OER: Opinion Externalization via Large Language Model Reasoning for Implicit Sentiment Analysis

Bingfeng Chen[1,2], Yongqi Luo[1], Shaobin Shi[1], Yuguang Yan[1], Ruichu Cai[1,3], Zhifeng Hao[1,4], and Boyan Xu[1,*]
[1]School of Computer Science, Guangdong University of Technology, Guangzhou, China
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]College of Mathematics and Computer, Shantou University, Shantou, China
Emails: chenbf@gdut.edu.cn, lyongqi001@gmail.com, d7inshi@gmail.com, ygyan@outlook.com, cairuichu@gmail.com, haozhifeng@stu.edu.cn, hpakyim@gmail.com

***Abstract*—Implicit Sentiment Analysis (ISA) requires models to infer sentiment polarity without explicitly given opinions. To accomplish this, models must either capture opinions from the surrounding context or leverage implicit clues for sentiment inference. Existing large language model (LLM)-based approaches decompose ISA through multi-step reasoning but suffer from hallucination and susceptibility to contextual interference. To address these challenges, we propose the Opinion Externalization via LLM Reasoning (OER) framework, which externalizes implicit opinions to enhance sentiment inference. OER consists of two key modules: (1) an *Explicit Opinion Extraction* module, which guides LLMs in extracting opinions while reducing hallucination through self-correcting in-context learning and a self-debate mechanism; and (2) an *Implicit Sentiment Clue Extraction* module, which retrieves sentiment-enhancing information to infer implicit opinions while mitigating contextual interference. Experimental results on two ISA benchmark datasets demonstrate that OER consistently outperforms existing methods in both overall performance and implicit sentiment accuracy.**

***Index Terms*—Implicit Sentiment Analysis**

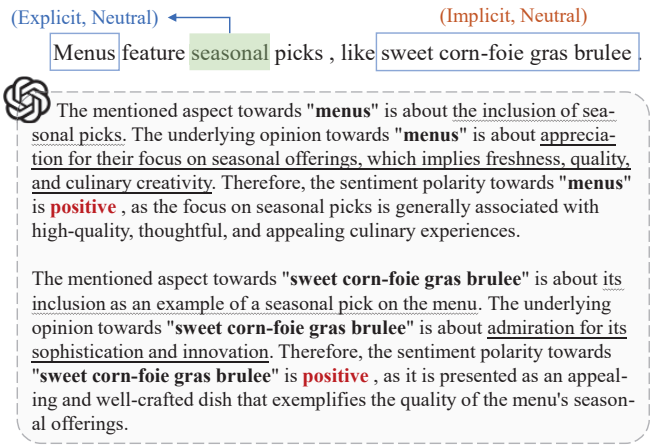

Fig. 1. This sentence encompasses two aspects, with one of the opinions being implicit. The opinion information here is sourced from [4].

I. Introduction

Implicit Sentiment Analysis (ISA) focuses on decoding aspect sentiment polarity from contexts lacking explicit opinion anchors. ISA poses two key challenges: contextual anchoring, where aspects need accurate links to explicit opinions for polarity determination. And implicit evaluative intent decoding, where models rely on deep semantic reasoning to infer polarity for phrases without direct evaluative terms.

Recent advances in large language models (LLMs) have brought new progress to ISA. THOR [1] proposes a CoT-based three-hop reasoning framework for implicit aspect, opinion and polarity inference, and subsequent methods like RVISA [2] and MT-ISA [3] further fine-tune LLMs on THOR's process data to achieve notable gains. Yet these methods still suffer from non-negligible limitations: the inherent hallucination of LLMs distorts polarity judgment (e.g., As illustrated in Figure 1, THOR misclassifies "menus" as positive by over-associating the neutral term "seasonal"), and implicit inference is susceptible to contextual explicit aspect interference, leading to cumulative reasoning errors. More importantly, these LLM-based methods are often computationally expensive and hard to deploy, making them impractical for real-world ISA scenarios that require lightweight models.

