Task-aware Contrastive Mixture of Experts for Quadruple Extraction in Conversations with Code-like Replies and Non-opinion Detection

Anonymous ACL submission

Abstract

This paper focuses on Dialogue Aspect-based Sentiment Quadruple (DiaASQ) analysis, aiming to extract structured quadruples from multi-turn conversations. Applying Large Language Models (LLMs) for this specific task presents two primary challenges: the accurate extraction of multiple elements and the understanding of complex dialogue reply structure. To tackle these issues, we propose a novel LLM-based multi-task approach, named **T**ask-**a**ware **Co**ntrastive **M**ixture **o**f **E**xperts (**TaCoMoE**), to tackle the DiaASQ task by integrating expert-level contrastive loss within task-oriented mixture of experts layer. TaCoMoE minimizes the distance between the representations of the same expert in the semantic space while maximizing the distance between the representations of different experts to efficiently learn representations of different task samples. Additionally, we design a Graph-Centric Dialogue Structuring strategy for representing dialogue reply structure and perform non-opinion utterances detection to enhance the performance of quadruple extraction. Extensive experiments are conducted on the DiaASQ dataset, demonstrating that our method significantly outperforms existing parameter-efficient fine-tuning techniques in terms of both accuracy and computational efficiency. The code is available at `https://anonymous.4open.science/r/TaCoMoE-08B4`.

1 Introduction

Dialogue Aspect-based Sentiment Quadruple (DiaASQ) is a newly-emergent task aiming to extract the sentiment quadruple (i.e., targets, aspects, opinions, and sentiments) from conversations (Li et al., 2023a), which plays a pivotal role in sentiment analysis (Cambria, 2016; Hu et al., 2020; Mao et al., 2024) and developing sentiment-support dialog systems (Merdivan et al., 2019; Zhou et al., 2022; Vlachos et al., 2024). The accurate dialogue quadruple extraction can benefit sentiment analysis, clinical treatment (Chen et al., 2020b; Tu et al., 2024), product and service feedback (Mukku et al., 2023), etc.

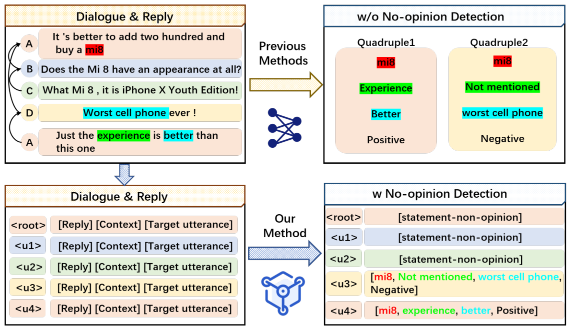


Figure 1: A concrete DiaASQ sample demonstrating how our approach with LLM architectures differs from traditional methods.

Compared to traditional Aspect-based Sentiment Analysis (ABSA) tasks that extracting opinions or sentiment preferences towards specific aspects from a single piece of text (Zhang et al., 2021b; Yan et al., 2021; Deng et al., 2023), the DiaASQ task is notably more challenging due to its complex multi-party dialogue structure and contextual dependencies. Recently, the research on dialogue aspect-based sentiment quadruple has been gradually gaining recognition, leading to a series of advancements (Li et al., 2023a; Luo et al., 2024b; Li et al., 2024). In addition, Large Language Models (LLMs) have demonstrated significant potential in Aspect-based Sentiment Analysis tasks (Fei et al., 2023; Varia et al., 2023; Wang et al., 2023). However, the effectiveness of LLMs on the DiaASQ task has not been effectively explored and existing studies for DiaASQ have several key limitations which prevent their performance.

Firstly, insufficient learning of cross-task shared features and knowledge. DiaASQ involves multiple tasks (e.g., single-element extraction, quadruple extraction), and traditional methods struggle to fully utilize the complementarity between tasks

(Chen et al., 2020a; Scaria et al., 2024), resulting in the model failing to achieve consistent performance across all tasks. Secondly, lack of effective modeling for dialogue reply dependency structures. Previous methods often require complex graph representation encoders to explicitly model these dependency structures (Zhang et al., 2023; Li et al., 2024), which increases computational overhead and complexity, especially when applied to large language models (Zhang et al., 2022; Fatemi et al., 2024). Thirdly, the impact of non-opinion utterances on DiaASQ performance has not been thoroughly investigated. These utterances often account for a significant proportion of the data and can interfere with the model's understanding and predictions (Larson et al., 2019; Zhang et al., 2024). Figure 1 illustrates a comparison between previous methods and our generative large model-based approach, in which we perform quadruple extraction and non-opinion detection for each utterance.

