic Error Mechanism Mining

CoT is a prompting technique that guides LLMs through step-by-step reasoning, enhancing their performance on complex tasks by encouraging them to demonstrate their reasoning process. To effectively identify and classify semantic errors in sentences, we employ the CoT prompting technique of LLMs. Specifically, we input sentences X from the training set into LLMs. For

sentences containing semantic errors (i.e., samples labeled as "是(yes)"), we encapsulate them into a predefined template. This template explicitly instructs the model to confirm the presence of semantic errors and requires it to analyze and enumerate specific types of semantic errors and their causes, as shown in Fig. 2. Subsequently, the LLMs output a binary label $C \in \{0, 1\}$, where 0 and 1 denote the absence and presence of semantic errors respectively, the semantic error classification T, and a detailed explanation of the error reason R. For each instance i in the training set where the LLMs outputs C equals 1 indicating semantic errors, we construct a structured template $Template_i = (X_i, T_i, R_i)$. These templates are aggregated into a set $\mathcal{T} = \{\text{Template}_i\}_{i=1}^{N}$, where N denotes the total number of semantic error instances.

Instruction	以下句子有语义错误，分析并列出具体的语义错误类型及原因。 *(The following sentences contain semantic errors. Analyze and identify the specific types of semantic errors and their causes.)*
I/O Format	Input: { X } Output:

Figure 2: Prompt template for semantic error mechanism mining. The text in parentheses provides the English translation of the original prompt.

3.3. Demonstration Sample Selection

For each sentence in both the training and test sets, we first perform word segmentation and extract its dependency tree. Then, for each dependency tree of sentences in the test set, we utilize tree kernel functions to calculate similarity scores between it and the dependency trees of all sentences in the training set, selecting the top K sentences with the highest similarity scores as its demonstration samples.

3.3.1. Construction of Dependency Syntax Trees

In this study, we employ the Language Technology Platform (LTP) [20], which is a comprehensive Chinese natural language processing toolkit providing extensive functionalities. Specifically, we utilize LTP for word segmentation and dependency parsing, constructing syntax trees to capture the grammatical and semantic relationships between word segments in sentences. For each sentence S in the training and test sets, we first segment it into a sequence of words $W = \{w_1, w_2, \ldots, w_n\}$, where each w_i represents a word token in the sentence. Subsequently, LTP identifies the dependency relations for each word token w_i and assigns it a dependency label d_i, which indicates the relationship type between w_i and its head word. These dependency relations can be formalized as a set of binary tuples $D = \{(h_i, r_i)\}_{i=1}^{n}$, where h_i denotes the head word index of w_i, and r_i represents the type of dependency relation. Based on this information, we can construct a syntax tree T, mathematically represented as a Directed Acyclic Graph (DAG) $T = (V, E)$, where the node set $V = \{v_1, v_2, \ldots, v_n\}$ corresponds to words w_i, and the edge set $E = \{(v_{h_i}, v_i, r_i)\}_{i=1}^{n}$ represents the dependency relation r_i from the head word v_{h_i} to the dependent word v_i.

3.3.2. Subtree Kernel Function

The subtree kernel function is a convolutional kernel method used to measure the similarity between two trees [21]. Its core idea is to compare the subtrees of two syntax trees and calculate

the number of shared substructures to quantify their similarity. For two dependency trees, T_1 and T_2, we utilize the following subtree kernel function formula to calculate their similarity score:

$$K(T_1, T_2) = \sum_{n_1 \in N_1} \sum_{n_2 \in N_2} k(n_1, n_2), \tag{1}$$

where N_1 and N_2 are the node sets of dependency trees $T1$ and $T2$, respectively, and $k(n_1, n_2)$ is the recursive formulation of the subtree kernel function that computes the similarity between subtrees rooted at nodes n_1 and n_2.

3.3.3. Recursive Definition

(1) If the labels of nodes n_1 and n_2 differ, then $k(n_1, n_2)$=0, ensuring that only nodes with the same label contribute to the similarity.