To address these limitations and enable efficient ISA via lightweight models, we propose an LLM-based Opinion Externalization Reasoning (OER) framework, which converts implicit or ambiguous opinion information into explicit sentiment cluesand uses these explicit clues to guide small models for ISA tasks. Specifically, OER addresses two core scenarios to generate high-quality explicit guidance clues: (1) accurate extraction of existing explicit opinions from contexts; (2) inference of implicit clues from external knowledge for opinion-free contexts. OER consists of two core modules to realize reliable opinion externalization: the Explicit Opinion Extraction (EOE) module guides LLMs to extract explicit opinions without additional fine-tuning and mitigates hallucination via self-correction in-context learning and a self-debate mechanism; the Implicit Sentiment Clue Extraction (ISCE) module infers implicit clues by retrieving an enhanced opinion corpus, mitigating contextual interference from explicit aspects. The explicit clues generated by OER are then used as supervised

*Corresponding author: hpakyim@gmail.com

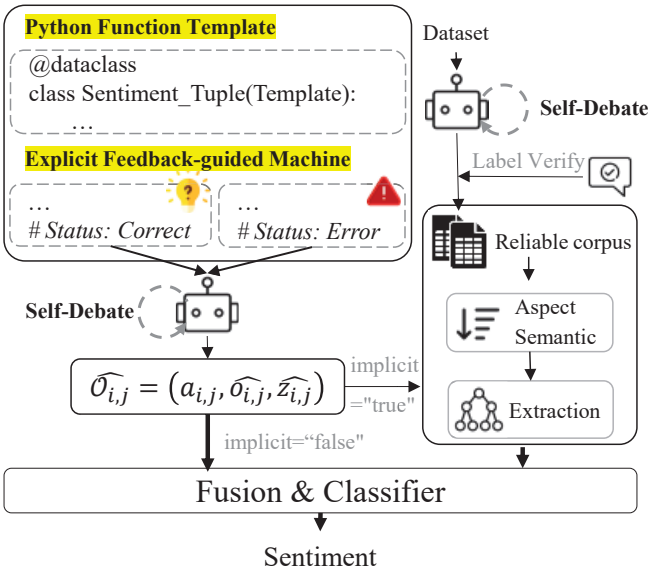

Fig. 2. OER Architecture

guidance to train or assist small models, enabling lightweight models to avoid the aforementioned limitations of LLMs and achieve accurate ISA inference. Experiments on two mainstream ISA benchmarks show that small models guided by OER outperform all state-of-the-art baselines, achieving superior overall performance and implicit sample accuracy while maintaining high efficiency.

II. Method

For the ISA task, the dataset is $\mathcal{D} = \{(x_i, y_{i,j})\}_{i=1}^{N}$, where x_i is the i^{th} sentence and $y_{i,j}$ the sentiment polarity of aspect $a_{i,j}$ in x_i. Standard models output $\hat{y}_{i,j} = f(x_i, a_{i,j})$, while our approach introduces aspect-associated intermediate sentiment elements $o_{i,j}$ to yield $\hat{y}_{i,j} = f(x_i, o_{i,j}, a_{i,j})$.

To ensure the quality of $o_{i,j}$, we design three modules to address key challenges: (1) reduce instability and improve reliability in $o_{i,j}$ generation; (2) extract hidden opinions closely related to target aspects; (3) integrate explicit and implicit opinion signals into a unified polarity prediction pipeline.

A. Explicit Opinion Extraction

To solve the first core challenge, we employ three strategies to ensure valid opinion generation: Python Function Template Construction, Explicit Feedback-guided Mechanism(EFGM), and Self-debate Mechanism, which guarantee generation stability and reliability from input to output.

1) Python Function Template Construction: Following the code-style prompting paradigm [5, 6, 7], we define the predicted opinion entity as $\hat{\mathcal{O}}_{i,j} = (a_{i,j}, \hat{o}_{i,j}, \hat{z}_{i,j})$, where $\hat{z}_{i,j} =$ True for implicit opinions and False for explicit ones. We map the task into executable Python function templates, where each code-style prompt $\mathcal{P}_k$ specifies output requirements via docstrings, and structured outputs are enforced by initialized empty data structures and formatting guidelines. This design aligns task semantics with model output, ensuring consistent and semantically grounded opinion extraction.

2) Examples Selection and Introduction: Carefully selected examples facilitate the model's task understanding, thus we propose the *Explicit Feedback-guided Mechanism*(EFGM) that integrates positive-negative examples with explicit feedback signals for model self-correction.

