MADAWSD: Multi-Agent Debate Framework for Adversarial Word Sense Disambiguation

Kaiyuan Zhang[1], Qian Liu[2], Luyang Zhang[1], Chaoqun Zheng[1], Shuaimin Li[3], Bing Xu[4], Muyun Yang[4], Xinxiao Qiao[1,*], Wenpeng Lu[1,*]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]University of Auckland, Auckland, New Zealand [3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China [4]Harbin Institute of Technology, Harbin, China

kaiyuan.zhang@foxmail.com, wenpeng.lu@qlu.edu.cn

Abstract

Word sense disambiguation (WSD) is a fundamental yet challenging task in natural language processing. In recent years, the advent of large language models (LLMs) has led to significant advancements in regular WSD tasks. However, most existing LLMs face two major issues that hinder their performance in WSD. Firstly, these models are often prone to misclassifying the correct meaning of an ambiguous word when confronted with contexts containing adversarial information. Secondly, there is a lack of sufficient adversarial WSD datasets, which severely limits the development and evaluation of adversarial WSD systems. To address these gaps, we propose a novel Multi-Agent Debate framework for Adversarial Word Sense Disambiguation (MADAWSD). The MADAWSD framework simulates a real-world debate environment where multiple agent roles, namely, the Debater, Moderator, Consensus-seeker, and Judge, engage in discussions about ambiguous words in the context of adversarial information. Through a collaborative mechanism among these agents, it achieves accurate WSD. Additionally, a novel dataset for Chinese adversarial WSD has been constructed, focusing on improving and evaluating the performance of WSD models in the Chinese language. Extensive experiments on both English and Chinese adversarial WSD datasets demonstrate that MADAWSD can seamlessly integrate with existing LLMs and significantly enhance their performance, showcasing broad generality and outstanding effectiveness.[1]

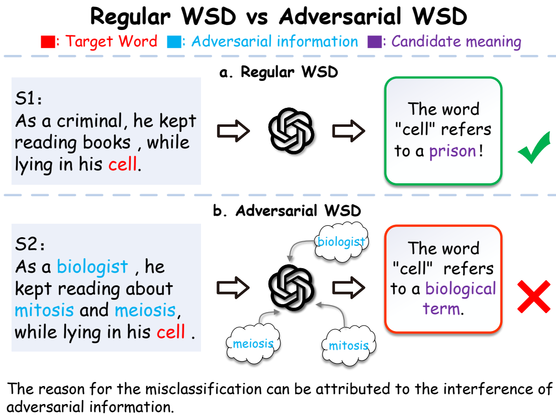

Figure 1: In the regular WSD task, a model demonstrates the ability to correctly identify the term *cell* as referring to the meaning *prison*. However, when confronted with contexts containing adversarial information, such as *biologist*, *mitosis* and *meiosis*, the model tends to misclassify *cell* as referring to a *biological term*. This highlights how the performance of WSD models is significantly affected by adversarial information.

1 Introduction

Word sense disambiguation (WSD) is the task of identifying the correct meaning of an ambiguous word in a specific context, which is a fundamental and challenging task in natural language processing (NLP) (Navigli, 2009; Wang et al., 2024a). WSD plays an essential role for human language understanding, which is critical for a range of downstream tasks, such as machine translation (Martelli et al., 2025; Tran et al., 2025), sentiment analysis (Zhang et al., 2023), and information retrieval (Dadure et al., 2024; Blloshmi et al., 2021).

Recently, large language models (LLMs) have emerged as the dominant paradigm for WSD (Kritharoula et al., 2023; Yang et al., 2024), surpassing earlier approaches such as lexicon-based methods (Agirre et al., 2014; Zhang et al., 2022; Bevilacqua and Navigli, 2020), machine learning (Wang et al., 2025), and fine-tuned pre-trained models (Huang et al., 2019). This success is largely attributed to the advanced abilities of LLMs for contextual understanding and generalization. However, despite their effectiveness, LLMs

* Corresponding author

[1]The full source code and datasets are publicly available at https://github.com/KaiyCheung/MADAWSD.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22295–22314
November 4-9, 2025

remain vulnerable to adversarial contexts, where misleading lexical cues can influence sense predictions. These cues may evoke irrelevant semantic associations, resulting in incorrect disambiguation. As shown in Figure 1, LLMs correctly identify *"cell"* as *"a prison room"* in a neutral context (S1); while, in an adversarial context (S2), the presence of distractors such as *"biologist"*, *"mitosis"*, and *"meiosis"* misleads the model into predicting the *"biological sense of cell"* instead. This highlights the need for adversarial WSD methods that are robust to semantic interference and capable of precise, context-sensitive reasoning.

To address this challenge, we propose a novel multi-agent framework for adversarial WSD, where agents engage in a structured debate focused on the rationality of different senses within adversarial contexts. Specifically, the process begins with the Debater, who presents arguments supporting a particular sense. The Moderator then facilitates the exchange and ensures the debate progresses logically. The Consensus-seeker assesses whether a consensus has been reached; if not, the discussion continues until a consensus is achieved or a maximum number of rounds is reached. In the latter case, a Judge is invoked to make the final decision. This debate-driven reasoning encourages the model to focus on rational support for each sense, thereby reducing susceptibility to misleading lexical cues.

Moreover, existing benchmarks for adversarial WSD remain limited, with only one English dataset named FOOL (Ballout et al., 2024) explicitly designed for adversarial evaluation. To advance this line of research, we introduce a high-quality Chinese adversarial WSD benchmark, named CAWSD, filling a critical gap in multilingual evaluation. In the FOOL dataset, ambiguous words have two meanings. Therefore, to maintain consistency with FOOL, we constructed the CAWSD dataset using the same approach, with ambiguous words also having two meanings. Our experiments show that the proposed framework achieves significant performance gains (up to 11.3%) across both open-source and proprietary LLMs in both English and Chinese settings, demonstrating its effectiveness and robustness. The main contributions of this work are summarized below:

- We propose a novel multi-agent debate framework for adversarial WSD (MADAWSD). It leverages LLM-based agents to engage in debates on ambiguous words, handling adversarial information in contexts and accurately differentiating the correct meaning. To the best of our knowledge, this is the first framework that leverages LLM-based multi-agents to achieve adversarial WSD.
- We construct and release a novel CAWSD dataset, a WSD dataset containing adversarial information. It effectively alleviates the problem of data scarcity for adversarial WSD datasets, enhances linguistic diversity, and significantly facilitates the evaluation of adversarial WSD systems.
- Extensive experiments on adversarial WSD datasets demonstrate that MADAWSD can seamlessly integrate with existing LLMs and significantly enhance their performance on both English and Chinese adversarial datasets, thereby showcasing its broad generality and outstanding effectiveness.

2 Related Work

2.1 Word Sense Disambiguation

WSD determines the most appropriate meaning for a target ambiguous word based on its context. Over the past few decades, numerous methods have been proposed to enhance WSD performance. For instance, prior research (Agirre et al., 2014) introduced a WSD algorithm based on random walks, grounded in a large lexical knowledge base. In recent years, machine learning techniques and deep neural networks have been extensively employed to improve WSD accuracy (Luo et al., 2018; Lu et al., 2019; Bevilacqua and Navigli, 2020; Zheng et al., 2021). For example, ESC (Barba et al., 2021) defined WSD as an extractive task, KELESC (Zhang et al., 2022) introduced local attention and knowledge enhancement models to improve the representation of meaning-related words for WSD. LTRS (Wang et al., 2025) employed a ranking learning method to account for the influence of similar word meanings, thereby enhancing WSD.

More recently, as LLMs continue to evolve, they are increasingly being employed to facilitate WSD and to enhance its accuracy. Kwon et al. (2023) utilized LLMs to generate word definitions, improving the accuracy of visual WSD. A recent study (Kritharoula et al., 2023) leveraged LLMs as a knowledge base to enhance a given phrase and employed Chain-of-Thought (CoT) prompting to facilitate the generation of explainable answers.

