

# TMAK-Plus at SIGHAN-2024 dimABSA Task: Multi-Agent Collaboration for Transparent and Rational Sentiment Analysis

Xin Kang<sup>1</sup>, Zhifei Zhang<sup>2</sup>, Jiazheng Zhou<sup>1</sup>,  
Yunong Wu<sup>3</sup>, Xuefeng Shi<sup>4</sup>, Kazuyuki Matsumoto<sup>1</sup>,

<sup>1</sup>Department of Computer Science, Tokushima University, Tokushima, Japan.

<sup>2</sup>Department of Computer Science and Technology, Tongji University, Shanghai, China.

<sup>3</sup>NLP Department, Dataa Robotics (Chengdu Branch), Chengdu, China.

<sup>4</sup>School of Artificial Intelligence and Computer Science, Nantong University, Nantong, China.

Correspondence: kang-xin@is.tokushima-u.ac.jp

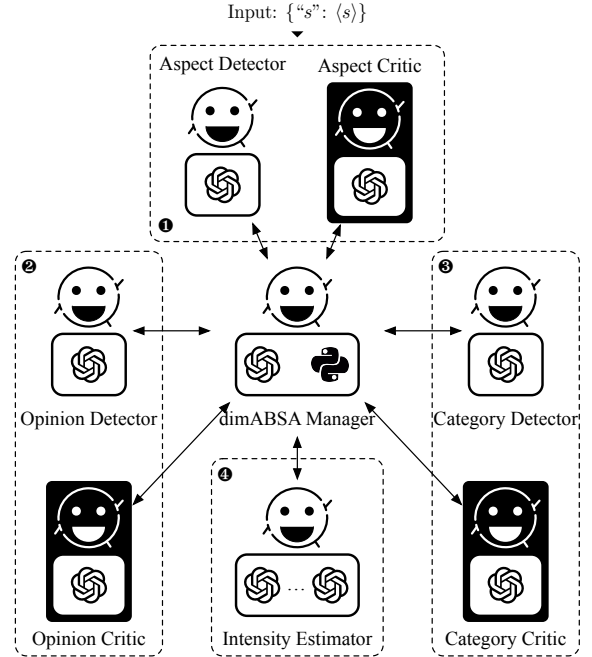
## Abstract

The TMAK-Plus team proposes a Multi-Agent Collaboration (MAC) model for the dimensional Aspect-Based Sentiment Analysis (dimABSA) task at SIGHAN-2024. The MAC model leverages Neuro-Symbolic AI to solve dimABSA transparently and rationally through symbolic message exchanges among generative AI agents. These agents collaborate on aspect detection, opinion detection, aspect classification, and intensity estimation. We created 8 sentiment intensity agents with distinct character traits to mimic diverse sentiment perceptions and average their outputs. The AI agents received clear instructions and 20 training examples to ensure task understanding. Our results suggest that the MAC model is effective in solving the dimABSA task and offers a transparent and rational approach to understanding the solution process.

## 1 Introduction

We consider the dimABSA task a challenging sentiment analysis problem (Cai et al., 2021; Pontiki et al., 2016) that requires a deep understanding of natural language and the ability to sense sentiments with distinct character traits. Specifically, aspect detection, opinion detection, and aspect classification account for distinct sentiment analysis abilities, while intensity estimation requires a comprehensive understanding of valence and arousal intensities corresponding to the empathetic sentiment sensitivity of different individuals.

To address the dimABSA task, we propose a MAC model that sequentially solves the aforementioned challenges, as shown in Figure 1. The model consists of GPT-4o (gpt-4o-2024-05-13) powered generative neural AI agents, including AD, OD, CD, AC, OC, CC, and IE. We pair AD and AC for collaborative aspect detection, OD and OC for collaborative opinion detection, and CD and CC for collaborative aspect classification. IE