In this paper, we propose a novel approach called **T**ask-**a**ware **Co**ntrastive **M**ixture **o**f **E**xperts (**TaCoMoE**) framework for the DiaASQ task, which integrates task-oriented mixture of experts layer into LLM with contrastive learning to learn distinct task-shared and -specific knowledge. Specifically, we first introduce the extraction of individual elements and the analysis of dialogue reply dependencies, in addition to the main task of quadruple extraction. On one hand, for all tasks that involve dialogue dependency inputs or target outputs, we design a formalized text description strategy to encourage large models to efficiently utilize dialogue reply dependencies. On the other hand, we treat utterances that do not contain any quadruples as recognition targets as well, as these utterances often constitute a significant proportion in real-world scenarios. Secondly, we perform utterance-level processing with task-oriented routing, which is integrated into the LLM, to learn separate sets of parameters for each task. Additionally, each expert is designed as two low-rank matrices to ensure parameter efficiency. Finally, we introduce contrastive learning into each task-oriented Mixture of Experts layer, treating outputs from the same expert as positive pairs and outputs from different experts as negative pairs to learn the distinct features of different tasks.

We conduct experiments on the public DiaASQ benchmark dataset, which includes both English and Chinese data. Results consistently demonstrate that our TaCoMoE significantly outperforms other state-of-the-art methods on the DiaASQ task, showing the effectiveness and superiority of our method. Additionally, our analysis indicates that considering non-opinion utterances in the DiaASQ task is essential and has a positive impact on quadruple extraction.

Our main contributions can be summarized as follows:

- We introduce a novel LLM-based approach for addressing the DiaASQ task by incorporating expert-level contrastive loss into task-oriented mixture of experts layer.
- We explore converting dialogues into a universal code-like format to represent reply dependency structures between utterances, eliminating the need for an additional graph encoder.
- We explicitly consider non-opinion utterances and validate that identifying these utterances also make a crucial contribution to the DiaASQ task.
- Extensive experimental results demonstrate that our method surpasses existing state-of-the-art (SOTA) approaches and validate the effectiveness of key components in our framework.

2 Related Work

The related work is provided in Appendix A.

3 Method

We begin by providing a formal definition of the DiaASQ task. A dialogue is represented as a sequence of utterances paired with their respective speakers: $D = \{(s_1, u_1), (s_2, u_2), \ldots, (s_{|D|}, u_{|D|})\}$, where $u_i = \{w_{i1}, w_{i2}, \ldots\}$ denotes the i-th utterance as a set of tokens, and s_i indicates the speaker of u_i. In addition, a reply list $L = \{l_1, l_2, \ldots, l_{|D|}\}$ is provided, where l_i identifies the current utterance u_i is replying to. The primary objective of this task is to extract a collection of **quadruples**: $C = \{(t_i, a_i, o_i, p_i)\}_{i=1}^{|C|}$, where t_i, a_i, o_i, and p_i are spans that correspond to the *target*, *aspect*, *opinion*, and *sentiment polarity*, respectively.

The proposed TaCoMoE consists of three main components: dialogue input engineering, task-oriented mixture of experts layer, and contrastive loss. The overall architecture of TaCoMoE is illustrated in Figure 2.

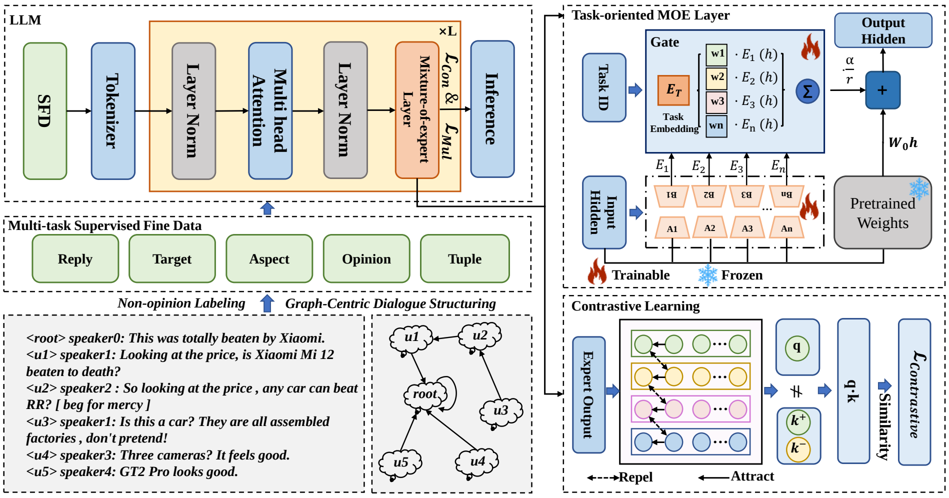


Figure 2: Illustration of the overall framework of TaCoMOE, which consists of three essential components: Dialogue Input Engineering, Task-oriented Mixture of Experts Layer, Contrastive Learning.

3.1 Dialogue Input Engineering

To enhance the model's understanding of dialogue reply relationships and improve the accuracy of element extraction, we introduced three single-element extraction tasks and a dialogue reply relationship analysis task in addition to the quadruple extraction task, aiming to capture multi-dimensional features.