PolCLIP (Yang et al., 2024) applied image-text complementarity strategy and semantic enhancement using LLMs for WSD. Despite their significant success, these approaches are primarily focused on regular WSD rather than adversarial WSD. When confronted with contexts containing adversarial information, they are prone to misclassifying the meanings of ambiguous words.

2.2 LLM-based Multi-agent Collaboration

In recent years, researchers have recognized the limitations of single models, which has prompted the development of various multi-agent systems (Xi et al., 2025). The key mechanism of LLM-based multi-agent collaboration lies in replicating human-like processes, specifically through role-playing (Hong et al., 2023) and collaboration (Li et al., 2023). Recent works have focused on enhancing performance by leveraging the collective strengths of multiple cognitive entities. Notably, the Solo Performance Prompting (SPP) (Wang et al., 2024b) has effectively utilized diverse perspectives by dynamically identifying and engaging various roles throughout the problem-solving process.

As an extension of these studies, recent research has explored the incorporation of adversarial cooperation strategies, including debates and negotiations among multiple agents to enhance performance. For example, MedAgents (Tang et al., 2024) employed LLM-based agents to participate in collaborative rounds of discussions in a role-playing environment, fully leveraging the potential of LLMs in certain specific domains and improving their proficiency and reasoning capabilities. In the MAD framework (Liang et al., 2024), multi-agents present their views in a tit-for-tat fashion, achieving significant advancements in machine translation and counter-intuitive reasoning tasks, while also pioneering a novel approach to multi-agent cooperation. Given the presence of adversarial information, we draw inspiration from this framework and adapt it for adversarial WSD.

3 Method

3.1 Task definition

Specifically, given the context of a target ambiguous word and a set of candidate meanings, the objective of WSD is to identify which candidate meaning best aligns with the context. Formally, we represent the context containing the target word as $C = \{w_1, \cdots, w_{target}, \cdots, w_m\}$, where w_i represents the i-th word in the context, w_{target} represents the target ambiguous word, and m represents the number of words in the context. The set of candidate meanings of the target word is denoted as $S = \{s_1, \cdots, \hat{s}, \cdots, s_n\}$, where s_j represents a candidate sense, $\hat{s}$ represents the correct sense of the target word in the given context, and n denotes the number of candidate senses.

3.2 MADAWSD Architecture

In this work, to tackle the WSD task, we organize the context C containing the target word into a debate topic and input it into the proposed MADAWSD framework. Through several rounds of argumentation among multiple agents, MADAWSD evaluates and compares the candidate meanings S, ultimately determining the meaning $\hat{s}$ that best fits the context as the final answer.

As shown in Figure 2, the MADAWSD framework consists of four key roles: *Debater*, *Moderator*, *Consensus-seeker*, and *Judge*. Each role is responsible for a different stage of the MADAWSD workflow. (1) *Debater*: This role is responsible for debating the meaning of the target ambiguous word in the given context. Debaters analyze the possible meaning of the ambiguous word according to contextual information and present arguments to support their viewpoints. They provide their own answers in each round of the debate and adjust them based on the viewpoints of other Debaters. (2) *Moderator*: This role is charged with guiding the progression of the debate. The Moderator analyzes the reasons and answers from all Debaters, concludes, and expresses its own preference for the correct meaning. The Moderator can guide the debate to proceed to the next round, or directly conclude the debate by determining the most appropriate meaning of the ambiguous word. Additionally, to ensure the standard format of the Moderator's output, we introduce a *Fixer* role, which monitors the output of the Moderator and converts it into a standard format, facilitating the operations of subsequent roles. (3) *Consensus-seeker*: This role is designed to Judge whether a consensus has been reached between the debating parties. If Debaters have reached a consensus on the meaning of the ambiguous word, the Consensus-seeker will adopt either side's answer as the final answer, and trigger the termination of the debate. (4) *Judge*: This role is responsible for rendering the final decision on the ambiguous word's meaning when the debate reaches its maximum number of rounds. The Judge

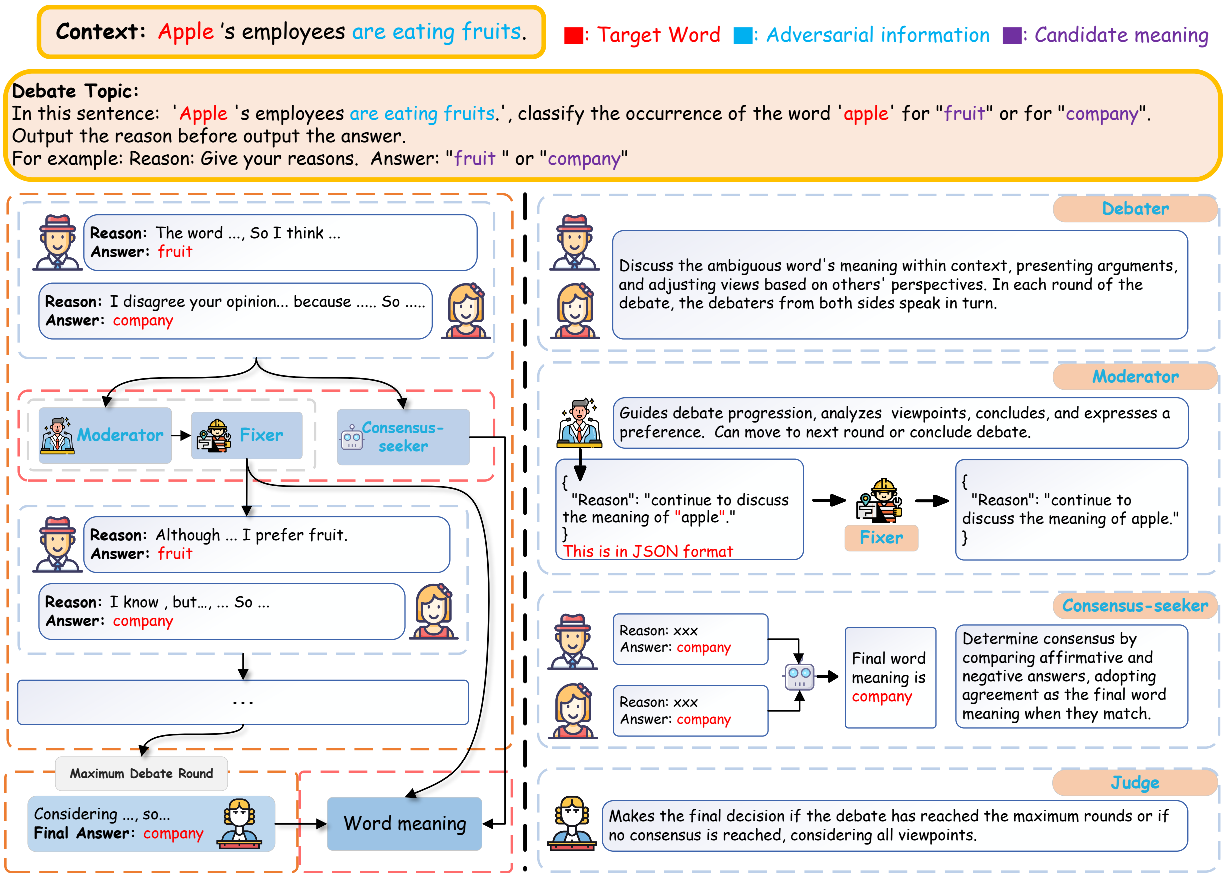

Figure 2: The MADAWSD framework consists of four key roles: (1) *Debater*: Debates the meaning of ambiguous words. (2) *Moderator*: Summarizes Debaters' arguments, expresses a preference, and guides or terminates the debate. (3) *Consensus-seeker*: Determines whether the debaters have reached an agreement and ends the debate if so. (4) *Judge*: Delivers the final decision if no consensus is reached within the set rounds.

considers the viewpoints of all Debaters and the preferences of the Moderator to determine the final meaning selection. To illustrate the workflow of the MADAWSD framework more clearly, a detailed example and the algorithm are given in Appendix G and Algorithm 1.