Output: {"s": "s"}, "r": [{"a": "a"}, {"c": "c"}, {"o": "o"}, {"i": "v">#(a)}]}

Figure 1: MAC model in the dimABSA Task. The central dimABSA Manager (DM) agent manages the overall collaboration, reading the input sentence, generating the dimABSA result, and coordinating other agents. These agents include the Aspect Detector (AD) and Opinion Detector (OD) for sentiment term extraction, the Category Detector (CD) for aspect classification, the Aspect Critic (AC), Opinion Critic (OC), and Category Critic (CC) for critical evaluation, and the Intensity Estimator (IE) for averaging sentiment intensities from 8 distinct estimators. All messages, including the input sentence, output quadruples, and intermediate results, are exchanged in JSON format.

averages the sentiment intensities from 8 estimators with distinct character traits. All agents exchange messages in JSON format, including the detected results or critical feedback, in a symbolic manner to ensure transparency and rationality.

Unlike previous approaches that focus on the end-to-end training of a single neural network with a particular training set (Chen et al., 2021; Mao

et al., 2021, 2022; Peng et al., 2020; Xu et al., 2020; Peper and Wang, 2022; Zhang et al., 2021), the MAC model is powered by multiple generative AI agents. These agents require only a few training examples and clear instructions to perform their tasks in a more robust manner and generalize easily to out-of-domain data. Our results demonstrate that MAC provides a generalizable and transparent Neuro-Symbolic AI framework for solving key phrase detection, multi-class classification, and regression tasks that require a deep understanding of natural language.

The main contributions are as follows:

- We propose a transparent and rational MAC model for the dimABSA task.
- We provide a new paradigm of Neuro-Symbolic AI powered by generative AI with symbolic collaboration.
- We demonstrate the effectiveness and generalizability of MAC in solving a challenging sentiment analysis task.

## 2 MAC Model for dimABSA

### 2.1 Formal Definition of dimABSA

The dimABSA task consists of 3 subtasks:

**Subtask 1: Intensity Prediction** involves predicting sentiment intensities (i) in the valence and arousal (v#a) dimensions for given aspect terms in a sentence.

**Subtask 2: Triplet Extraction** requires extracting sentiment triplets composed of an aspect term (a), an opinion term (o), and their corresponding intensity (i).

**Subtask 3: Quadruple Extraction** focuses on extracting sentiment quadruples that include an aspect term (a), an aspect category (c), an opinion term (o), and their intensity (i).

We use the following running example throughout this paper. The example is presented in traditional Chinese with an English translation for clarity and analysis purposes.

```
{
  "k": "R0645:S125",
  "s": "這牛排外面裹著一層麵包粉看起來蠻粉嫩的，吃下去外皮酥脆卡滋卡滋真的好吃。",
  e_s: "This steak is coated with a layer of breadcrumbs on the outside, making it look quite tender. When you bite into it, the crust is crispy and crunchy and really delicious.",
  "a": ["牛排", "外皮", "牛排"],
  e_a: ["steak", "crust", "steak"],
```

```
"c": ["食物#品質", "食物#品質", "食物#品質"],
e_c: ["food#quality", "food#quality", "food#quality"],
"o": ["蠻粉嫩的", "酥脆", "真的好吃"],
e_o: ["quite tender", "crispy", "really delicious"],
"i": ["6.25#5.75", "6.62#6.0", "6.88#6.62"]
}
```

In this example, the aspect terms (a) are 牛排 (steak) and 外皮 (crust), the opinion terms (o) are 蠻粉嫩的 (quite tender), 酥脆 (crispy), and 真的好吃 (really delicious), and the sentiment intensities (i) are expressed as valence and arousal scores (v#a) with  $v, a \in [1, 9]$ .

### 2.2 MAC Model Architecture

Figure 2 illustrates the MAC model architecture for the dimABSA task. This model integrates multiple generative AI agents that collaborate to perform the subtasks required for dimABSA. The agents involved in this process include DM, AD, AC, OD, OC, CD, CC, and various IE agents.

The DM agent manages the overall collaboration, reading input sentences, generating results, and coordinating the other agents. The AD agent detects all sentiment aspects in a given sentence, while the AC agent evaluates the performance of AD and provides constructive feedback. The OD agent detects sentiment opinions associated with each aspect, and the OC agent evaluates the performance of OD, offering feedback. The CD agent classifies each aspect into predefined categories, and the CC agent evaluates the performance of CD, ensuring accuracy and consistency.