The first challenge is how to align the dialogue reply dependencies with the sequence format or structure required by LLMs. Building upon previous work addressing the alignment between graphs and text (Wang et al., 2024), we propose a *Graph-Centric Dialogue Structuring* (GCDS) strategy to transform the dialogue into a simple code-like format. Formally, given one dialogue d $\in D$, we denote M(.) as the structured format verbalizer, and the original graph can be mapped into a sequence as $C_i = M(d)$. For each utterance in the dialogue, we assigned it a sequence identifier <u> indicating its position in the dialogue. For the fundamental format, all utterances are listed as a sequence with entity_list, while all reply dependencies are listed as a sequence with variable triple_list. The specific example is shown in Figure 3.

After obtaining the structured textual representation of dialogue reply dependencies, we decompose the tasks into two different graph-centric instruction tasks: element extraction tasks $\mathcal{E}$ and dialogue reply dependency analysis task $\mathcal{R}$. $\mathcal{E}$ corresponds to the extraction of three single elements and tuple extraction (i.e., pair extraction and quadruple extraction) in Figure 2. Additionally, in the quadruple extraction task, we prompt the model to first determine whether each utterance is a non-opinion. For the $\mathcal{E}$, both the dialogue and its structured textual representation are provided as inputs to help the LLM better utilize the dialogue reply dependency information. For the $\mathcal{R}$, only the dialogue is given as input, while the structured textual representation of the dialogue reply dependencies is used as the target output. This aims to enhance the LLM's ability to analyze the structure of the dialogue. Finally, given one dialogue $d \in D$, the LLM can be optimized by maximum likelihood with:

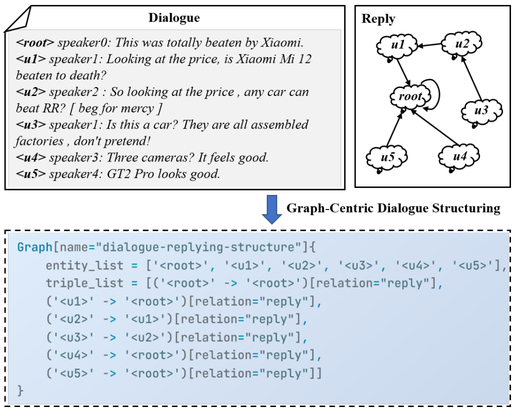


Figure 3: A specific sample to illustrate the transformation process of the Graph-Centric Dialogue Structuring strategy.

$$\mathcal{L}(\mathcal{T}_j) = -\sum_{i=1}^{N_j} \log \pi_\theta(\mathcal{Y}_i = \mathcal{A}_i | \mathcal{X}_i), \quad (1)$$

where π_θ denotes the LLM with trainable parameters θ, $\mathcal{Y}$ is the model output, $\mathcal{X}$ and $\mathcal{A}$ respectively represent the input sequence and reference label, which depends on the specific task definition.

3.2 Task-oriented Mixture of Experts Layer

Existing studies demonstrate that task-related information is helpful for improving model performance (Liu et al., 2024; Tian et al., 2024). We assume that there is task-shared knowledge among element extraction tasks and dialogue reply dependency analysis task, and by learning this knowledge, the model can achieve better performance in each task. To learn task-shared knowledge better, we replace each dense layer in the LLM with a task-oriented mixture of experts (MoE) layer.

In the task-oriented MoE layer, every expert can be denoted as $\{E_i\}_{i=1}^{N}$ and is constructed as two decomposed low-rank matrices, where N denotes the number of experts. For the samples from task $\mathcal{T}_j \in \{\mathcal{E}, \mathcal{R}\}$, the output of intermediate LLM layers can be expressed as during the forward process of a linear layer paired with the task-oriented MoE layer, Specifically, each task is assigned a unique task identifier token. Then the task identifier token is fed into the task-motivated gate network. Upon identifying a task $\mathcal{T}_j$, we extract the j-th column of E, which serves as the representation vector for that task, symbolized as $e_j \in \mathbb{R}^{d_\mathcal{T}}$, where $\mathbb{R}^{d_\mathcal{T}}$ represents the dimension of the task embedding. Additionally, a linear transformation is applied to determine the contribution weights for task $\mathcal{T}_j$. This calculation is represented by the following equation:

$$\omega_j = \mathrm{Softmax}(W^{\mathcal{T}} e_j), \tag{2}$$

where $\omega_j \in \mathbb{R}^N$ represents the contribution weight vector tailored for task $\mathcal{T}_j$. The transformation matrix is denoted as $W_\mathcal{T} \in \mathbb{R}^{N \times d_\mathcal{T}}$. To avoid excessively large weights, a softmax operation is leveraged to normalize the contribution weights. Based on this structure, the forward process of a linear layer paired with a task-oriented MoE layer for samples from task $\mathcal{T}_j$ is expressed as:

$$\begin{aligned} \mathbf{h}_j &= \mathbf{W}_0 \mathbf{x}_j + \frac{\alpha}{r} \cdot \sum_{i=1}^{N} \omega_{ji} \cdot E_i(\mathbf{x}_j) \\ &= \mathbf{W}_0 \mathbf{x}_j + \frac{\alpha}{r} \cdot \sum_{i=1}^{N} \omega_{ji} \cdot \mathbf{B}_i \mathbf{A}_i \mathbf{x}_j, \end{aligned} \tag{3}$$

where $\mathbf{h}_j$ and $\mathbf{x}_j$ represent the input and output of intermediate LLM layers for samples from $\mathcal{T}_j$. The matrices $\mathbf{B}_i \in \mathbb{R}^{d_{in} \times \frac{r}{N}}$ and $\mathbf{A}_i \in \mathbb{R}^{\frac{r}{N} \times d_{out}}$ form the expert E_i. The hyper-parameter N denotes the number of experts in MOELoRA, and for each expert, the rank of matrices $\mathbf{A}$ and $\mathbf{B}$ is $\frac{r}{N}$.

3.3 Expert-Level Contrastive Learning

In the task-oriented mixture of experts layer, we aim to reduce feature redundancy between tasks and allow experts to focus on handling distinct task characteristics, thereby improving the overall efficiency of the model. To enhance expert differentiation and representation learning, we incorporate contrastive learning into the mixture of experts layer. Inspired by previous work (He et al., 2020; Luo et al., 2024a), our approach encourages representations of inter-expert to be more discriminative while maintaining intra-expert consistency.

Given a input sample x, let $E(x) = \{E_1(x), \ldots, E_n(x)\}$ denote the set of expert outputs, where $E_i(x) \in \mathbb{R}^{L \times D}$, L is the sequence length activated by E_i and D is the hidden dimension. We first compute the gating activation for each expert via element-wise product:

$$\mathbf{G} = \mathrm{MeanPool}(E(x)) \odot \omega_j, \tag{4}$$

where $\omega_j \in \mathbb{R}^N$ represents the contribution weight vector same as in Equation 2. Then, we construct a binary mask to select activated tokens per expert using: $\mathbf{M} = (\mathbf{G} > \epsilon)$, where ϵ denotes the threshold. Each token's expert representation is then L2-normalized: $\hat{\mathbf{E}(\mathbf{x})} = \frac{\mathbf{E}(\mathbf{x})}{\|\mathbf{E}(\mathbf{x})\|_2}$ to ensure numerical stability in contrastive similarity computations.

In terms of the contrastive pair construction, the outputs of the same expert are treated as positive samples, while the outputs of different experts are considered negative samples. We define the binary mask matrix $\mathbf{P} \in \{0,1\}^{N \times L \times L}$ as:

$$P_{q,k} = \begin{cases} P_{q,k^+}, & \text{if } q,k \text{ belong to the same expert} \\ P_{q,k^-}, & \text{otherwise} \end{cases} \tag{5}$$

To construct the similarity matrix and stabilize training and prevent numerical overflow, we compute:

$$\hat{\mathbf{S}} = \exp\left(\frac{\mathbf{S}}{\tau}\right), \mathbf{S} = \hat{\mathbf{E}(\mathbf{x})} \cdot \hat{\mathbf{E}(\mathbf{x})}^{\top}, \tag{6}$$

where τ represents the temperature coefficient. To compute the final contrastive probability distribution, we normalize the similarity scores within each row:

$$p_{q,(k^+,k^-)} = \frac{\hat{\mathbf{S}}_{q,k^+} \cdot P_{q,k^+}}{\sum_{k^-} \hat{\mathbf{S}}_{q,k^-} \cdot P_{q,k^-}}, \tag{7}$$

the contrastive loss is then formulated as:

$$\mathcal{L}_{\text{contrastive}} = -\sum_{q \neq k_+} \log(p_{q,(k^+,k^-)}). \tag{8}$$

This contrastive loss forces representations of tokens assigned to the same expert to be close in the learned space while separating representations assigned to different experts. The final training objective is a combination of the contrastive loss and the objective function for multi-task fine-tuning:

$$\mathcal{L} = \mathcal{L}(\mathcal{T}_j) + \lambda \mathcal{L}_{\text{contrastive}}, \tag{9}$$

where λ is a hyperparameter controlling the trade-off between the primary extraction task and contrastive expert learning.

4 Experimental Settings

4.1 Dataset

We evaluate TaCoMoE using the DiaASQ dataset (Li et al., 2023a), the first multilingual dataset designed for dialogue-level aspect-based sentiment analysis. The raw data is sourced from the largest Chinese social media platform, comprising 1,000 dialogues available in both Chinese and English. Specifically, the dataset features multipart, multi-turn conversations centered primarily on mobile phone-related topics. More detail is in Appendix B.