3.2.1 Debater

In the MADAWSD framework, Debaters are responsible for debating the meaning of the target ambiguous word in the given context. Debaters present arguments to support their viewpoints and provide answers in each round of the debate.

To simulate a real-world debate environment, we adopt a dialectical architecture to construct the MADAWSD framework. We rigorously divide debate participants into an affirmative side and a negative side, forming a dualistic debate structure denoted as $\mathcal{D} = \{d_{aff}, d_{neg}\}$, where d_{aff} represents the affirmative side and d_{neg} represents the negative side. This dualistic structure aligns with classical dialectical principles, providing a solid foundation for analyzing the debate process. To ensure fairness in the debate and mitigate the impact of power asymmetries, we adopt a strict alternating turn-taking mechanism. This mechanism guarantees equal debating opportunities for both the affirmative and negative sides. After both sides present their respective arguments, the content of a single round of the debate can be represented as $\mathcal{OD} = \{od_{aff}, od_{neg}\}$, where od_{aff} and od_{neg} denote the arguments from the affirmative and negative sides, respectively, in that round of the debate.

$$\begin{aligned} od_{aff} &= \mathrm{LLM}(H, \mathrm{prompt}_{\mathrm{d_{aff}}}), \\ od_{neg} &= \mathrm{LLM}(H, \mathrm{prompt}_{\mathrm{d_{neg}}}), \end{aligned} \tag{1}$$

where H refers to the debate history, $\mathrm{prompt}_{\mathrm{d_{aff}}}$ and $\mathrm{prompt}_{\mathrm{d_{neg}}}$ are prompts for Debaters, detailed in Appendix B.3.

During the debate process, we maintain a debate history record H to track the progress of the debate. This debate history H includes all information provided by the Debaters and the Moderator in each round, including arguments, rebuttals, and the Moderator's rulings. The debate history H is updated in real time as the debate progresses, serving as a crucial basis for subsequent analysis and decision-making.

Algorithm 1: MADAWSD.

Input : Initial prompts $\{\text{prompt}_{\text{d}_{\text{aff}}}, \text{prompt}_{\text{d}_{\text{neg}}}, \text{prompt}_{\text{mod}}, \text{prompt}_{\text{fix}}, \text{prompt}_{\text{jug}}\}$, context C, maximum debate round t.
Output : Word meaning

```
// Initialize variables
flag ← False, debate_topic ← C
round ← 0
while flag is False and round < t do
    round ← round + 1 ;
    od_aff ← LLM(H, prompt_d_aff)
    od_neg ← LLM(H, prompt_d_neg)
    OD ← {od_aff, od_neg}
    OM ← LLM(H, OD, prompt_mod)
    if OM is incorrect form then
        OM ← LLM(OM, prompt_fix)
    end
    if OM's preference is True then
        flag ← True
        Word meaning ← OM's Answer
    end
    OC ← ans_aff is equal to ans_neg
    if OC is True then
        flag ← True
        Word meaning ← ans_aff
    end
end
if flag is False then
    OJ ← LLM(H, prompt_jud)
    Word meaning ← OJ's Answer
end
```

Result: Word meaning

3.2.2 Moderator

In our framework, the Moderator is responsible for guiding the progression of the debate by analyzing the arguments and answers from all Debaters and expressing a preference for the correct meaning. The Moderator guides the debate to proceed to the next round, or directly concludes the debate by determining the most suitable meaning.

The Moderator analyzes the potential meanings of ambiguous words based on the Debaters' arguments and expresses a preference for the meaning of the ambiguous words. The Moderator's intervention is primarily aimed at assessing the progress of the debate and determining whether it is possible to derive the most suitable meaning of the ambiguous word and conclude the debate. Specifically, after both Debaters present their respective arguments $\mathcal{OD}$ in each round, the Moderator evaluates them and expresses their own opinion $\mathcal{OM}$.

$$\mathcal{OM} = \text{LLM}(H, \mathcal{OD}, \text{prompt}_{\text{mod}}), \quad (2)$$

where $\text{prompt}_{\text{mod}}$ is the prompt for Moderator, detailed in Appendix B.3.

The Moderator decides whether to terminate the debate based on their opinion. If the Moderator expresses a clear preference, indicating that a relatively reliable answer has been found, the debate will conclude. Otherwise, the debate will proceed to the next round.

Considering the potential hallucination issues that can arise when LLMs generate text, and to ensure the stable operation of the MADAWSD framework, we introduce the role of a Fixer as an affiliate of the Moderator. The primary function of the Fixer is to process the output of the Moderator and convert it into a standardized format. If the Moderator's output deviates from the predefined standard JSON format, the Fixer intervenes.

$$\mathcal{OM} = \begin{cases} \mathcal{OM}, & \text{correct form} \\ \text{LLM}(\mathcal{OM}, \text{prompt}_{\text{fix}}), & \text{incorrect form} \end{cases} \quad (3)$$

where $\text{prompt}_{\text{fix}}$ is the prompt for Moderator, detailed in Appendix B.3.

3.2.3 Consensus-seeker

In our framework, the Consensus-seeker is responsible for determining whether the debating parties have reached an consensus on the meaning of the ambiguous word. Once consensus is achieved, the Consensus-seeker adopts the consensus as the final answer and ends the debate.

The consensus status is denoted as $\mathcal{OC}$, and its formula can be expressed as:

$$\mathcal{OC} = \begin{cases} \text{True}, & ans_{aff} = ans_{neg} \\ \text{False}. & ans_{aff} \neq ans_{neg} \end{cases} \quad (4)$$

It is important to note that the Consensus-seeker is not implemented with an LLM-based agent, but rather as a mechanism that we implement through logical judgment.

3.2.4 Judge

In the MADAWSD framework, to ensure the debate process does not continue indefinitely, we implement a maximum round limit for debates.

If the debate reaches the prescribed maximum number of rounds without the Consensus-seeker confirming an agreement between the both sides, and the Moderator has not expressed a definitive preference, the Judge intervenes. The Judge's responsibility is to evaluate all existing candidate answers and select the final interpretation that most accurately reflects the target word's meaning in context, thereby terminating the debate. We denote the Judge's output as $\mathcal{OJ}$, representing the conclusive determination of the word's meaning.

$$\mathcal{OJ} = \text{LLM}(H, \text{prompt}_{\text{jud}}), \tag{5}$$

where $\text{prompt}_{\text{jud}}$ is the prompt for Judge, detailed in Appendix B.3.

4 Experiments and Analysis

4.1 Datasets

To comprehensively evaluate the MADAWSD framework, we conduct extensive experiments on two challenging datasets: the FOOL dataset and our self-constructed CAWSD dataset.

4.1.1 FOOL Dataset

The FOOL dataset is a newly emerged English WSD dataset (Ballout et al., 2024), specifically designed to investigate the robustness of LLMs in both regular and adversarial WSD scenarios. The dataset is divided into four subsets: the first two are designed for regular WSD, while the latter two focus on adversarial WSD. Currently, it is the only dataset specifically tailored for WSD tasks involving adversarial contexts. However, it is monolingual and supports English only.

4.1.2 Self-constructed CAWSD Dataset

To alleviate the problem of data scarcity for adversarial WSD and enhance linguistic diversity, we have constructed a novel Chinese adversarial WSD dataset. Similar to the organization of existing CoarseWSD-20 (Loureiro et al., 2021) and FOOL (Ballout et al., 2024) datasets, we carefully selected 20 ambiguous words, each word with two different meanings and 25 sentences for each meaning. We chose ambiguous words such as "眼线", which can refer to both "间谍" *(spy)* and "化妆术语" *(eyeliner)*; and "炒鱿鱼", which can mean both "食物" *(food)* and "开除" *(dismiss)*. These deliberately selected ambiguous words make the CAWSD dataset more challenging and effectively evaluate the model's WSD capabilities in adversarial contexts.