(2) If both nodes are leaf nodes and their labels match, then $k(n_1, n_2)$=λ, where λ ($0 < \lambda < 1$) is a decay factor that controls overall similarity's impact by matching leaf nodes.

(3) If the labels of two nodes are the same and they have child nodes, then:

$$k(n_1, n_2) = \lambda \prod_{i=1}^{l} (1 + k(c_{1i}, c_{2i})), \tag{2}$$

where l is the number of child nodes, and c_{1i} and c_{2i} represent the i-th child nodes of n_1 and n_2, respectively.

3.4. Template Filling and Output

To facilitate effective interaction with LLMs in semantic error detection tasks, we construct a predefined template to standardize the input format. This template aims to provide clear instructions to LLMs while incorporating K demonstration samples selected through DSS. Our template format is shown in Fig. 3. Finally, we input the constructed template into the large model to allow the model to determine whether the test sample contains semantic errors.

Instuction	你是一个中文语义错误检测系统，能够识别句子中是否含有语义错误。 *(You are a Chinese semantic error detection system capable of identifying whether a sentence contains semantic errors.)*
ICL example	Input: {X_1} 语义错误类型：{T_1} 原因：{R_1} Output: 是 Input: {X_2} 语义错误类型：{T_2} 原因：{R_2} Output: 是 ⋮ Input: {X_K} 语义错误类型：{T_K} 原因：{R_K} Output: 是 *(Input: {X_1} Semantic Error Type: {T_1} Reason: {R_1} Output: Yes* *Input: {X_2} Semantic Error Type: {T_2} Reason: {R_2} Output: Yes* ⋮ *Input: {X_K} Semantic Error Type: {T_K} Reason: {R_K} Output: Yes)*
I/O Format	Input: {Test sample} Output:

Figure 3: Prompt template for semantic error detection. The text in parentheses provides the English translation of the original prompt.

4. Experiments

4.1. Experimental Settings and Datasets

We evaluate DISCERN and other ICL methods on ChatGLM3-6B[1] and DeepSeek-R1[2] [22]. Both models are selected for their strong Chinese understanding and ICL abilities, achieving a balance between computational efficiency and reasoning depth. We perform experimental testing using the Chinese Semantic Error Diagnosis Recognition (CSED-R) dataset built by Sun et al. [1]. This dataset contains 45,248 training samples, 2,160 validation samples, and 2,000 test samples, which contains richer semantic error types compared to other existing datasets. All experiments are conducted on an RTX8000.

4.2. Evaluation Metrics

In all experiments, precision, recall, F1-score, and accuracy are regarded as the most crucial performance metrics. Higher values of these metrics correspond to better performance.

4.3. Comparison Methods

To assess the effectiveness of DISCERN, we perform comparative analyses using several established methods in this field.

Universal Pre-trained Model-Based Methods. Universal pre-trained model-based methods refer to the capability of universal pre-trained language models to perform semantic error detection tasks without explicit training on specific tasks or dataset examples [23].

Syntax-Infused Fine-tuning Methods. Models [24] enhanced with syntactic information serve as the foundational architecture. These models are subsequently fine-tuned on the CSED-R dataset to improve their adaptation to the semantic error detection tasks. To evaluate the effectiveness of different approaches, various fine-tuning strategies are employed, including CSER-DSA [5] as well as Syntax-RoBERTa [25] and its variants that incorporate additional syntactic information.

RoBERTa Pre-training with Syntax-related Task Approach. RoBERTa models are initially pre-trained using one million Wikipedia articles, utilizing the LTP tool to perform syntactic parsing and obtain sentence dependency structures. Subsequently, specialized pre-training tasks, such as DP and DP$^+$, are designed based on these dependency structures, enabling the models to acquire syntactic knowledge during the pre-training phase. For experiments on the CSED-R dataset, various pre-training strategies, including RoBERTa+DP and RoBERTa+DP$^+$ [1], are implemented and their performance is compared against baseline models.