The 8 sentiment intensity estimators in the MAC model reflect distinct human characters, grounded in psychological and linguistic theories. According to the Big Five personality traits model (John et al., 1999), human personalities can be categorized into dimensions such as openness, conscientiousness, extraversion, agreeableness, and neuroticism, each influencing how individuals perceive and react to emotional stimuli. Additionally, the circumplex model of affect (Russell, 1980) provides a framework for understanding emotions in a valence-arousal space. Integrating these perspectives ensures the model captures a broad range of human emotional responses, enhancing its robustness and generalizability. Specifically, the IE agents estimate sentiment intensities analytically (IE<sup>A</sup>), empathetically (IE<sup>E</sup>), critically (IE<sup>C</sup>), optimistically (IE<sup>O</sup>), realistically (IE<sup>R</sup>), pessimistically (IE<sup>P</sup>), balanced (IE<sup>B</sup>), and intuitively (IE<sup>I</sup>).

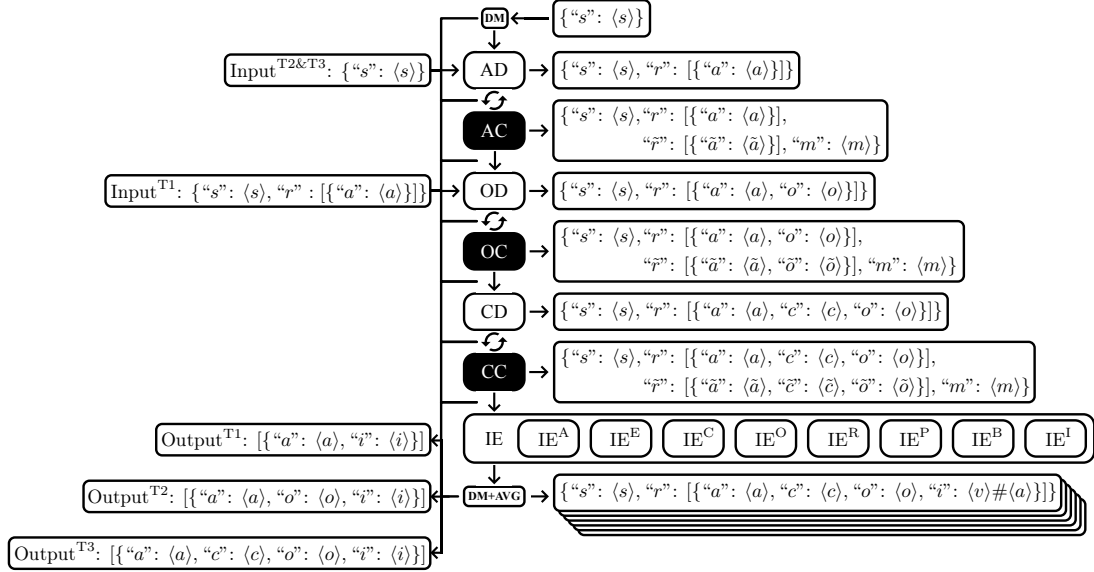


Figure 2: MAC Model Architecture. This diagram illustrates the MAC prediction process, involving agents DM, AD, AC, OD, OC, CD, CC, and intensity estimators (IE<sup>A</sup>, IE<sup>E</sup>, IE<sup>C</sup>, IE<sup>O</sup>, IE<sup>R</sup>, IE<sup>P</sup>, IE<sup>B</sup>, IE<sup>I</sup>). Inputs and outputs of subtasks T1, T2, and T3 are depicted, with T1 connecting to OD and T2&T3 connecting to AD. Outputs are generated by DM + AVG, showing the entire prediction process. JSON-formatted messages flow between agents, managed by DM.