The construction of the CAWSD dataset was entirely carried out manually by Chinese native-speaking master's and doctoral students majoring in computer science, with a total of 6 annotators participating in the annotation process. For the CAWSD dataset, it is necessary to initially select ambiguous words manually from everyday environments and gather or create sentences containing these words. Subsequently, each sentence must be transformed into sentences that contain adversarial information and could potentially occur in the real world. Everyone is required to spend a minimum of 5 minutes on each sentence. For each sentence, we used a voting mechanism to verify the reasonableness of a sentence. If the approval rate is below 0.8, we will discard that sentence, ensuring the data quality and reliability. The detailed ambiguous words, meanings, and examples are given in Appendix A.

4.1.3 Statistics of Experimental Datasets

In our experiments, we adopt the FOOL dataset and the CAWSD dataset to evaluate the performance of our MADAWSD framework. The statistics of experimental datasets are shown in Table 1. Specifically, the first two sets (Set1 and Set2) of the FOOL dataset are used to validate effectiveness in regular WSD scenarios, while the fourth set (Set4) and the CAWSD datasets are utilized to validate performance in adversarial WSD scenarios. It is important to note that the third set (Set3) of the FOOL dataset is not included. The reason is that it is constructed with artificial adversarial contexts instead of realistic ones. For example, the sentence "*A young man jumped in and swam to the innovative bank*" is problematic. Clearly, while *bank* refers to *a riverbank*, its modification with *innovative* is deemed inappropriate. Therefore, Set3 of the FOOL dataset is excluded in our experiments.

Dataset	Word	Sentence	Adversarial
Set1 (FOOL)	20	1810	No
Set2 (FOOL)	20	1810	No
Set4 (FOOL)	20	1019	Yes
CAWSD	20	1000	Yes

Table 1: Statistics of experimental datasets.

4.2 Experiment Settings

4.2.1 Evaluation Metrics and Baselines

For each instance in the experimental datasets, we employ the MADAWSD framework to determine which of the two candidate meanings of the am-

Method	FOOL									CAWSD		
	Set1 (Regular)			Set2 (Regular)			Set4 (Adversarial)			CAWSD (Adversarial)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
GPT-3.5-Turbo												
Zero-shot	98.4	93.4	95.0	98.6	95.1	96.3	71.2	83.0	73.8	61.8	69.4	61.0
Zero-shot CoT	97.5	96.7	96.8	98.1	96.0	96.4	**76.8**	71.0	72.5	67.0	69.8	65.6
Zero-shot CoT+SC	**99.2**	97.8	**98.4**	**99.7**	96.9	97.7	71.7	73.2	68.1	**71.7**	73.2	**68.1**
MADAWSD	97.3	**97.8**	97.4	98.0	**98.7**	**98.2**	72.4	**86.3**	**78.0**	63.2	**75.2**	67.0
Qwen-Turbo												
Zero-shot	98.8	98.7	98.6	**99.5**	97.9	98.5	75.7	81.3	77.1	65.0	80.4	69.0
Zero-shot CoT	**98.8**	99.2	**99.0**	99.0	**99.4**	**99.2**	62.9	78.3	69.0	68.6	79.6	72.1
Zero-shot CoT+SC	98.7	96.3	96.4	99.0	95.6	96.0	66.2	78.1	70.1	**69.0**	83.6	**73.9**
MADAWSD	98.0	**99.3**	98.5	98.5	98.9	98.6	**77.2**	**87.0**	**81.1**	67.5	**84.4**	72.8
Qwen-Plus												
Zero-shot	99.5	99.9	99.7	99.9	99.8	**99.9**	86.4	88.5	86.6	76.7	85.8	77.7
Zero-shot CoT	99.6	99.8	99.7	**99.9**	99.6	99.8	85.4	86.6	85.7	80.5	79.2	76.1
Zero-shot CoT+SC	99.7	97.7	98.4	99.9	96.9	97.8	85.3	89.1	86.5	77.5	**89.0**	80.7
MADAWSD	**99.7**	**99.9**	**99.8**	99.8	**99.8**	99.8	**86.9**	**94.6**	**90.3**	**82.3**	84.6	**80.9**
Qwen-2.5-7B-Instruct												
Zero-shot	99.0	96.4	97.4	99.1	97.4	98.0	75.4	74.6	73.4	**74.6**	60.6	60.5
Zero-shot CoT	99.4	96.0	97.2	98.9	95.6	96.5	70.7	63.0	65.4	66.0	78.8	69.7
Zero-shot CoT+SC	**99.4**	95.4	96.4	**99.6**	96.1	97.1	75.4	76.9	74.6	69.3	**79.8**	71.4
MADAWSD	98.9	**97.9**	**98.3**	98.9	**98.2**	**98.5**	**78.7**	**78.8**	**77.6**	71.9	79.2	**71.8**
DeepSeek-V3												
Zero-shot	99.2	**99.8**	99.5	99.7	99.7	99.7	82.6	86.8	84.0	71.7	90.6	78.8
Zero-shot CoT	99.3	93.3	96.1	99.3	96.6	97.2	76.0	81.1	77.8	74.2	93.0	81.8
Zero-shot CoT+SC	99.4	96.6	97.1	99.4	96.3	96.8	77.2	85.2	79.9	73.6	**94.6**	**82.1**
MADAWSD	**99.5**	99.4	**99.5**	**99.7**	**99.8**	**99.7**	**85.6**	**90.2**	**87.4**	**76.0**	91.4	81.7

Table 2: Comparison of disambiguation performances with baselines on two datasets.

biguous word corresponds to the given context. To evaluate WSD performance, we use three evaluation metrics: Precision, Recall, and F1-score.

To investigate the effectiveness of the MADAWSD framework, we select five LLMs as baseline models, including GPT-3.5-Turbo-0125[2], Qwen-Turbo-0211[3], Qwen-Plus-0125[4], Qwen2.5-7B-Instruct[5], and DeepSeek-V3-241226[6] (Liu et al., 2024).

All experiments were conducted in a zero-shot setting. The models directly predict the correct meaning of ambiguous words according to the given context. This setup allows for a more effective evaluation of the models' generalization capabilities when confronted with new vocabulary and contexts, as well as their ability to understand contextual information. Models accessible through public APIs have been utilized, with the following baselines established:

- **Zero-shot:** We employed a zero-shot (Kojima et al., 2022) approach by directly inputting the context requiring disambiguation and the candidate meanings into the LLMs, enabling them to deduce the meaning of the ambiguous word without intermediate reasoning.

- **CoT:** We facilitate reasoning in LLMs by directly employing prompts of the form "Let's think step by step" to prompt the models. This approach encourages the LLMs to express their thought processes via the Chain of Thought (CoT) (Kojima et al., 2022) before arriving at the meaning of an ambiguous word.

- **Self-Consistency:** We adopt the Self-Consistency (Wang et al., 2023) approach, which builds on the zero-shot CoT method to sample various reasoning paths. The final answer is determined by selecting the most consistent outcome across all sampled paths.

4.2.2 Implementation Details

To ensure the stability of the outputs from LLMs, we configured the models by setting the temperature parameter to 0 throughout our experimental

[2] https://www.chatgpt.com

[3] https://chat.qwen.ai

[4] https://chat.qwen.ai

[5] https://modelscope.cn/models/Qwen/Qwen2.5-7B-Instruct

[6] https://chat.deepseek.com

procedures[7]. This adjustment minimizes the randomness in the models' predictions, thereby enhancing the stability of the experimental results. Furthermore, to maintain a structured and controlled debate environment within our experiments, we established a limit on the number of debate rounds. Specifically, we have set the maximum number of debate rounds to 3[8]. All experiments were repeated three times and averaged. The detailed prompt used in the experimental procedures is given in Appendix B.

4.3 Results

The detailed experimental results are shown in Table 2. According to the table, we have the following key observations.