Jaccard-based ICL. Jaccard-based ICL [26] method refers to a word similarity-based example selection approach for ICL, which leverages the Jaccard coefficient to identify the most relevant demonstration examples for test samples by computing the ratio between intersection and union of segmented sentence sets.

[1]https://huggingface.co/THUDM/chatglm3-6b-base
[2]https://www.ollama.com/SIGJNF/deepseek-r1-671b-1.58bit

Model	P	R	F_1	ACC
RoBERTa [24]	72.9	72.4	72.6	72.7
MacBERT [27]	72.3	75.3	73.7	73.1
SLA [28]	72.8	73.0	72.9	72.9
Syntax-RoBERTa[25]	73.3	74.3	73.8	73.6
RoBERTa+DP [1]	74.2	74.4	74.3	74.3
RoBERTa+DP$^+$ [1]	73.2	75.8	74.8	74.1
SLA+DP [1]	72.1	77.1	74.5	73.6
SLA+DP$^+$ [1]	72.0	76.9	74.4	73.5
Syntax-RoBERTa+DP [1]	73.7	75.9	74.8	74.4
Syntax-RoBERTa+DP$^+$ [1]	73.6	76.1	74.8	74.4
CSER-DSA [5]	75.6	74.8	75.2	75.3
ChatGLM3	51.2	98.8	67.5	52.4
Jaccard-based ICL (ChatGLM3) [26]	51.1	96.2	66.8	52.2
DISCERN (ChatGLM3)	52.6	94.5	67.6	54.6
DISCERN w/o SEMM (ChatGLM3)	51.7	98.5	67.8	53.2
DeepSeek-R1	51.6	72.8	60.4	52.1
Jaccard-based ICL (DeepSeek-R1)	53.1	92.7	67.5	55.5
DISCERN (DeepSeek-R1)	54.7	87.2	67.2	57.5
DISCERN w/o SEMM (DeepSeek-R1)	52.1	90.8	66.2	53.7

Table 1: Performance comparison of different models

4.4. Comparison Experimental Results

The comparison experimental results are shown in Table 1. The results demonstrate that LLMs exhibit relatively inferior performance in Chinese semantic error detection tasks. For instance, ChatGLM3 achieved an accuracy of 52.4% and an F1-score of 67.5%, showing a significant performance gap compared to specialized models such as CSER-DSA and Syntax-RoBERTa+DP. For the DeepSeek-R1 model, its accuracy and F1-score are also significantly lower than those of the specialized models. This indicates that purpose-designed pre-trained language models possess substantial advantages in Chinese semantic error detection tasks.

In our analysis of different enhancement approaches for ChatGLM3, the Jaccard-based ICL (ChatGLM3) achieved an accuracy of 52.2% and an F1-score of 66.8%, showing slight decreases of 0.2% and 0.7% respectively compared to the original ChatGLM3 model. In contrast, DISCERN (ChatGLM3) attained an accuracy of 54.6% and an F1-score of 67.6%, representing improvements of 2.2% and 0.1% respectively over the original ChatGLM3 model. For the DeepSeek-R1 model, DISCERN (DeepSeek-R1) achieved an accuracy of 57.5% and an F1-score of 67.2%, marking significant improvements of 5.4% and 6.8% respectively compared to the original DeepSeek-R1 model. When compared to Jaccard-based ICL (DeepSeek-R1), which achieved an accuracy of 55.5% and an F1-score of 67.5%, DISCERN (DeepSeek-R1) demonstrated a 2.0% increase in accuracy while maintaining comparable F1-score performance.

Overall, DISCERN demonstrates significant enhancement in model performance, surpassing traditional Jaccard-based ICL approaches. This improvement can be attributed to DISCERN's comprehensive consideration of both semantic and syntactic relationships during example selection, as well as the integration of additional semantic information through CoT prompting technique, thereby enhancing the model's logical understanding and detection capabilities for semantic errors.