4.2 Comparison Methods

SpERT (Eberts and Ulges, 2019) features entity recognition and filtering, as well as relation classification with a context representation.
CRFExtract (Cai et al., 2021) adapts one of the representative aspect-opinion co-extraction system.
ParaPhrase (Zhang et al., 2021a) reveals a more comprehensive and complete aspect-level sentiment structure.
Span-ASTE (Xu et al., 2021) considers the interaction between the whole spans of targets and opinions when predicting their sentiment relation.
Meta-WP (Li et al., 2023a) manages to incorporate rich dialogue-specific and discourse feature representations.
SADD (Luo et al., 2024b) proposes a multi-granularity denoising generation model for denoising and a distribution-based solution for debiasing.

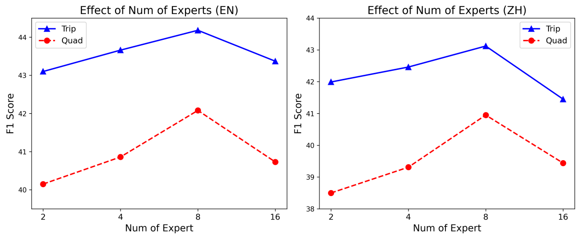


Figure 4: The results of experiments for expert number.

DMIN (Huang et al., 2024) enhances utterance interactions at the token level and introduces a novel integrator to address the challenge of data integration.
H2DT (Li et al., 2024) leverages unified discourse features and triadic interaction for dialogue sentiment quadruple extraction.
ChatGPT4 (OpenAI, 2023) is a large language model developed by OpenAI, capable of understanding and generating human-like text across diverse tasks and domains.
ChatGLM (GLM et al., 2024) is an open-source, bilingual large language model, designed for dialogue and general-purpose language understanding tasks.

4.3 Evaluation Metrics

Following previous work (Li et al., 2023a, 2024), we mainly measure the performances in terms of four angles: span match (i.e., Target, Aspect, and Opinion), pair extraction (i.e., Target-Aspect, Aspect-Opinion, and Target-Opinion), triplet detection (i.e., Target-Aspect-Opinion), quadruple extraction (i.e., Target-Aspect-Opinion-Sentiment), and non-opinion detection through precision, recall, and F1 score metrics.

4.4 Implementation Details

TaCoMoE uses ChatGLM3-6B[1] as the robust backbone model comprising 28 transformer layers, which are implemented in the Huggingface Transformers library (Wolf et al., 2020) and utilizes low rank adaptation (LoRA) (Hu et al., 2021) to perform parameter-efficient learning with rank = 16 and set the rank of each expert to 2. Specifically, we conduct dedicated experiments to investigate the impact of the number of experts on quadruple extraction performance. As shown in the experimental results in the Figure 4, we observe that the model achieved the best score when the number of experts is set to 8. Therefore, we ultimately set the

[1] https://huggingface.co/THUDM/chatglm3-6b

Data	Methods	Entity (F1)			Pair (F1)			Triplet			Quadruple		
		T	A	O	T-A	T-O	A-O	P	R	F	P	R	F
ZH	CRF-Extract	91.11	75.24	50.06	32.47	26.78	18.90	/	/	9.25	/	/	8.81
	SpERT	90.69	76.81	54.06	38.05	31.28	29.05	/	/	14.19	/	/	13.00
	ParaPhrase	/	/	/	37.81	34.32	27.76	/	/	27.98	/	/	23.27
	Span-ASTE	/	/	/	44.13	33.42	32.21	/	/	30.85	/	/	27.42
	Meta-WP	90.23	76.94	59.35	48.61	43.31	45.44	/	/	37.51	/	/	34.94
	SADD	/	/	/	51.13	46.72	47.87	/	/	41.05	/	/	37.80
	DMIN	/	/	/	**57.62**	51.65	**56.16**	/	/	**47.50**	/	/	**44.49**
	H2DT	**91.72**	76.93	61.87	50.48	48.39	52.40	45.40	40.50	42.81	**42.78**	38.17	40.34
	LLM-based												
	ChatGPT4$_{4\text{-shot}}$	36.89	37.69	39.03	19.72	18.78	22.85	11.59	14.50	12.88	10.67	13.36	11.86
	ChatGLM3$_{\text{LoRA}}$	68.06	65.73	47.88	46.86	32.28	36.89	20.48	28.12	23.70	20.48	28.12	23.70
	ChatGLM3$_{\text{KTO}}$	68.84	64.42	48.17	46.49	33.12	37.44	30.81	26.97	28.77	27.91	24.68	26.20
	TaCoMoE	91.18	**81.48**	**64.63**	55.85	**52.48**	52.55	**45.87**	**42.49**	44.12	42.58	**39.44**	40.95
EN	CRF-Extract	88.31	71.71	47.90	34.31	21.90	19.21	/	/	12.80	/	/	11.59
	SpERT	87.82	74.65	54.17	28.33	23.64	23.64	/	/	13.38	/	/	13.07
	ParaPhrase	/	/	/	37.22	32.19	30.78	/	/	26.76	/	/	24.54
	Span-ASTE	/	/	/	42.19	30.44	45.90	/	/	28.34	/	/	26.99
	Meta-WP	88.62	74.71	60.22	47.91	45.58	44.27	/	/	36.80	/	/	33.31
	SADD	/	/	/	50.82	49.64	49.70	/	/	43.32	/	/	38.87
	DMIN	/	/	/	53.49	52.66	52.09	/	/	42.31	/	/	39.22
	H2DT	88.69	73.81	62.61	48.69	48.84	52.47	**44.36**	40.23	42.19	41.01	37.20	39.01
	LLM-based												
	ChatGPT4$_{4\text{-shot}}$	47.63	29.07	37.17	22.72	27.40	18.45	12.55	20.18	15.48	11.61	18.77	14.34
	ChatGLM3$_{\text{LoRA}}$	70.76	61.99	52.25	46.92	41.07	40.33	33.01	31.09	32.02	29.39	27.80	28.57
	ChatGLM3$_{\text{KTO}}$	73.28	61.39	53.57	47.04	42.69	41.35	35.82	32.71	34.19	31.62	28.94	30.22
	TaCoMoE	**91.04**	**77.02**	**63.13**	**54.53**	**52.86**	**53.71**	44.09	**44.27**	**44.18**	**41.99**	**42.16**	**42.08**