The input for subtask T1 is processed by the OD agent, as it already contains the aspect (a). In contrast, the input for subtasks T2 and T3, which only contain the sentence (s), is processed by the AD agent. Outputs of T1, T2, and T3 are generated by the bottom agent DM + AVG, showing that all final outputs are predicted through the entire process, with variations only in their format.

Symbolic messages in JSON format are exchanged among these neural generative AI agents, as depicted in Figure 2, ensuring a transparent and rational approach to solving the dimABSA task. Critic agents provide feedback messages (m) indicating suggested results for aspects ( $\tilde{a}$ ), categories ( $\tilde{c}$ ), and opinions ( $\tilde{o}$ ), while the original results from detector agents do not include these symbols. This new paradigm of Neuro-Symbolic AI not only addresses the dimABSA problem but also provides a general framework for solving key phrase detection, multi-class classification, and regression tasks.

### 3 Experimental Results

#### 3.1 Experimental Setup

The dimABSA task provides a dataset with 3,000 training examples, each manually annotated with aspect terms, aspect categories, opinion terms, and sentiment intensities. There are also two test sets, each containing 2,000 examples for subtask T1 and

subtasks T2&T3. The detailed annotation guidelines and data splits are described in the official summary paper of the dimABSA task (Lee et al., 2024).

For T1, the evaluation metrics include Mean Absolute Error (MAE) to measure the accuracy of predicted  $v$  and  $a$  scores, and Pearson Correlation Coefficient (PCC) to assess the correlation between predicted and actual values. For T2&T3, Precision, Recall, and F1-score assess the accuracy of the extracted triplets (a, o, i) and quadruples (a, c, o, i). Detailed experimental setup can be found in Appendix A.

#### 3.2 Experimental Results

We report our experimental results in Tables 1 and 2, where V, A, and VA represent results related to Valence, Arousal, and both Valence and Arousal, respectively. Due to format issues, only the results of T2 are officially ranked, while the results of T1 and T3 were post-processed and evaluated through the post-evaluation process and are provided for reference. Detailed result comparisons with other participating teams are available in the summary paper (Lee et al., 2024).

#### 3.3 Result Analysis

We demonstrate the transparency and rationality of the MAC model through the analysis of agent col-

Subtask	V-MAE	V-PCC	A-MAE	A-PCCA
T1 <sup>post</sup>	0.4706	0.9266	0.4618	0.6745

Table 1: Experimental Results for dimABSA subtask 1.

Subtask	V-P	V-R	V-F1	A-P	A-R	A-F1	VA-P	VA-R	VA-F1
T2	25.64	28.24	26.88	29.31	32.28	30.72	14.97	16.49	15.69
T3 <sup>post</sup>	23.85	26.19	24.97	27.68	30.40	28.98	14.15	15.54	14.81

Table 2: Experimental Results for dimABSA subtasks 2 and 3, in percentage.

laborations based on the running example provided in Section 2.1. For a detailed analysis, please refer to Appendix B.

Given the input restaurant review sentence, the AD agent detects aspect terms (a) 牛排 (steak) and 外皮 (crust), consistent with the ground truth. The AC agent evaluates AD’s performance, provides critical feedback, and suggests results ( $\bar{r}$ ) with feedback messages (m).

Once AD and AC reach a consensus, DM forwards it to the OD agent to detect opinion terms (o). The OC agent then evaluates OD’s performance, pointing out that 好吃 (delicious) should be 真的好吃 (really delicious), aligning with the ground truth. DM forwards the consensus to the CD agent to classify aspect terms into predefined categories (c), and the CC agent evaluates CD’s performance and concurs with its output.