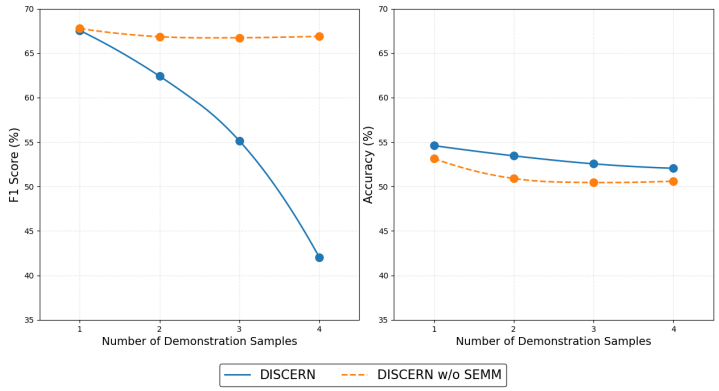

Figure 4: Variations in the performance of each method in response to changes in the K value.

Template A 请识别下列句子中是否含有语义错误，如果有请回答是，否则请回答否。 Please identify whether there are any semantic errors in the following sentence. If yes, please answer “yes”; if no, please answer “no”.
Template B 你的任务是检测以下句子中是否存在语义错误。若存在则回答是，若句子没有问题，则回答否。 Your task is to detect whether there are any semantic errors in the following sentence. If an error exists, answer ”yes”; if the sentence is correct, answer ”no”.

Table 2: Different prompt templates for semantic error detection

4.5. Parameter Study Experiment

We further investigate the impact of the number of ICL demonstration samples K on the performance of the model. For this purpose, we conduct experiments with K values ranging from 1 to 4. At each value of K, we test under consistent computational resource configurations and model parameter settings to ensure the comparability of the results. These experiments are conducted on the ChatGLM3-6B. As shown in Figure 4, the results indicate that, with the increase in K, the DISCERN framework equipped with the SEMM module exhibits a significant monotonic decrease in both the F1 score and accuracy metrics. In contrast, the DISCERN framework without the SEMM module (DISCERN w/o SEMM) shows only a slight decrease in these metrics, with overall performance remaining relatively stable. This phenomenon reveals that after incorporating the SEMM module, increasing the number of ICL demonstration samples actually leads to an information redundancy effect (IRE). The introduction of excessive samples brings additional feature noise, negatively impacting the model’s performance in semantic error detection tasks. This finding, which has significant practical implications for selecting the appropriate number of ICL samples, suggests that when utilizing the ICL technique, it is necessary to balance the trade-off between sample scale and model performance to avoid performance degradation caused by excessive samples.

Sentences	Ans.	Orig.	Ours
Case 1			
无论做什么事情，保持清醒的头脑和认真的态度都是成功的关键所在。 Whatever you do, maintaining a clear mind and serious attitude is the key to success.	是 yes	否 no	是 yes
形象的核心，是为风和土地、河流和黎明不懈地歌唱，死后连羽毛也奉献给土地的一只多情鸟。 The core of the image is a passionate bird that tirelessly sings for the wind, land, rivers, and dawn, and even dedicates its feathers to the earth after death.	是 yes	否 no	是 yes
中国是嫦娥的故乡，火箭的发源地，诞生了人类'真正的航天始祖'万户的国度。 China is the homeland of Chang'e, the birthplace of rockets, and the nation that gave birth to Wan Hu, humanity's 'true pioneer of space travel'.	是 yes	否 no	是 yes
Case 2			
孔子所创立的儒家学术成为了中国古代社会的正统思想。他还将鲁国史官所记《春秋》加以删修，成为我国第一部编年体史书。 Confucius established Confucianism, which became the orthodox ideology of ancient Chinese society. Additionally, he edited and revised The Spring and Autumn Annals, originally recorded by the official historians of the State of Lu, transforming it into China's first chronicle-style historical work.	是 yes	否 no	否 no
我们纪念季羡林先生的最好方式，莫过于学习他的独立人格、淡泊心态与严谨治学，传承季老具有的学术精神。 The best way to commemorate Mr. Ji Xianlin is to learn from his independent personality, detached mindset, and rigorous scholarship, and to carry forward his academic spirit.	是 yes	否 no	否 no
为遏制不断攀升的事故数量，新版《机动车驾驶证申领和使用规定》通过提高违法成本的方式引导司机文明驾驶，确实必要。 To curb the rising number of accidents, the new 'Regulations on the Application and Use of Motor Vehicle Driver's Licenses' guides civilized driving by increasing the cost of violations, which is indeed necessary.	是 yes	否 no	否 no