Table 1: Performance (%) evaluation metrics for entity, pair, triplet, and quadruple extraction in both ZH (Chinese) and EN (English) datasets. The best results are highlighted in **bold** and the second best results are underlined. '/' means that the results are unavailable from the original paper. The results of all LLM-based methods are derived from experiments conducted using self-constructed instruction data.

number of experts to 8. The optimizer is AdamW (Loshchilov and Hutter, 2017) in all stages with initial learning rates of 2e-4. The maximum length is set as 2048 and batch size is set to 16. The TaCoMoE is trained on 4×24G NVIDIA RTX4090 GPUs. For all experiments, we report the results as the average over three runs with different random seeds.

5 Results and Discussions

5.1 Comparison with Baseline Models

The overall performance of all the compared baselines and proposed TaCoMoE on the DiaASQ dataset is presented in Table 1.

Item Extraction We observe that our method outperforms all previous models on the item detection task for both datasets. This is attributed to the fact that our method, in contrast to previous approaches, adopts a multi-task framework and incorporates the single-element extraction task. On the English dataset, our method achieves improvements of 2.35%, 2.31%, and 0.52% over the previous state-of-the-art for the three sub-element extraction tasks, respectively. On the Chinese dataset, TaCoMoE achieves marked improvements of 4.54% and 2.76% on the aspect and opinion extraction.

Pair Extraction TaCoMoE achieves improvements on all metrics in pair extraction compared with SADD and H2DT, indicating that it has excellent ability in pairing binary relationships. In terms of the English dataset, significant improvements are observed in the T-A and A-O pair detection, with gains of 1.04% and 1.24% in F1 scores, respectively. The T-O pair detection also demonstrates a smaller improvement of 0.20%. In terms of the Chinese dataset, the T-O pair detection showcases improvements of 0.83% in F1 score.

Triplet and Quadruple Extraction Regarding triplet extraction (i.e., Identification F1), TaCoMoE surpasses DMIN and SADD by 1.87% and 0.86% on English dataset, demonstrating the superiority of our proposed method in entity extraction and triplet correspondence. In the quadruple extraction task, TaCoMoE consistently obtains the best micro F1 score over comparison methods. Specifically, TaCoMoE obtains 2.86% absolute improvements on English dataset. Experimental results demonstrate that TaCoMoE achieves the new state-of-the-art performances on English dataset.

Discussion on Suboptimal Performance We observe that, compared to DMIN, TaCoMoE yields slightly lower results on the Chinese dataset. The main reason is that DMIN employs different backbone models for Chinese and English datasets; specifically, it utilizes a customized pre-trained language model optimized for Chinese on the Chinese dataset, which already exhibits strong performance in tuple extraction tasks. In contrast, our method uses a large language model as the backbone, which, as shown in the experimental results, performs relatively poorly in quad-tuple extraction tasks—even after supervised fine-tuning, it cannot match previous state-of-the-art results. Nevertheless, TaCoMoE achieves competitive performance while jointly handling data in two different languages, which we believe makes it a fair and meaningful comparison to DMIN. Naturally, our approach is applicable to large language models of varying scales, and we plan to conduct further investigations in this direction.

Compared to LLM-based Methods In addition to the comparisons with the aforementioned SOTA results, we also observe that our method demonstrates superior efficiency when compared with LLM-based approaches. It consistently outperforms the compared supervised fine-tuning and reinforcement learning methods in both item extraction and tuple extraction tasks.