Finally, DM forwards the consensus to 8 IE agents, each with distinct character traits, to estimate sentiment intensities and average their outputs as the consensus intensity (i). The final prediction, shown below, correctly identifies aspects, categories, and opinions, and provides intensity values close to the ground truth.

```
{
  "s": same_as_above,
  "r": [
    {"a": "牛排", "c": "食物#品质",
     "o": "蛮粉嫩", "i": "6.06#5.56"},
    {"a": "外皮", "c": "食物#品质",
     "o": "酥脆卡滋卡滋", "i": "6.71#6.22"},
    {"a": "牛排", "c": "食物#品质",
     "o": "真的好吃", "i": "7.19#6.72"},
  ]
}
```

## 4 Conclusion

In this paper, we proposed a MAC model for the dimABSA task, representing a new paradigm of Neuro-Symbolic AI. Our approach employs multiple generative AI agents, each specializing in different subtasks, ensuring a robust and transparent

workflow. The MAC model operates in a transparent and rational manner, demonstrated by its accurate identification of aspects, categories, opinions, and sentiment intensities. By incorporating agents with distinct character traits, we mimic the sentiment receptions of different individuals, enhancing the model’s effectiveness. Additionally, the use of generative AI agents with few-shot learning enables MAC to easily generalize to out-of-domain data.

Future work will focus on addressing the hallucination problem within the Neuro-Symbolic AI framework, improving reliability and consistency of outputs, and extending the model’s applicability to other domains.

## Limitations

This study presents several limitations that should be considered. Firstly, few-shot learning with only 20 examples may not capture the full variability and nuances of the data, potentially leading to less robust models compared to those fine-tuned with the entire dataset. While fine-tuning with the entire dataset could improve task-specific performance, it may reduce the generalization ability of the agents, making them less adaptable to unseen data or different domains. Future research could investigate hybrid learning approaches that integrate the strengths of both few-shot and full dataset methods to enhance model robustness and generalizability.

Although the critic multi-agent collaboration (MAC) framework effectively mitigates error propagation, the sequential nature of the model could still lead to cumulative errors if initial detections are flawed. Future efforts could focus on developing more sophisticated error correction mechanisms and exploring alternative architectures that reduce the dependency on initial accuracy, thereby further minimizing the risk of error propagation.

Utilizing advanced models like GPT-4o requires



substantial computational resources, which might not be accessible to all researchers or practitioners. Additionally, the cost of calling the GPT-4o API may be a limitation for refining the agents or the MAC framework. Future research could explore fine-tuning more recent open-source large language models, such as LLaMA-3 (Meta, 2024) and Phi-3 (Abdin et al., 2024), as cost-effective alternatives. Fine-tuning these models could mitigate the financial and computational constraints while maintaining high performance and accessibility.

Despite efforts to ensure transparency, the complexity of the multi-agent system might make it challenging to interpret individual agent decisions and their contributions to the overall output. Enhancing model interpretability remains a crucial area for future work, potentially through improved visualization techniques and the development of methods to clearly attribute specific decisions to individual agents within the system.

Furthermore, the scalability of the proposed method to other languages, domains, or larger datasets has not been fully explored and might present additional challenges. Future research could test the scalability and adaptability of the MAC framework across various languages, domains, and dataset sizes to evaluate its broader applicability and performance.

## Acknowledgments

This work was supported by JKA and its promotion funds from KEIRIN RACE and National Natural Science Foundation of China (Grant Number 61906137).

## References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. 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)*, pages 340–350.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12666–12674.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian Liao. 2024. Overview of the sighthan 2024 shared task for chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*, pages 1–10. Association for Computational Linguistics.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13543–13551.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8600–8607.

Joseph Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6089–6095, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. pages 19–30, San Diego, California.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. [Towards generative 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 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

## A Detailed Experimental Setup

This section details the comprehensive templates utilized for various agents in our study, along with methods for hyper-parameter configuration and sample selection strategies. The templates standardize procedures and outputs, ensuring consistency and reproducibility across different experiments.