First, by comparing model performance on regular WSD datasets (Set1 and Set2) with that on adversarial WSD datasets (Set4 and CAWSD), we find that all LLM baselines exhibit significantly degraded performance in adversarial settings. This confirms the high susceptibility of existing LLM-based methods to adversarial perturbations, frequently leading to disambiguation failures. It further highlights the challenge and research potential of adversarial WSD.

Second, our MADAWSD framework outperforms most baselines on both adversarial WSD datasets while maintaining stable performance on regular WSD datasets. In contrast, baseline methods (e.g., zero-shot and zero-shot CoT) demonstrate substantially inferior performance. We attribute this superiority to MADAWSD's unique design: it enhances robustness against adversarial information by simulating real-world debate dynamics, thereby fully leveraging LLMs' contextual reasoning capabilities.

Third, introducing CoT prompting into adversarial WSD tasks unexpectedly degrades the performance of most LLMs. This phenomenon is absent in regular WSD scenarios. CoT prompting seems to propagate and enhance initial errors. We analyze the underlying causes in Appendix C.

Notably, our MADAWSD framework integrates both efficiency and generality, which not only effectively handles adversarial perturbations in WSD tasks, but also seamlessly integrates with existing LLMs and is applicable to different languages.

[7]Specifically, for the SC method, we set the number of iterations to 5 and the temperature to 0.7.

[8]A more detailed analysis of the maximum number of debate rounds is provided in Appendix E.

Model	Variant	1-Round	2-Round	3-Round
GPT-3.5-Turbo	*w/o* Cs	23.6%	0.5%	76.0%
	w/ Cs	92.4 %(+68.8)	1.3% (+0.8)	6.3% (-69.7)
Qwen-Turbo	*w/o* Cs	89.2%	9.1%	1.7%
	w/ Cs	91.6% (+2.4)	6.1% (-3.0)	2.4% (+0.7)

Table 3: Ablation study of MADAWSD from the perspective of the number of debate rounds required to reach a final decision.

4.4 Ablation Study

We conduct an ablation study to analyze the rationality and effectiveness of the components designed in MADAWSD.

For the Debater and Moderator, a complete debate in MADAWSD requires the participation of two Debaters and one Moderator. To maintain the integrity of the debate process, both the Debaters and the Moderator must be retained.

For the Judge, we initially attempted to remove Judge from the framework. However, without Judge, when the maximum number of debate rounds is reached, both Debaters persist in their arguments, and the Moderator is unable to make a decision, leading to an endless debate. We therefore retained the Judge and manually reviewed the debate history. We found that some cases exhibited the aforementioned situation; however, in these cases, the Judge made the final judgment, successfully concluding the debate. Specifically, we conducted experiments using GPT-3.5-Turbo and Qwen-Turbo on Set4, where the examples involving the Judge accounted for approximately 5% of the cases in GPT-3.5-Turbo and 1% in Qwen-Turbo, respectively. Hence, the Judge is necessary.

For the Consensus-seeker, we conducted experiments on Set4 using GPT-3.5-Turbo and Qwen Turbo, and the results are presented in Table 3. The debate round completion statistics demonstrate that when the Consensus-seeker is incorporated, both models show significantly higher percentages of debates concluding in the first round compared to configurations without it. This indicates that the Consensus-seeker can substantially reduce the required debate duration, thereby making our approach more efficient.

Furthermore, as shown in Table 4, the model performance results reveal consistent improvements across all metrics when the Consensus-seeker is incorporated. This suggests that excessive debate rounds may potentially degrade performance, further validating the necessity of incorporating the Consensus-seeker role.

Model	Variant	Precision	Recall	F1-score
GPT-3.5-Turbo	*w/o* Cs	71.8	83.9	76.6
	w/ Cs	72.4 (+0.6)	86.3 (+2.4)	78.0 (+1.4)
Qwen-Turbo	*w/o* Cs	73.2	89.0	79.8
	w/ Cs	77.2 (+4.0)	87.0 (-2.0)	81.1 (+1.3)

Table 4: Ablation study of MADAWSD from the perspective of the performance.

4.5 Error Analysis

We conducted a human evaluation to identify the limitations of the MADAWSD framework. For the error cases, we traced the corresponding debate processes, aiming to determine the prevalent defects and issues within the framework, thereby providing a foundation for future improvements.

We categorize these errors into three groups: (1) *Confusion due to semantic similarity*. In some instances, two meanings of a target word are highly semantically similar, making it difficult even for humans to distinguish between them. For example, in the sentence "When he received the envelope, he carefully read every letter on it", the two meanings of "letter" (i.e., "a letter as a written communication" and "a letter of the alphabet") are both highly applicable in context, making it challenging for the framework to make an accurate judgment. (2) *Judgment bias by the Moderator/Judge*. Despite sufficient arguments provided by both sides of the debate, the Moderator or Judge sometimes still makes erroneous judgments, designating the incorrect meaning as the correct answer. This may stem from the Moderator/Judge's misunderstanding of contextual information or an unreasonable weighting of the arguments presented by both sides. (3) *Limited knowledge base of LLMs*. The knowledge base of LLMs is inherently limited and derived from their training datasets. Consequently, certain words that are predominantly employed in one sense and less frequently in another may be readily interpreted as the more common meaning during the debate process.

5 Conclusion

In this paper, we focus on two significant gaps limiting WSD advancement: LLMs are prone to misclassify ambiguous words when dealing with WSD instances involving adversarial information, and the lack of sufficient adversarial WSD datasets severely restricts the development and evaluation of WSD systems. To address these gaps, we propose an LLM-based MADAWSD framework and introduce a new Chinese adversarial WSD dataset named CAWSD. Owing to its unique design, MADAWSD not only performs well on adversarial WSD datasets but also maintains stable performance on regular WSD datasets. Extensive experiments on two adversarial WSD datasets demonstrate that MADAWSD can seamlessly integrate with existing LLMs, significantly outperforming most baseline methods, thus demonstrating its robust performance and broad generality. Future work will focus on exploring more efficient multi-agent frameworks to further improve efficiency and reduce inference costs.

Limitations

Although our work is the first to apply a multi-agent debate framework for adversarial WSD and has achieved better performance than existing LLMs, it has the following limitations. First, our framework still struggles to distinguish between highly semantically similar meanings, and its ability to discern subtle differences between semantics is still limited. Second, we have yet to implement our MADAWSD framework on a large number of LLMs due to the cost constraints of inference APIs. Finally, our framework is currently capable of handling only words with two candidate senses. Therefore, improving the model's ability to distinguish subtle differences between semantically similar meanings, as well as exploring more multi-agent frameworks to enhance effectiveness and reduce inference costs, are promising directions for future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62376130, No.62502249), Shandong Provincial Natural Science Foundation (No.ZR2022MF243, No.ZR2024QF054), Program of New Twenty Policies for Universities of Jinan (No.202333008), Program of Innovation Improvement for Small and Medium-sized Enterprises of Shandong (No.2024TSGC0039), the Open Project of the Key Laboratory of Computing Power Network and Information Security, Ministry of Education (No.2023ZD027), and the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2024ZDZX08).

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Mohamad Ballout, Anne Dedert, Nohayr Abdelmoneim, Ulf Krumnack, Gunther Heidemann, and Kai-Uwe Kühnberger. 2024. FOOL ME IF YOU CAN! an adversarial dataset to investigate the robustness of LMs in word sense disambiguation. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing*, pages 5042–5059.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Rexhina Blloshmi, Tommaso Pasini, Niccolò Campolungo, Somnath Banerjee, Roberto Navigli, and Gabriella Pasi. 2021. IR like a SIR: sense-enhanced information retrieval for multiple languages. In *Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing*, pages 1030–1041.

Pankaj Dadure, Partha Pakray, and Sivaji Bandyopadhyay. 2024. Mathematical information retrieval: A review. *ACM Computing Surveys*, 57(3):1–34.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, et al. 2023. MetaGPT: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3509–3514.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 22199–22213.