5.2 Ablation Study

In this section, we perform ablation studies to analyze the effects of critical modules in our TaCoMoE, detailed in Table 2.

Effects of Contrastive Learning To study the effect of contrastive learning, we remove the $\mathcal{L}_{Con}$. Experimental results show that the performances of TaCoMoE$_{\text{w/o}\mathcal{L}_{Con}}$ decrease in all metrics on both English and Chinese datasets. The performances on both datasets prove the effectiveness of expert-level contrastive learning. The visual demonstration of the further analysis comparing the impact of contrastive loss on the distribution of expert outputs in the semantic space is provided in Appendix C.1.

Effects of Non-opinion Detection To analyze the impact of non-opinion detection (NOD), we ignore the identification of utterances that do not contain opinions during the fine-tuning process and focus solely on quadruple extraction. As shown in 2, the performances of TaCoMOE$_{\text{w/o NOD}}$ fall sharply in all metrics. Taking the English dataset as an example, the model's performance on triplet and quadruple extraction decreased by 10.17% and 9.32%, respectively. The results prove the importance and superiority of considering non-opinion detection detection. A more detailed comparison with other LLM-based methods will be presented in Section 5.3.

Methods	Chinese (F1)		English (F1)	
	Trip.	Quad.	Trip.	Quad.
TaCoMoE	**43.12**	**40.95**	**44.18**	**42.08**
w/o $\mathcal{L}_{Con}$	$40.75_{\downarrow 2.37}$	$38.66_{\downarrow 2.29}$	$42.57_{\downarrow 1.61}$	$39.87_{\downarrow 2.21}$
w/o NOD	$31.16_{\downarrow 11.96}$	$29.66_{\downarrow 11.29}$	$34.01_{\downarrow 10.17}$	$32.76_{\downarrow 9.32}$
w/o GCDS	$40.30_{\downarrow 2.82}$	$38.51_{\downarrow 2.44}$	$41.59_{\downarrow 2.54}$	$40.00_{\downarrow 2.08}$
- w/o Structure	$41.51_{\downarrow 1.61}$	$39.63_{\downarrow 1.32}$	$42.64_{\downarrow 1.54}$	$40.82_{\downarrow 1.26}$
- w/o $\mathcal{T}^{Reply}$	$41.67_{\downarrow 1.45}$	$39.12_{\downarrow 1.83}$	$43.05_{\downarrow 1.13}$	$41.03_{\downarrow 1.05}$

Table 2: Performance (%) comparison on Chinese and English datasets (F1 score).

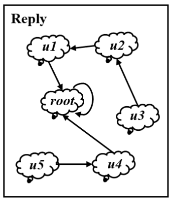
Effects of Graph-Centric Dialogue Structuring Since we utilize Graph-Centric Dialogue Structuring strategy in both the task of dialogue reply relationship analysis and the dialogue input, we implement three variants: TaCoMoE$_{\text{w/o}\mathcal{T}^{Reply}}$, TaCoMoE$_{\text{w/o Structure}}$, and TaCoMoE$_{\text{w/o GCDS}}$. These three variants respectively represent the removal of the dialogue reply relationship analysis task, the exclusion of the reply relationship, and the elimination of both the dialogue reply relationship analysis task and the reply relationship. Experimental results demonstrate that the performances of these three variants drop considerably on both English and Chinese datasets. The experimental results of our further validation of the GCDS strategy in understanding context and leveraging reply relationships are detailed in the Appendix C.3.

5.3 Analysis of Non-opinion Detection

To rigorously investigate the contribution of non-opinion detection, we conduct experiments in two settings: training without non-opinion detection (w/o NOD) and with non-opinion detection (w NOD). The results are displayed in Table 3.

Since there has been no prior work specifically analyzing non-opinion utterances in the DiaASQ task, we conduct comparative experiments with ChatGPT-4$_{\text{4shot}}$ and ChatGLM3$_{\text{LoRA}}$ (GLM et al., 2024). Examples of instruction templates for few-shot and fine-tuning can be found in the Appendix D. It is evident that TaCoMoE achieves results that far exceed those of the other two methods, regardless of whether non-opinion detection is performed. For intra-method, we find that the fine-tuned method performs better when considering non-opinion detection compared to not consider-

<root>speaker0: The positioning of 12p seems to be quite embarrassing, I saw it is recommended to either 12 or 12 pm
<u1>speaker1: That's for sure... The difference between pm and p is only 800, better battery life, better photography, bigger screen...
<u2>speaker0: I originally wanted to buy the size of a pro, but its positioning, I feel like I have to give up weight for the camera
<u3>speaker2: Hahaha if you are interested in photography, then pm, I don't pay attention to this aspect mainly for convenience
<u4>speaker2: Yes, I found a lot of people say this, but the PM is really too big and my hands are small. Now I'm in a dilemma
<u5>speaker0: Me too, the kind with small hands, I want to buy PM for the camera, but it will really be too heavy like a brick