We construct the example set by sampling 10% of the training data, identifying model failure cases via few-shot prompting and polarity analysis, and manually annotating correct opinion triplet relations and common error patterns (e.g., polarity confusion, cross-sentence entity mismatches). Semantic similarity is used to form positive-negative instance pairs $O_{i,j}^{\pm} = (a_{i,j}, o_{i,j}, z_{i,j})$, and explicit feedback signals $\hat{F}_{i,j}^{\pm}$ are added: $\hat{F}_{i,j}^{+}$ is labeled "#Status: Correct" for valid predictions, and $\hat{F}_{i,j}^{-}$ is marked "#Status: Error. Expected: D" for erroneous ones to indicate discrepancies and correct outputs. These pairs and signals form dual-channel prompts $P_i^{\pm} = (O_{i,j}^{\pm}, \hat{F}_{i,j}^{\pm})^{N_{x_i^{\pm}}}$, and cascaded prompt concatenation guides opinion generation:

$$\hat{\mathcal{O}}_{i,j} = \text{LLM}(\mathcal{P}_i \parallel \mathcal{P}_i^{+} \parallel \mathcal{P}_i^{-}) \tag{1}$$

where $\parallel$ denotes sequential concatenation.

3) Self-debate Mechanism: We introduce the *Self-debate Mechanism* to address hallucination risks in generated opinions via textual entailment-based verification and iterative correction. Taking the initial predicted $\hat{\mathcal{O}}_{i,j}$ as input, the system reformulates it into a natural language assertion using a verification template $\mathcal{V}_{i,j}$ for entailment evaluation. An LLM judges the semantic consistency between the original context x_i and the reconstructed statement: if consistent, verification passes; if contradictory, a conflict marker $\mathcal{E}_{i,j}^{(t)}$ is injected to trigger corrective iteration. The process is formalized as:

$$(e_{i,j}^{(t)}, \mathcal{E}_{i,j}^{(t)}) = \text{LLM}\left(\mathcal{V}_{i,j}\left(\hat{\mathcal{O}}_{i,j}^{(t-1)}, x_i\right)\right) \tag{2}$$

$$\hat{\mathcal{O}}_{i,j}^{(t)} = \text{LLM}\left(\mathcal{P}_i \parallel \mathcal{P}_i^{+} \parallel \mathcal{P}_i^{-} \parallel \mathcal{E}_{i,j}^{(t)}\right) \tag{3}$$

where $t \in \{1, \ldots, T\}$. The iteration stops when $e_{i,j}^{(t)} = 1$, and the verified $\hat{\mathcal{O}}_{i,j}^{(T)}$ is output. For explicit opinions ($z_{i,j} =$ False), $O_{i,j}^{(T)}$ is directly used for subsequent prediction; for implicit ones ($z_{i,j} =$ True), the Implicit Sentiment Clue Extraction (ISCE) module is further employed.

B. Implicit Sentiment Clue Extraction

To solve the second core challenge of extracting target-aspect-related hidden opinions, we use sentiment-domain corpora to mitigate inference difficulties caused by implicit emotional expressions.

1) Corpus Construction: We construct the opinion corpus $\mathcal{D}^T$ from the training splits of public datasets, adopting the same Python function template and self-debate verification framework as the EOE module. Specifically, we require the LLM to explicitly generate opinion tuples (aspect, opinion, polarity) alongside their corresponding polarity labels. To

ensure data reliability, we perform a polarity consistency check and automatically filter entries with conflicting predicted sentiment and extracted opinions, or those lacking sentiment cues. This self-verification yields a refined, low-noise corpus of high-quality opinion expressions, enhancing the reliability of subsequent clue extraction.

2) Clue Extraction: We propose a dual-channel framework for implicit sentiment clue extraction, integrating semantic enhancement and syntactic constraints to improve model performance in identifying implicit sentiment associations.