Below are the templates for different agents, including detection agents and critic agents. Each template specifies the agent’s name, task description, input and output formats, and examples. The JSON formats ensure structured input data and output results. For critic agents, additional instructions on identifying potential errors and providing feedback are included.

```
Agent Name: {AgentName}
Task Description: {TaskDescription}
1. Read the input in the following JSON
   format: {JSONInput}
2. Assess and identify {AgentTargets}.
3. Output the results strictly in the
   following JSON format: {JSONOutput}
Examples:
[{{Example}}]
```

```
Agent Name: {CriticAgentName}
Task Description: {CriticTaskDescription}
1. Read the input in the following JSON
   format:{CriticJSONInput}
2. Identify potential {ErrorTypes}.
3. Provide feedback and suggestions in the
   following JSON format: {CriticJSONOutput}
Examples:
[{{CriticExample}}]
```

The JSON formats for inputs and outputs can be found in Appendix B, with sample strings replaced by symbols: ⟨s⟩ for sentence, ⟨a⟩ for aspect, ⟨c⟩ for category, ⟨o⟩ for opinion, and ⟨i⟩ for intensity. For critic JSON, critic strings are represented as ⟨ã⟩, ⟨c̃⟩, ⟨õ⟩, and ⟨ĩ⟩. The message from the critic agent is denoted as ⟨m⟩.

Samples for few-shot learning are selected randomly from the training dataset and fixed using a seed value of 41. This modified stratified sampling method respects the original distribution of different aspect categories and ensures the inclusion of all aspect categories. Additionally, the

“NULL” aspect is included, which is unique to the dimABSA task. All agents share this common set of samples. The GPT-4o (gpt-4o-2024-05-13) model with the above system messages is used with the default temperature and top-p values as hyper-parameters.

## B Detailed Result Analysis

This section analyzes the results from the MAC model, focusing on transparency, rationality, and error analysis.

We demonstrate the transparency and rationality of the MAC model by analyzing the collaborations among the agents based on the running example provided in Section 2.1. The collaborative messages are JSON-formatted as shown in Figure 2, and we use the same format with English translations for clarity and analysis purposes.

```
{
  "s": "這牛排外面裹著一層麵包粉看起來蠻粉嫩的，吃下去外皮
        酥脆卡滋卡滋真的好吃。",
  "e_s": "This steak is coated with a layer of
          breadcrumbs on the outside, making it
          look quite tender. When you bite into
          it, the crust is crispy and crunchy
          and really delicious.",
  "r": [
    {"a": "牛排"},
    {"e_a": "steak"},
    {"a": "外皮"},
    {"e_a": "crust"}
  ]
}
```

The AC agent evaluates the performance of AD and provides critical feedback, with suggested results ( $\tilde{r}$ ) and feedback message (m), as follows.

```
{
  "s": same_as_above,
  "r": same_as_above,
  "r̃": [
    {"ã": "牛排"},
    {"e_ã": "steak"},
    {"ã": "外皮"},
    {"e_ã": "crust"}
  ],
  "m": "Correctly identified aspects."
}
```

Since AD and AC have reached a consensus, DM forwards this consensus to the OD agent to detect opinion terms (o), as follows.

```
{
  "s": same_as_above,
  "r": [
    {"a": "牛排", "o": "蠻粉嫩"},
    {"e_a": "steak", "e_o": "quite tender"},
    {"a": "外皮", "o": "酥脆卡滋卡滋"},
    {"e_a": "crust", "e_o": "crispy and crunchy"},
    {"a": "牛排", "o": "好吃"},
    {"e_a": "steak", "e_o": "delicious"}
  ]
}
```

```
]
}
```

The OC agent evaluates the performance of OD and critically points out that the opinion term 好吃 (delicious) should be 真的好吃 (really delicious), which aligns with the ground truth.

```
{
  "s": same_as_above,
  "r": same_as_above,
  "f": [
    {"a": "牛排", "o": "蠻粉嫩"},
    {"e_ã": "steak", e_õ: "quite tender"},
    {"a": "外皮", "o": "酥脆卡滋卡滋"},
    {"e_ã": "crust", e_õ: "crispy and crunchy"},
    {"a": "牛排", "o": "真的好吃"},
    {"e_ã": "steak", e_õ: "really delicious"}
  ],
  "m": "Incomplete opinion: '好吃' should be '真的好吃'."
  m_e: "Incomplete opinion: 'delicious' should be 'really delicious'."
}
```