Table 3: A case study of model outputs before and after integrating DISCERN

4.6. Investigation of Different Templates

In the context of LLMs' limited semantic understanding capabilities, the design of prompt templates plays a key role in determining whether models can correctly understand and solve semantic error detection tasks. To this end, we designed two different prompt templates and explored their impact on model output.

Experimental results show that when using the templates in Table 2, regardless of the value of K, the predicted results of the model are consistently "否(no)", suggesting that in semantic error detection tasks the model is very sensitive to changes in prompt words and may not correctly understand the task requirements. In contrast, using our initially designed prompt template that includes examples, the model can correctly differentiate sentences with semantic errors. This emphasizes the critical importance of accurate template design in improving model performance.

4.7. Case Study

We conduct case studies to investigate the effectiveness and limitations of DISCERN. As shown in Table 3, Case 1 presents three examples where DISCERN succeeds in providing correct answers whereas the original model falls short. In Case 2, neither DISCERN nor the original model provides the correct answer for the three examples. The results from Case 1 demonstrate that DISCERN can significantly improve LLMs' performance in semantic error detection through two key mechanisms. DISCERN selects example sentences that are highly similar both semantically and syntactically to the test samples by computing similarity scores based on dependency parse trees. LLMs can then perform ICL through these example sentences, thereby improving their performance in semantic error detection tasks. Additionally, the integration of CoT prompting technique provides the model with additional semantic information, enhancing its understanding of complex semantic relationships. However, the analysis of Case 2 reveals several limitations of DISCERN. It still faces challenges in handling complex semantic relationships in lengthy sentences. For instance, in the first sentence of Case 2, the model fails to

correctly resolve references between entities. This indicates that the model has limited capability in understanding deep semantic relationships such as anaphora, ellipsis, implicit logic, and discourse cohesion.Additionally, when encountering sentences with intricate semantic relationships, CoT reasoning often fails to resolve the issue in a single step. If the analytical content is suboptimal or redundant, it may hinder the model's grasp of the task requirements. In the future, vector-based semantic representations from pre-trained language models can be integrated to compensate for the deficiencies of traditional dependency-based similarity measures in complex sentences. Furthermore, by designing task-decomposition-based CoT strategies, complex problems that are difficult to resolve in one step can be broken down into smaller, interpretable reasoning chains, thus improving the depth and reliability of the reasoning process.

5. Conclusion

In this paper, we propose a chain-of-thought-augmented syntactic-based in-context learning framework DISCERN for Chinese semantic error detection tasks. The proposed framework integrates the CoT prompting technique with dependency syntax tree-based similarity calculation to select appropriate demonstration examples, aiming to enhance LLMs' performance in detecting complex semantic errors. Experimental results on the CSED-R dataset demonstrate that DISCERN achieves promising results in improving the accuracy of semantic error detection, particularly when the original model output is inaccurate. While showing effectiveness, the framework still faces challenges in example selection strategies and handling complex semantic relationships, which provides valuable directions for future research to further optimize the performance of LLMs in semantic error detection tasks.

Author Contributions

Bowen Ruan: Writing-Original draft preparation, Methodology, Investigation, Writing-Reviewing and Editing. Hongyan Wu: Methodology, Writing-Reviewing and Editing. Yitong Han: Writing-Reviewing and Editing. Shengyi Jiang: Writing-Reviewing and Editing. Lianxi Wang: Writing-Reviewing and Editing. Nankai Lin: Conceptualization of this study, Methodology, Writing-Reviewing and Editing.