Semantic enhancement enriches the input context by retrieving semantically relevant opinion tuples from an external corpus $\mathcal{D}^T$. To ensure relevance to the target aspect, we employ an aspect-aware retrieval mechanism based on a hybrid similarity measure, which avoids task-specific fine-tuning and thus enhances generalizability. The similarity between the input instance $(x_i, a_{i,j})$ and a candidate tuple $(x_m, a_{m,n})$ from $\mathcal{D}^T$ is computed as:

$$\mathcal{S}_{m,n} = \text{SBert}(x_i, x_m) + \text{SBert}(a_{i,j}, a_{m,n}), \tag{4}$$

where $x_m \in \mathcal{D}^T$, $a_{m,n} \in x_m$ and SBert denotes Sentence-BERT. The most relevant sentiment reference is then selected as the implicit opinion clue $\hat{\mathcal{O}}^z_{i,j} = \arg\max\{\mathcal{S}_{m,n}\}$. Meanwhile, we extract the corresponding reference sentence x_k to facilitate subsequent syntactic constraint-based filtering.

Syntactic constraints ensure that only contextually and structurally relevant adjectives are considered as sentiment clues[8]. We first extract all adjectives in x_k and retain only those within a k-hop dependency distance from the aspect term. This dependency-driven filtering guarantees syntactic proximity and semantic plausibility. Specifically, we define extraction rules that construct a candidate set of adjectives, selecting those that are either within k hops in the dependency tree or directly connected via a short dependency path to the aspect.

Together, these two channels provide complementary signals: semantic enhancement brings in external knowledge grounded in aspect-aware relevance, while syntactic constraints offer local structural precision. By combining these controlled auxiliary cues, our framework effectively supports implicit sentiment reasoning with enriched yet well-constrained contextual information, generating the implicit clue $\hat{\mathcal{O}}^z_{i,j}$ for the final integration step.

C. Integration and Polarity Prediction

To solve the third core challenge, we integrate verified explicit opinions $\hat{\mathcal{O}}^{(T)}_{i,j}$ and implicit clues $\hat{\mathcal{O}}^z_{i,j}$ into a unified pipeline via an enriched input representation for polarity prediction. The input is constructed by concatenation with the delimiter "|":

$$\mathcal{X}_{i,j} = \text{“the opinion of } a_{i,j} \text{ is } \hat{\mathcal{O}}_{i,j} \mid a_{i,j} \mid x_i\text{”} \tag{5}$$

where $\hat{\mathcal{O}}_{i,j}$ is dynamically selected by $\hat{z}_{i,j}$: $\hat{\mathcal{O}}^z_{i,j}$ for $\hat{z}_{i,j} =$ True, and $\hat{\mathcal{O}}^{(T)}_{i,j}$ for $\hat{z}_{i,j} =$ False. Context-aware semantic encoding is performed via RoBERTa to generate deep contextualized representations $h_{i,j}$, followed by BiLSTM for long-range dependency modeling and a softmax classifier for final prediction. Finally, we adopt cross-entropy loss for model optimization:

$$h_{i,j} = \text{RoBERTa}(\mathcal{X}_{i,j}) \tag{6}$$

$$\hat{y}_{i,j} = \text{Softmax}\big(\text{BiLSTM}(h_{i,j})\big) \tag{7}$$

$$\mathcal{L}_{\text{ce}} = -\sum_{(i,j)\in D}\sum_{c\in C} y_{(i,j),c} \log \hat{y}_{(i,j),c} \tag{8}$$

where D is the training dataset and C is the set of sentiment classes (positive, negative, neutral). Minimizing this loss unifies the three core modules into a complete ISA solution.

TABLE I
STATISTICS OF THE REST 14 AND LAP 14 DATASET. THE "IMPLICIT" COLUMN DENOTES THE NUMBER OF SENTENCES WITH IMPLICIT SENTIMENT EXPRESSION AND "AVG. LEN" REFERS TO THE AVERAGE SENTENCE LENGTH.

Dataset	Positive	Negative	Neutral	Implicit	Avg. len
Lap 14_{train}	987	460	866	715	19.46
Lap 14_{test}	341	169	128	174	15.92
Rest 14_{train}	2164	633	805	1030	17.17
Rest 14_{test}	728	196	196	267	16.46

III. EXPERIMENTS

A. Settings

1) Dataset: We perform experiments on two publicly datasets from SemEval 2014 Task 4 [9]: Rest 14 and Lap 14. These datasets were categorized into explicit and implicit sentiments by Li[10]. The statistics of Rest 14 and Lap 14 are presented in Table I.