DM then forwards the consensus to the CD agent to classify the aspect terms into predefined categories (c), as follows.

```
{
  "s": same_as_above,
  "r": [
    {"a": "牛排", "c": "食物#品質", "o": "蠻粉嫩"},
    {"e_a": "steak", e_c: "food#quality", e_o: "quite tender"},
    {"a": "外皮", "c": "食物#品質", "o": "酥脆卡滋卡滋"},
    {"e_a": "crust", e_c: "food#quality", e_o: "crispy and crunchy"},
    {"a": "牛排", "c": "食物#品質", "o": "真的好吃"},
    {"e_a": "steak", e_c: "food#quality", e_o: "really delicious"}
  ]
}
```

The CC agent evaluates the performance of CD and concurs with its output, as follows.

```
{
  "s": same_as_above,
  "r": same_as_above,
  "f": [
    {"a": "牛排", "c": "食物#品質", "o": "蠻粉嫩"},
    {"e_ã": "steak", e_õ: "food#quality", e_õ: "quite tender"},
    {"a": "外皮", "c": "食物#品質", "o": "酥脆卡滋卡滋"},
    {"e_ã": "crust", e_õ: "food#quality", e_õ: "crispy and crunchy"},
    {"a": "牛排", "c": "食物#品質", "o": "真的好吃"},
    {"e_ã": "steak", e_õ: "food#quality", e_õ: "really delicious"}
  ],
  "m": "Correctly classified categories."
}
```

Finally, DM forwards the consensus to 8 IE agents, each with distinct character traits, to estimate sentiment intensities and average their outputs as the consensus intensity (i). We use superscripts to denote intensity predictions given by the IE agents with distinct character traits, as follows.

```
{
  "s": same_as_above,
  "r": [
    {"a": "牛排", "c": "食物#品質", "o": "蠻粉嫩", "i": "6.06#5.56", "i^A": "6.00#5.50", "i^E": "6.00#5.00", "i^C": "6.00#5.75", "i^O": "6.00#6.00", "i^R": "6.50#5.75", "i^P": "5.50#5.00", "i^B": "6.00#5.50", "i^I": "6.50#6.00"},
    {"a": "外皮", "c": "食物#品質", "o": "酥脆卡滋卡滋", "i": "6.71#6.22", "i^A": "6.75#6.00", "i^E": "6.50#6.00", "i^C": "6.75#6.25", "i^O": "7.00#7.00", "i^R": "7.20#6.50", "i^P": "5.75#5.50", "i^B": "6.75#6.00", "i^I": "7.00#6.50"},
    {"a": "牛排", "c": "食物#品質", "o": "真的好吃", "i": "7.19#6.72", "i^A": "7.00#6.50", "i^E": "7.00#6.50", "i^C": "7.00#6.50", "i^O": "8.00#8.00", "i^R": "8.00#7.00", "i^P": "6.00#5.75", "i^B": "7.00#6.50", "i^I": "7.50#7.00"}
  ]
}
```

Analyzing the final predictions reveals a notable discrepancy in the opinion term 酥脆卡滋卡滋 (crispy and crunchy) predicted by the OD agent compared to the ground truth 酥脆 (crispy). This term was justified as reasonable by the OC agent, highlighting the variability in sentiment perception among humans, which AI agents reflect. Conversely, the OC agent correctly criticized another OD prediction, where 好吃 (delicious) was adjusted to 真的好吃 (really delicious), aligning with the ground truth. These observations suggest that critic multi-agent collaboration effectively mitigates the error propagation problem, which is more prevalent in sequential models. However, accurately mimicking the sentiment perception of a group of human beings, as reflected in the dimABSA data annotations, remains challenging. This issue could be mitigated by further fine-tuning the agents with the entire training dataset, although this approach is costly and may reduce the generalization capability of the agents.

The valence and arousal intensities predicted by the IE agents with distinct character traits also exhibit variability. These differences underscore the subjective nature of sentiment analysis, influenced by individual perspectives. The deviations between the IE agents' predictions and the ground truth values illustrate the difficulty in accurately

mimicking the diversity of sentiment perception in human beings.