Acknowledgements

Our work is supported by the National Social Science Fund of China (No. 22BTQ045), the Research Fund of National Language Commission (No. YB145-123) and the College Students' Innovative Entrepreneurial Training Plan Program of Guangdong University of Foreign Studies.

References

[1] B. Sun, et al., Improving pre-trained language models with syntactic dependency prediction task for chinese semantic error recognition, 2022. arXiv:2204.07464.

[2] W. Huang, et al., Csec: A chinese semantic error correction dataset for written correction, in: B. Luo, L. Cheng, Z.-G. Wu, H. Li, C. Li (Eds.), Neural Information Processing, Springer Nature Singapore, 2024, pp. 383–398.

[3] R. Zhang, et al., Research on proofreading method of semantic collocation error in chinese, Advances in Artificial Intelligence and Security (2021).

[4] L. Weihua, L. Zhensheng, G. Xiaojin, Semantic error checking in automatic proofreading for chinese texts, in: IEEE International Conference on Systems, Man and Cybernetics, volume 7, 2002, pp. 1–5.

[5] H. Wu, et al., Pseudo-label data construction method and syntax-enhanced model for Chinese semantic error recognition, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 5391–5402. URL: https://aclanthology.org/2025.coling-main.361/.
[6] Y. Wu, et al., Research and realization of chinese text semantic correction based on rule, in: Proceedings of the 2015 3rd International Conference on Education, Management, Arts, Economics and Social Science, Atlantis Press, 2015/12, pp. 1394–1404.
[7] X. Tan, G. Deng, X. Hu, Multi-granularity context semantic fusion model for chinese event detection, in: 2021 10th International Conference on Internet Computing for Science and Engineering, ICICSE 2021, Association for Computing Machinery, 2022, p. 1–7.
[8] A. Garcia-Garcia, et al., A review on deep learning techniques applied to semantic segmentation, 2017. arXiv:1704.06857.
[9] Brown, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.
[10] P. Liu, et al., Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023).
[11] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, 2021. arXiv:2012.15723.
[12] J. Liu, et al., What makes good in-context examples for gpt-3?, 2021. arXiv:2101.06804.
[13] A. Bar, et al., Visual prompting via image inpainting, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 25005–25017.
[14] Z. Zhang, et al., Speak foreign languages with your own voice: Cross-lingual neural codec language modeling, 2023. arXiv:2303.03926.
[15] S. Huang, et al., Language is not all you need: Aligning perception with language models, 2023. arXiv:2302.14045.
[16] J. Wei, et al., Emergent abilities of large language models, 2022. arXiv:2206.07682.
[17] S. An, et al., Skill-based few-shot selection for in-context learning, 2023. arXiv:2305.14210.
[18] O. Rubin, J. Herzig, J. Berant, Learning to retrieve prompts for in-context learning, 2022. arXiv:2112.08633.
[19] Y. Lu, et al., Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, 2022. arXiv:2104.08786.
[20] W. Che, Z. Li, T. Liu, Ltp: a chinese language technology platform, in: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, COLING '10, Association for Computational Linguistics, 2010, p. 13–16.
[21] M. Collins, N. Duffy, Convolution kernels for natural language, in: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01, MIT Press, 2001, p. 625–632.
[22] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. arXiv:2501.12948.
[23] J. Devlin, et al., Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
[24] J. Yuan, J. Vig, N. Rajani, isea: An interactive pipeline for semantic error analysis of nlp models, in: Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22, Association for Computing Machinery, 2022, p. 878–888.
[25] J. Bai, et al., Syntax-BERT: Improving pre-trained transformers with syntax trees, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, 2021, pp. 3011–3020.
[26] J. Wang, et al., Knowledgeable in-context tuning: Exploring and exploiting factual knowledge for in-context learning, in: Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, 2024, pp. 3261–3280.
[27] Y. Cui, et al., Revisiting pre-trained models for chinese natural language processing, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020.
[28] Z. Li, et al., Improving BERT with syntax-aware local attention, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, 2021, pp. 645–653.