2) Evaluation metrics: We use Accuracy and F1-score to evaluate sentiment polarity classification, and introduce two complementary metrics for implicit sentiment analysis: ALL_F (Macro-F1 on the full dataset) and ISE_A (Accuracy on implicit samples).

3) Implementation Details: We adopt GPT-4o-mini for opinion generation and set the maximum self-debate iterations to three for a balance of effectiveness and computational efficiency. Under this setting, the assertion conflict ratios on Lap 14 and Rest 14 are only 3.44% and 2.41%, respectively, meaning such inconsistencies are negligible. For classification, RoBERTa-large is fine-tuned with the Adam optimizer at a learning rate of 3×10^{-5}. SBERT uses the all-mpnet-base-v2 model for sentence embedding, and SpaCy is employed for syntactic processing (adjective extraction and dependency tree construction) with a 2-hop threshold ($k = 2$) for syntactic relevance in dependency graphs. All experiments are conducted on a consumer-grade NVIDIA RTX 3090 (24GB) GPU, verifying the approach's feasibility without large-scale computing infrastructure.

TABLE II
MAIN RESULT. RESULTS MARKED WITH ♮, †, AND ◇ ARE FROM [11], [2], AND OUR REPRODUCTION, RESPECTIVELY. **ALL ARE RE-IMPLEMENTED AT COMPARABLE PARAMETER SCALES**. 'ALL' AND 'ISE' REFER TO RESULTS USING FULL DATA AND IMPLICIT SAMPLES. SUBSCRIPTS A AND F DENOTE ACCURACY AND MACRO-F1 SCORES.

	Lap 14		Rest 14	
	ALL$_F$	**ISE$_A$**	**ALL$_F$**	**ISE$_A$**
RGAT♮	75.95	75.42	80.17	68.55
BERTAsp♮	77.54	73.71	81.07	67.05
ASA-WD♮	77.38	72.57	80.92	68.53
CLEAN	77.25	78.29	81.40	69.66
ABSA-ESA	79.34	80.00	81.74	70.78
THOR†	77.51	74.29	81.10	68.54
RVISA	77.13	75.43	78.92	66.67
MT-ISA	79.86	80.57	82.45	70.41
Feedback-ICL◇	61.03	62.89	68.39	48.64
ICL zero-shot	74.15	64.79	75.85	62.85
ICL random-5 shot	78.58	71.41	79.94	71.16
ICL semantic-5 shot	79.30	73.62	81.32	73.12
SFT Qwen 2.5	77.57	73.71	81.03	71.53
OER (Ours)	**81.83**	**84.21**	**84.04**	**76.77**

TABLE III
ABLATION STUDIES

	Lap 14		Rest 14	
	ALL$_F$	**ISE$_A$**	**ALL$_F$**	**ISE$_A$**
OER	81.83	84.21	84.04	76.77
w/o ISCE	80.29	81.71	83.18	74.90
w/o ISCE & EOE	79.34	80.52	82.64	73.03
- Self-Debate	79.61	77.71	82.91	73.96
- EFGM	80.18	80.00	82.83	73.03

4) Baseline: We compare our model with a range of SOTA ISA methods, spanning T5-based, BERT-based, and LLM-based approaches. T5-based methods: We evaluate THOR[1], RVISA[2], MT-ISA[3] and ABSA-ESA[11]; BERT-based methods: We include R-GAT[12] (graph attention with dependency relations), ASA-WD[13] (memory networks for long-range dependencies), BERTAsp[4] (contrastive post-training), and CLEAN[14] (causal inference modeling). All results are re-implemented at comparable parameter scales; LLMs: We also compare with GPT-4o-mini and Qwen2.5-7B-Instruct.

B. Result Analysis

1) Main Result: We demonstrate the performance of our model and the baseline models in Table II. It can be observed that OER achieves the best ISE$_A$ and ALL$_F$ scores on both the Lap 14 and Rest 14 datasets, with improvements of 4.51% and 7.32% in ISE$_A$ values compared to the second-best results, respectively. This indicates that OER excels in identifying the sentiment polarity of given aspects in reviews containing implicit sentiments. Furthermore, it demonstrates that uncovering opinion information can further assist the model in understanding implicit sentiments within sentences, thereby effectively enhancing model performance.