Anastasia Kritharoula, Maria Lymperaiou, and Giorgos Stamou. 2023. Large language models and multimodal retrieval for visual word sense disambiguation. In *Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing*, pages 13053–13077.

Sunjae Kwon, Rishabh Garodia, Minhwa Lee, Zhichao Yang, and Hong Yu. 2023. Vision meets definitions: Unsupervised visual word sense disambiguation incorporating gloss information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1583–1598.

Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing*, pages 180–192.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904.

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

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Wenpeng Lu, Fanqing Meng, Shoujin Wang, Guoqiang Zhang, Xu Zhang, Antai Ouyang, and Xiaodong Zhang. 2019. Graph-based chinese word sense disambiguation with multi-knowledge integration. *Computers, Materials and Continua*, 61(1):197–212.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 23rd Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. Dibimt: A gold evaluation benchmark for studying lexical ambiguity in machine translation. *Computational Linguistics*, pages 1–71.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 599–621.

Van-Hien Tran, Raj Dabre, Hour Kaing, Haiyue Song, Hideki Tanaka, and Masao Utiyama. 2025. Exploiting word sense disambiguation in large language models for machine translation. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages*, pages 135–144.

Hansi Wang, Yue Wang, Qiliang Liang, and Yang Liu. 2025. LTRS: Improving word sense disambiguation via learning to rank senses. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1934–1942.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The 11th International Conference on Learning Representations*.

Yue Wang, Qiliang Liang, Yaqi Yin, Hansi Wang, and Yang Liu. 2024a. Disambiguate words like composing them: A morphology-informed approach to enhance chinese word sense disambiguation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15354–15365.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024b. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 257–279.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Qihao Yang, Yong Li, Xuelin Wang, Fu Lee Wang, and Tianyong Hao. 2024. PolCLIP: A unified image-text word sense disambiguation model via generating multimodal complementary representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10676–10690.

Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4061–4070.

Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. 2023. Neuro-symbolic sentiment analysis with dynamic word sense disambiguation. In *Proceedings of the 61st Association for Computational Linguistics*, pages 8772–8783.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021. Leveraging word-formation knowledge for Chinese word sense disambiguation. In *Proceedings of the 26th Conference on Empirical Methods in Natural Language Processing*, pages 918–923.

A Words, Meanings, and Example Sentences in the CAWSD dataset

To more intuitively illustrate how our self-constructed Chinese Adversarial Word Sense Disambiguation dataset (CAWSD) contributes to evaluating more realistic adversarial WSD, a portion of the words, meanings, and example sentences is shown in Table 5. Each word is represented by two rows: the first row displays an example sentence for the first meaning, and the second row showcases an example sentence for the second meaning. Adversarial information is highlighted in blue.

Words	Meanings	Example Sentences
笔记本	电脑 (laptop)	她把日记本摊在笔记本旁边，对比手写体与电子字体的微妙差异。 *She spread out the diary next to the laptop, comparing the subtle differences between the handwritten script and the electronic font.*
	纸质本子 (paper notebook)	咖啡馆里，女孩对着发光的电脑屏幕，钢笔在笔记本上快速游走。 *In the café, the girl faced the glowing laptop screen, with her fountain pen swiftly gliding across the paper notebook.*
炒鱿鱼	开除 (dismiss)	公司年度晚会的鱿鱼烧烤摊前，两个偷吃食材的兼职学生收到了炒鱿鱼警告。 *At the squid barbecue stall during the company's annual party, two part-time students who were caught stealing ingredients received a warning of dismissal.*
	食物 (food)	昨天他做了一份特别美味的炒鱿鱼，今天得知邻居收到了被开除的通知。 *Yesterday he made a particularly delicious stir-fried squid and today he learned that his neighbor received a notice of dismissal.*
出轨	背叛 (betray)	他们曾经的幸福就像一列快车，直到他的出轨让这列列车彻底翻覆。 *Their past happiness was like a fast train, until his betrayal derailed it.*
	脱轨 (derail)	出轨现场的列车残骸，混着他忘在客房的结婚戒指在新闻镜头里一闪而过。 *The wreckage of the train at the scene of the derailment, mixed with his forgotten wedding ring in the guest room, flashed briefly across the news.*
火箭	篮球队 (team)	看着飞行器飞向太空，球迷们也像看火箭队飞速进攻一样，心情激动。 *Watching the rocket fly into space, fans are also excited, like watching the Houston Rockets launch a rapid attack.*
	飞行器 (rocket)	篮球队员们在火箭发射的过程中，体会到了一种与比赛相似的紧张感和期待。 *During the rocket launch, the basketball players experienced a sense of tension and anticipation similar to that of a game.*
满月	月亮 (moon)	满月移过百日宴的八仙桌，银匙敲击瓷碗的脆响惊醒了熟睡的婴孩。 *The full moon moved across the Eight Immortals table of the hundred-day feast, and the crisp sound of a silver spoon striking a porcelain bowl startled the peacefully sleeping infant.*
	满一个月 (time)	庭院石臼盛着接来的露水，满月礼的银铃铛在廊下轻响，与荷叶上滚动的月华共振。 *The stone mortar in the courtyard held the dew that had been collected, and the silver bells of the full-month celebration softly chimed beneath the corridor, resonating with the moonlight rolling off the lotus leaves.*
台风	舞台风格 (stage presence)	大自然的狂风横扫大地，但她的台风却在舞台上独树一帜，无法忽视。 *The fierce winds of nature swept across the land, but her stage presence was unique on the stage, impossible to ignore.*
	天气 (typhoon)	舞台监督的哨声穿透风雨："台风是今晚最抢戏的临时演员！" *The stage manager's whistle cut through the wind and rain: "The typhoon is tonight's most attention-grabbing guest star!"*
铁饭碗	工作 (job)	洗碗机轰鸣声中，妻子突然说："要不你也去考个事业编，端个铁饭碗？" *Amid the hum of the dishwasher, his wife suddenly said, "How about you take the civil service exam and get yourself a stable job?"*
	餐具 (bowl)	这只铁饭碗总让他联想到自己那份稳定的工作，每天安定的生活带给他莫大的安心。 *This iron rice bowl always reminded him of his stable job, and the steady life it provided brought him immense peace of mind.*
小米	食物 (food)	她用红米云存储她参观小米种植基地的照片，以确保她不会丢失任何记忆。 *She uses Redmi Cloud to store her photos from visiting the Xiaomi planting base, ensuring that she won't lose any memories.*
	公司 (company)	围绕小米新产品发布的兴奋之情，就像冬日里第一口美味粥品一样令人温暖和振奋。 *The excitement surrounding Xiaomi's new product launch is as warming and uplifting as the first spoonful of a delicious porridge on a winter day.*
眼线	间谍 (spy)	精心设计的妆容掩饰了她内心的紧张，因为她的眼线给她传递的情报格外重要。 *Her meticulously designed makeup concealed the nervousness inside, as the information relayed by her spy was especially important to her.*
	化妆术语 (eyeliner)	他观察她的眼线，却看到她的眼睛中却透露出一种极其警觉的情绪，似乎在寻找某个线索。 *He observed her eyeliner, but saw a look of heightened alertness in her eyes, as if she were searching for a clue.*

Table 5: Words, meanings, and example sentences in the CAWSD dataset. Adversarial information is marked in blue.

B Prompts used in Experiments

B.1 Prompt of a single LLM (Zero-shot)

Following the work of the FOOL dataset (Ballout et al., 2024), we adopt the following prompt to utilize LLMs in a zero-shot setting, as shown in Table 6.

Zero-shot: In this sentence: <*sentence*>, classify the occurrence of the word <*target word*> for <$meaning_1$> or for <$meaning_2$>. Answer only by one of these options: <$meaning_1$> or <$meaning_2$>. Don't output irrelevant content.

Table 6: Prompt given to an LLM in zero-shot setting.

B.2 Prompt of a single LLM (Zero-shot CoT)

To assess the performance of the LLM's most basic Chain-of-Thought (CoT) method (Kojima et al., 2022) on WSD tasks, we employ the prompt *"let's think step by step"* to guide the models.