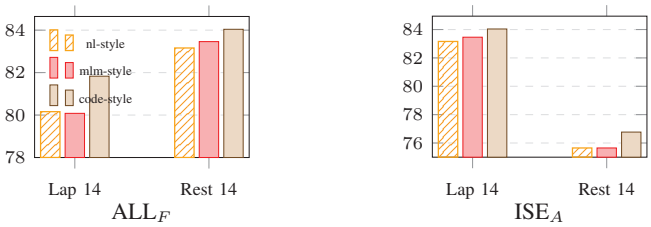

Fig. 3. The effects of different prompt settings. Bars (left to right) correspond to natural language-based task descriptions, masked language modeling (fill-in-the-blank format), and structured code-style prompts.

2) Ablation Studies: In Table III, we present the results of systematic ablation studies conducted on the Lap 14 and Rest 14 datasets. The experimental results show that removing the ISCE module led to a measurable decrease in model performance, which aligns with the module's design goal of extracting implicit sentiment cues. This trend is consistent with Lap 14's higher proportion of implicit cases (30.12%), as the ISCE module plays a more direct role in capturing implicit emotional signals—though the magnitude of performance decline between Lap 14 and Rest 14 remains moderate. This outcome confirms that accurately extracting implicit sentiment cues is crucial for revealing the true emotional tone behind text and enhancing the model's ability to understand complex emotional expressions. Additionally, when all opinion information was removed (condition: w/o ISCE & EOE), ISE_A dropped by 4.38 and 4.87 points on Lap 14 and Rest 14, respectively. This underscores the importance of opinions as semantic anchors in guiding implicit sentiment reasoning: the EOE module, by providing explicit opinion information, not only helps the model better understand explicit opinions but also supports deeper analysis of implicit emotions.

To further analyze the contribution of individual components within the EOE module, we conducted additional ablation studies by separately removing the self-debate mechanism and the EFGM. Compared to the full model, both ablation conditions (- Self-Debate, - EFGM) resulted in decreased overall performance and lower accuracy on implicit samples in Lap 14. This suggests that opinions generated by LLMs may contain hallucinations that can mislead the model—an issue that is more pronounced when the self-debate mechanism is removed, confirming its critical role in ensuring robustness during semantic verification. Furthermore, the decline in performance when removing EFGM highlights its value in enhancing the reliability of opinion extraction. Together, these findings emphasize that both the self-debate and EFGM modules are vital for maintaining model performance, especially in complex sentiment analysis scenarios involving high proportions of implicit cases.

3) Effect of Different Prompt Template: The experimental results shown in Figure 3 indicate a convergence in performance between the nl-style and mlm-style prompt templates. In contrast, the structured code-template prompt significantly enhances the accuracy of capturing sentiment elements. This

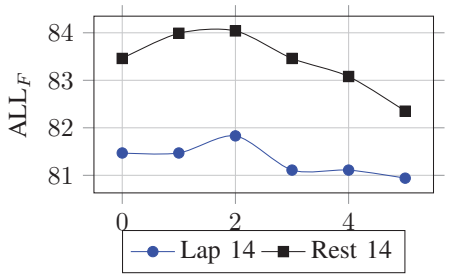

Fig. 4. The Impact of K Value Settings on the Results

phenomenon verifies the superiority of using a code-style prompt to guide LLMs in opinion extraction: the strong structural characteristics of code-style prompts are highly consistent with the target patterns of information extraction tasks. Their strict syntactic constraints help achieve precise alignment between task requirements and model understanding, not only enhancing the model's comprehension of specific tasks but also increasing the accuracy of opinion information extraction. This finding aligns with our designed Python function-style prompt template, and it is this structural design that underpins the stable opinion extraction performance of the OER model, thus forming a rigorous closed loop between our experimental results and method design.

4) The Impact of k: We illustrate the fluctuation of the ALL$_F$ metric with the change of parameter k in Figure 4. It can be observed from the graph that as the value of k increases, the curve for both datasets initially rises slowly and then drops sharply, with the optimal performance achieved at $k = 2$. When the value of k is small, the ALL$_F$ metric gradually increases, which may be attributed to the loss of some detailed information when relying on LLMs for extracting opinion information. Fortunately, such omissions are relatively rare and have a minor impact on the overall results. Specifically, $k = 2$ strikes a perfect balance: it preserves valid syntactic associations between sentiment clues and target aspects while avoiding the introduction of irrelevant emotional noise. In contrast, as k exceeds 2, the metric declines significantly because larger k values incorporate words with weak syntactic correlations to the target aspect, thereby introducing excessive noise and leading to a dramatic reduction in the F1 score. Therefore, selecting an appropriate k value is crucial for balancing model performance and noise control.

C. Further Discussion

1) Effect on MAMs: To further verify OER's robustness in handling complex contextual interference, we evaluate OER on the MAMs dataset[17] (Table IV), which consists of multi-aspect sentences with conflicting polarities. In this setting, precise opinion-word identification is critical to resolve aspect-sentiment entanglement. CoT outperformed zero-shot baselines yet underperformed few-shot methods, a shortcoming mainly caused by generation hallucinations. In contrast, OER enhances the reliability of opinion extraction via EFGM, and its self-debate mechanism verifies the semantic relevance between opinions and target aspects—this dual design effectively avoids cross-aspect opinion-word misallocation and mitigates contextual explicit interference, a common issue that plagued both few-shot and SFT RoBERTa models. By taking opinion words as auxiliary anchors to strengthen aspect-opinion associations, OER ultimately achieves notable performance gains on this challenging dataset.

TABLE IV
EXPERIMENTAL RESULTS ON MAMS.

Baseline	**Acc**	F_1
APSCL[15]	84.06	83.50
CVIB[16]	84.66	84.03
ChatGPT-4o-mini$_{zero-shot}$	54.49	53.02
ChatGPT-4o-mini$_{few-shot}$	67.44	50.96
ChatGPT-4o-mini$_{COT}$	58.83	44.10
SFT Roberta	83.79	83.56
OER(Ours)	84.61	84.33

TABLE V
THE IMPACT OF DIFFERENT PARAMETERS.

Model		P	ALL$_F$	
			Lap 14	Rest 14
Ours	RoBerta$_{base}$	125M	79.77	82.52
	RoBerta$_{large}$	355M	81.83	84.04
ABSA-ESA	T5$_{base}$	220M	79.34	81.74
MT-ISA	Flan-T5$_{base}$	250M	79.86	82.45

2) The impact of different parameters: Two state-of-the-art T5-based baselines are compared in Table V, with RoBERTa-base added for parameter scale matching. Our model shows advantages at comparable scales (<1B parameters): RoBERTa-base delivers performance close to larger T5 models, and scaling to 355M parameters yields a notable boost ($\Delta > 1.5$). At matched scales, our method outperforms baselines in overall F1 and key implicit discrimination accuracy, validating its effectiveness and efficiency for ISA.

3) Token Cost And Time Efficiency Analysis: We report the token-level inference cost analysis of our method and baselines in Table VI. Despite incorporating multi-step self-debate reasoning, our method maintains favorable cost-effectiveness by limiting the maximum number of iterations and only invoking the lightweight GPT-4o-mini model. The average time cost of each stage is 2.64 s for EOE, 0.09 s for ISCE, and 0.08 s for final inference. The EOE module constitutes the time

TABLE VI
TOKEN COST PER SAMPLE IN THE TEST DATASET

		ALL$_F$	Token cost	
			Input	Output
Lap 14	5-shot	74.15	314	21
	Ours	81.83	989	107
Rest 14	5-shot	75.85	267	16
	Ours	84.04	643	74

bottleneck, while the overall per-sample inference time is still below 3 seconds. Although our method slightly increases token consumption and inference time cost, such cost is fully acceptable, and the significant performance improvement fully justifies it, especially on hard implicit sentiment samples.