Since the output generated by the CoT method does not yield a single meaning, we use another LLM to extract the word meaning from the CoT output for performance evaluation.

The prompt provided in Table 7 is used for this purpose.

CoT: Let's think step by step. In this sentence: <*sentence*>, classify the occurrence of the word <*target word*> for <$meaning_1$> or for <$meaning_2$>. The answer can only be chosen from <$meaning_1$> or for <$meaning_2$>. Word modification is not allowed.
Answer extraction: Here is a passage. You need to extract one word from it as the answer. The answer can only be chosen from <$meaning_1$> or <$meaning_2$>. Word modification is not allowed. Just output the answer and do not output any other content: <*The output of CoT*>.

Table 7: Prompt given to an LLM in zero-shot CoT setting.

B.3 Prompt of MADAWSD

Our MADAWSD framework comprises four key roles, namely the Debater, the Moderator, the Consensus-seeker, and the Judge, as well as an auxiliary role of the Moderator, the Fixer. Since the Consensus-seeker is realized through logical judgment, it does not involve a prompt. We present the prompts for the Debater, the Moderator, the Fixer, and the Judge. Prompt of MADAWSD are presented in Table 8.

Debate topic: In this sentence: <*sentence*>, classify the occurrence of the word <*target word*> for <$meaning_1$> or for <$meaning_2$>. Answer only by one of these options: <$meaning_1$> or <$meaning_2$>. Output the reason before output the answer. For example: Reason: Give your reasons. Answer: <$meaning_1$> or <$meaning_2$> Don’t output irrelevant content.
Meta prompt of Debater: You are a debater. Welcome to the word sense debate competition. You can agree or disagree with other’s viewpoint, and you can use context analysis, syntax analysis, dependency parsing or other methods, it is forbidden to copy the views of one’s opponents, as our objective is to find the correct word sense. The debate topic is stated as follows: <*Debate topic*>
Meta prompt of Moderator: You are a moderator. There will be two debaters involved in a debate. They will present their answers and discuss their perspectives on the following topic: <*Debate topic*> At the end of each round, you will evaluate answers and decide which is correct.
Affirmative: <*Debate topic*>
Negative: The other debater’s answer is: <*Affirmative’s answer*> Provide your reasons and answer.
Moderator: Now the <*round*> round of debate for both sides has ended. Affirmative side arguing:<*Affirmative’s answer*> Negative side arguing: <*Negative’s answer*> You, as the moderator, will evaluate both sides’ answers and determine if there is a clear preference for an answer candidate. If so, please summarize your reasons for supporting affirmative/negative side and give the final answer that you think is correct, and the debate will conclude. If not, the debate will continue to the next round. Don’t give your final answer unless you are absolutely sure. Keep the debate going as long as you can. Now please output your answer in this format, with the format as follows: { "Whether there is a preference": "Yes or No", "Supported Side": "Affirmative or Negative", "Reason": <*Moderator’s reasons*>, "Debate answer": <*Moderator’s answer*> }. Please strictly output in this format, do not output irrelevant content.
Fixer: You need to determine if the text you receive belongs to this format: { "Whether there is a preference": "Yes or No", "Supported Side": "Affirmative or Negative", "Reason": <*Moderator’s reasons*>, "Debate answer": <*Moderator’s answer*> }. If yes, print original content, if no, organize the text into the format and print it, the use of quotation marks must be taken care of to ensure the correct JSON format. For example "aaa": "bbb" is correct and "aaa": "b"b"bb" is wrong. Don’t output text like ``` json either. Don’t output other irrelevant content.
Judge: Affirmative side arguing: <*Affirmative’s answer*> Negative side arguing: <*Negative’s answer*> Now, what answer candidates do we have? Present them without reasons. Therefore, <*Debate topic*> Please summarize your reasons and give the final answer that you think is correct. Now please output your answer in this format, with the format as follows: { "Reason": <*Judge’s reasons*>, "Debate answer": <*Final answer*> }. You must give the only final answer and allow no neutrality, such as no preference. Please strictly output in this format, do not output irrelevant content.

Table 8: Prompt given to MADAWSD.

C Analysis on the CoT Method

We uncover a counterintuitive phenomenon: incorporating Chain-of-Thought (CoT) prompting in zero-shot settings paradoxically degrades performance compared to the standard zero-shot baseline, as shown in Table 12.

Our analysis reveals that in adversarial WSD contexts, CoT approaches may amplify model hallucinations. Specifically, adversarial perturbations in the input context propagate through the reasoning chain, causing cascading errors that ultimately misclassify ambiguous words.

These failures can be attributed to the susceptibility of individual LLMs to adversarial information. In adversarial WSD tasks, encouraging excessive CoT reasoning increases the likelihood of the model being misled. In contrast, our MADAWSD framework addresses this vulnerability by decoupling from the limitations of a single LLM, thereby significantly enhancing robustness and accuracy in adversarial WSD scenarios.

D Additional Error Analysis on CAWSD

Interestingly, on the CAWSD dataset, the MADAWSD method is not consistently the best. To investigate this, we conducted a more in-depth error analysis. Specifically, we examined the F1-score for each word sense and identified two types of words that contribute to errors:

For certain words, there is a significant frequency disparity among their candidate senses, and LLMs tend to favor the more common sense. This causes Debaters to incorrectly select the more frequent sense, resulting in the conclusion of the debate with an incorrect answer. Even if, in the first round of debate, one Debater selects the correct answer without being influenced by this preference, while the other selects the wrong answer due to bias, the two sides fail to reach a consensus. In subsequent debates, the Debater biased toward the incorrect answer is more easily persuaded by the other's arguments. In contrast, the CoT+SC method ensures that each reasoning chain remains independent, and does not alter the correct answer due to external influences.

As shown in Table 9, the word '下课' *(xiàkè)* has two senses: a high-frequency sense ('after class') and a low-frequency sense ('dismissal from position'). Our additional error analysis revealed that MADAWSD's F1-score for this word is significantly lower than that of the CoT+SC method.

Model	MADWSD	CoT + SC
GPT-3.5-Turbo	13.8	**21.4**
Qwen-Turbo	14.3	**32.2**
DeepSeek-V3	56.4	**74.5**

Table 9: Error analysis on the word '下课'.

When the two sides of the debate hold opposing views, the Moderator/Judge's judgment bias can lead to an incorrect decision. Since there is no Moderator/Judge in CoT+SC, it remains unaffected. The most notable characteristic of such words is that, in certain specific LLMs, the presence of the Moderator/Judge causes MADAWSD's performance to be significantly lower than that of CoT+SC, while in other LLMs, MADAWSD's F1-score is generally higher than that of CoT+SC.

Model	MADWSD	CoT + SC
GPT-3.5-Turbo	75.4	**89.4**
Qwen-Turbo	**68.0**	64.3
DeepSeek-V3	**94.3**	85.2

Table 10: Error analysis on the word '上火'.

Specifically, as illustrated in Table 10, in GPT-3.5-Turbo, the word '上火' is an anomaly,

Model	MADWSD	CoT + SC
GPT-3.5-Turbo	**78.9**	78.1
Qwen-Turbo	75.5	**90.6**
DeepSeek-V3	**85.3**	84.7

Table 11: Error analysis on the word '火星'.

Table 11 shows that, in Qwen-Turbo, '火星' is an anomaly, while no such anomalies are found in DeepSeek-V3.

This is consistent with the error analysis in Section 4.5. After removing the anomalous words for each model, we find that MADAWSD outperforms the CoT+SC method in all cases.