IV. Conclusion

This paper proposes the OER framework to address implicit sentiment analysis externalizing aspect-related opinions. OER introduces two key innovations: (1) in the *Explicit Opinion Extraction* module, a self-correction mechanism and a self-debate process mitigate hallucination in opinion extraction, and (2) in the *Implicit Sentiment Clue Extraction* module, sentiment-enhancing corpora reduce cumulative errors, ensuring more accurate implicit sentiment inference. Experimental results demonstrate that OER outperforms baselines, achieving an overall F1 improvement of 1.59 to 1.97 points and a 6.19% increase in implicit sentiment case accuracy. These findings confirm the effectiveness of OER in complex contexts and highlight its potential as a novel solution for ISA.

Acknowledgment

This research was supported in part by Natural Science Foundation of China (U24A20233, 62406078, 62476163), National Science and Technology Major Project (2021ZD0111502), the Guangdong Basic and Applied Basic Research Foundation (2023B1515120020), Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)(GML-KF-24-23), Collaborative Education Project of the Ministry of Education (202407), and CCF-DiDi GAIA Collaborative Research Funds (CCF-DiDi GAIA 202521).

References

[1] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T.-S. Chua, "Reasoning implicit sentiment with chain-of-thought prompting," 2023.

[2] W. Lai, H. Xie, G. Xu, and Q. Li, "Rvisa: Reasoning and verification for implicit sentiment analysis," 2024.

[3] W. Lai, H. Xie, G. Xu, and Q. Li, "Multi-task learning with llms for implicit sentiment analysis: Data-level and task-level automatic weight learning," 2024.

[4] Z. Li, Y. Zou, C. Zhang, Q. Zhang, and Z. Wei, "Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training," 2021.

[5] P. Li, T. Sun, Q. Tang, H. Yan, Y. Wu, X. Huang, and X. Qiu, "Codeie: Large code generation models are better few-shot information extractors," 2023.

[6] O. Sainz, I. García-Ferrero, R. Agerri, O. L. de Lacalle, G. Rigau, and E. Agirre, "GoLLIE: Annotation guidelines improve zero-shot information-extraction," in *The Twelfth International Conference on Learning Representations*, 2024.

[7] Y. Mo, J. Liu, J. Yang, Q. Wang, S. Zhang, J. Wang, and Z. Li, "C-icl: Contrastive in-context learning for information extraction," 2024.

[8] B. Chen, Q. Ouyang, Y. Luo, B. Xu, R. Cai, and Z. Hao, "S^2GSL: Incorporating segment to syntactic enhanced graph structure learning for aspect-based sentiment analysis," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 13366–13379, Association for Computational Linguistics, Aug. 2024.

[9] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (P. Nakov and T. Zesch, eds.), (Dublin, Ireland), pp. 27–35, Association for Computational Linguistics, Aug. 2014.

[10] R. Li, H. Chen, F. Feng, Z. Ma, X. Wang, and E. Hovy, "Dual graph convolutional networks for aspect-based sentiment analysis," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong, F. Xia, W. Li, and R. Navigli, eds.), pp. 6319–6329, Association for Computational Linguistics, Aug. 2021.

[11] J. Ouyang, Z. Yang, S. Liang, B. Wang, Y. Wang, and X. Li, "Aspect-based sentiment analysis with explicit sentiment augmentations," 2023.

[12] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3229–3238, 2020.

[13] Y. Tian, G. Chen, and Y. Song, "Enhancing aspect-level sentiment analysis with word dependencies," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (P. Merlo, J. Tiedemann, and R. Tsarfaty, eds.), (Online), pp. 3726–3739, Association for Computational Linguistics, Apr. 2021.

[14] S. Wang, J. Zhou, C. Sun, J. Ye, T. Gui, Q. Zhang, and X. Huang, "Causal intervention improves implicit sentiment analysis," 2022.

[15] P. Li, P. Li, and X. Xiao, "Aspect-pair supervised contrastive learning for aspect-based sentiment analysis," *Know.-Based Syst.*, vol. 274, Aug. 2023.

[16] M. Chang, M. Yang, Q. Jiang, and R. Xu, "Contrastive variational information bottleneck for aspect-based sentiment analysis," *Know.-Based Syst.*, vol. 284, Jan. 2024.

[17] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, "A challenge dataset and effective models for aspect-based sentiment analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 6280–6285, Association for Computational Linguistics, Nov. 2019.