Question: Apple has snacks in the kitchen for all employees, classify the occurrence of the word 'apple' for fruit or for company.
Zero-shot Answer: company (✓)
Zero-shot CoT Answer: fruit (✗)
Zero-shot CoT Rationale:
To determine the correct classification for the word "Apple" in the given sentence, let's analyze it step by step:
1. Contextual clues: The sentence mentions "snacks in the kitchen for all employees."
This suggests that Apple is likely referring to something tangible or accessible in a physical space.
2. Logical reasoning:
   - If "Apple" refers to the company, it would be unusual to place a company in a kitchen.
   - If "Apple" refers to fruit, it makes more sense in the context of snacks and a kitchen setting.
3. Conclusion: Given the context of snacks and a kitchen, "Apple" most likely refers to the fruit.
Therefore, the classification for the word "Apple" in this sentence is fruit. (✗)

Table 12: Examples showing the influence of the CoT method.

E Analysis of the Maximum Number of Debate Rounds

To investigate the impact of the maximum number of debate rounds in the MADAWSD framework, we conducted a dedicated analysis to evaluate the rationale behind setting a limit and to explore the characteristics of debating behavior under different round configurations. The corresponding results are illustrated in Figure 3. We experimented with various maximum debate round settings on Set4, ranging from 1 to 4. Our findings indicate that most LLMs achieve optimal performance when the maximum number of rounds is set to 3. However, when the limit is increased to 4, the performance of all LLMs—except for Qwen-2.5-7B-Instruct—declines to varying degrees. Based on these observations, we set the maximum number of debate rounds to 3 in our subsequent experiments. This setting implies that if consensus on the most appropriate meaning of the target word is not reached within 3 rounds, the Judge will intervene, select the answer that best reflects the target word's meaning based on the debate history, and forcibly terminate the debate. This mechanism is designed to prevent debates from continuing indefinitely while ensuring the overall efficiency of the framework.

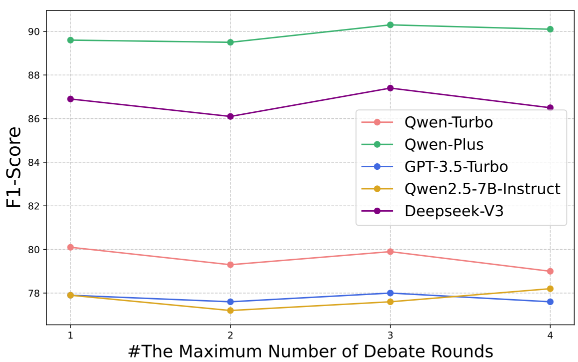

Figure 3: The impact of maximum number of debate rounds.

Furthermore, we conducted a detailed statistical analysis on the number of debate rounds required to reach a final decision, as shown in Table 13. According to the results, most debates either reach a consensus in the first round or are resolved directly by the Moderator. Only a small proportion of debates proceed to the maximum number of rounds (i.e., 3). This finding supports the reasonableness of setting the maximum number of debate rounds to 3, as it balances the need for sufficient interaction between Debaters with the overall efficiency of the framework.

Model	1-Round (%)	2-Round (%)	3-Round (%)
GPT-3.5-Turbo	92.44	1.28	6.28
Qwen-Turbo	91.56	6.08	2.36
Qwen-Plus	89.1	5.5	5.4
Qwen-2.5-7B-Instruct	98.23	1.38	0.39
DeepSeek-V3	77.04	14.52	8.44

Table 13: Statistical analysis on the number of debate rounds required to reach a final decision.

F Efficiency Analysis

To investigate the computational efficiency of the MADAWSD framework, we conducted a comprehensive analysis. Specifically, we used GPT-3.5-Turbo and Qwen-Turbo on Set4 to measure the average number of API calls and the average number of output characters required for our framework to complete one disambiguation process, and compared the results with those of the CoT+SC method.

Figure 4 illustrates the average number of API calls in the MADAWSD framework compared with the CoT+SC method. The experimental results indicate that our approach requires approximately 30% fewer API calls to complete one disambiguation process than the CoT+SC method.

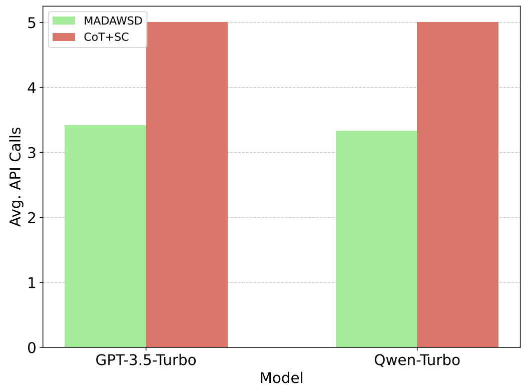

Figure 4: Statistical analysis of the average number of API calls required to complete one disambiguation process.

In addition, Figure 5 presents the average output character length of the MADAWSD framework and the CoT+SC method when completing one disambiguation process. The results demonstrate that our framework requires a shorter average output character length than the CoT+SC method for a single disambiguation process.

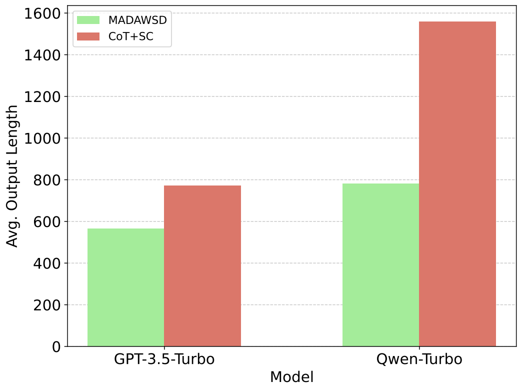

Figure 5: Statistical analysis of the average number of output length required to complete one disambiguation process.

Based on these experimental findings, we conclude that our framework achieves higher efficiency and superior performance at a lower computational cost.

G Case Study

Table 14 illustrates the debate process within our framework regarding the sentence *"As a biologist, he kept reading about mitosis and meiosis while lying in his cell."* We observed that the affirmative consistently argued for the wrong answer, while the negative provided the correct answer along with supporting reasons. After two rounds of debate, the moderator ultimately selected the negative's answer as the final word meaning.

Debate Topic: In this sentence: ‘As a biologist , he kept reading about mitosis and meiosis , while lying in his cell’. classify the occurrence of the word ‘cell’ for "biology" or for "prison".
Output the reason before output the answer. For example:
Reason: Give your reasons. Answer: "biology" or "prison"

Round 1:

Affirmative: Reason: The sentence mentions "mitosis and meiosis," which are biological processes related to cell division. Additionally, the subject is identified as a biologist, which further supports the biological context. The phrase "lying in his cell" is more likely to metaphorically refer to a biological cell rather than a prison cell, given the surrounding context.
Answer: biology

Negative: Reason: Although the sentence includes terms like "mitosis and meiosis" related to biology, "cell" is more likely referring to a prison cell given the context. The common usage of "cell" in the context of a prison setting makes it the preferred interpretation here. The biological terms don’t change the more typical understanding of "cell" as a prison cell in this sentence structure.
Answer: prison

Moderator: Because both sides have certain reasons, the debate will continue. (Format Error ✗)

Fixer:

```
{
  "Whether there is a preference": "No",
  "Supported Side": "",
  "Reason": "Because both sides have certain reasons, the debate will continue.",
  "Debate answer": ""
}
```

Consensus-Seeker: The two sides failed to reach a consensus, so the debate continued.

Round 2:

Affirmative: Reason: The biological interpretation is more fitting. The subject is a biologist, and "mitosis and meiosis" are clearly related to biological cells. "Lying in his cell" can be metaphorically understood as deep engagement with his research in a lab setting.
Answer: biology

Negative: Reason: The phrase "lying in his cell" is more likely to refer to a prison cell, as it describes a confined space where someone can physically lie down. The biological context of "mitosis and meiosis" doesn’t fit well with someone "lying in" a microscopic cell.
Answer: prison

Moderator:

```
{
  "Whether there is a preference": "Yes",
  "Supported Side": "negative",
  "Reason": "The directness and natural fit of the prison interpretation outweigh the biological context.",
  "Debate answer": "prison"
}
```

Final word meaning: prison

Table 14: A detailed case study of the MADAWSD framework.