ID	TaCoMoE		w/o NOD		Ground Truth
<root>	(12p, positioning, quite embarrassing, neg)	✓	(12p, positioning, quite embarrassing, neg)	✓	(12p, positioning, quite embarrassing, neg)
<u1>	(pm, battery life, better, pos) (pm, photography, better, pos) (pm, screen, bigger, pos)	✓	(pm, battery life, better, pos) (pm, photography, better, pos) (pm, screen, bigger, pos)	✓	(pm, battery life, better, pos) (pm, photography, better, pos) (pm, screen, bigger, pos)
<u2>	statement-non-opinion	✓	(pro, weight, give up, neg) (pro, camera, give up, neg)	✗	statement-non-opinion
<u3>	(pm, photography, interested, pos) ✓ (12p, Not mentioned, don't pay attention, neg) ✗		(pm, photography, don't pay attention, neu) (pm, Not mentioned, convenience, pos)	✗	(pm, photography, interested, pos)
<u4>	statement-non-opinion	✓	(PM, Not mentioned, too big, neg)	✗	statement-non-opinion
<u5>	(PM, camera, too heavy like a brick, neg)	✗	(PM, camera, too heavy like a brick, neg)	✗	(PM, camera, want to buy, pos)

Figure 5: Case study. The primary target, aspect, and opinion in the dialogue are highlighted in different colors.

ing it. Additionally, after performing non-opinion detection, the model shows a more significant improvement in handling both quadruple and non-opinion utterances. This indicates that the model is better able to distinguish whether utterances contain opinions, thereby achieving improved results in quadruple extraction.

Train	Methods	With-O		With-O + Non-O	
		Trip.	Quad.	Trip.	Quad.
EN					
w/o NOD	ChatGPT4$_{4shot}$	23.70	22.47	18.09	17.14
	ChatGLM3$_{LoRA}$	32.48	30.39	25.93	23.86
	TaCoMoE	43.47	41.88	34.01	32.76
	$\triangle$ (TaCoMoE)	10.99↑	11.49↑	8.08↑	8.90↑
w NOD	ChatGPT4$_{4shot}$	19.04	17.71	15.48	14.34
	ChatGLM3$_{LoRA}$	33.40	30.77	30.58	28.61
	TaCoMoE	46.12	43.93	44.18	42.08
	$\triangle$ (TaCoMoE)	12.72↑	13.16↑	13.60↑	13.47↑
ZH					
w/o NOD	ChatGPT4$_{4shot}$	18.99	17.59	15.06	13.94
	ChatGLM3$_{LoRA}$	29.04	27.19	22.84	21.39
	TaCoMoE	38.84	37.59	31.16	29.66
	$\triangle$ (TaCoMoE)	9.80↑	10.40↑	8.32↑	8.27↑
w NOD	ChatGPT4$_{4shot}$	14.89	13.72	12.88	11.86
	ChatGLM3$_{LoRA}$	29.14	27.30	27.95	26.20
	TaCoMoE	44.94	41.74	43.12	40.95
	$\triangle$ (TaCoMoE)	15.80↑	14.44↑	15.17↑	14.75↑

Table 3: Performance (%) comparison of different methods in w NOD and w/o NOD scenarios. With-O refers to utterances that contain opinions, while Non-O refers to utterances that do not contain opinions.

5.4 Case Study

To better understand how non-opinion detection affects the quadruple extraction results, we present a specific case in Figure 5.

Intuitively, we can observe that when considering non-opinion detection, our method correctly identifies *<u2>* and *<u4>* as "statement-non-opinion." In contrast, the model without performing non-opinion detection incorrectly extracts quadruples from these utterances. Actually, taking the *<u4>* as an example, it describes a dilemma in making a choice rather than explicitly expressing sentiment toward a specific Target-Aspect. Aside from this, we also observe that models that do not handle non-opinion cases tend to more easily misinterpret the speaker's opinion, leading to incorrect extraction of the final quadruples. Taking *<u3>* as an example, TaCoMoE correctly identifies the quadruples in the sentence but additionally extracts an incorrect quadruple, whereas TaCoMOE$_{w/o\ NOD}$ incorrectly identifies two quadruples. In this utterance, 'pay attention' and 'convenience' do not refer to any product, but rather express the speaker's attitude.

6 Conclusion

In this paper, we propose an LLM-based approach that integrates contrastive learning to the task-oriented mixture of experts. Additionally, we define non-opinion utterances that contain no opinion associated with targets or aspects and incorporate non-opinion detection. For modeling dialogue response relations, we employ a Graph-Centric Dialogue Structuring strategy, enabling the LLM to understand dialogue reply structure. Experimental results and analyses illustrate the effectiveness of our proposed TaCoMoE.

7 Limitations

Although the proposed TaCoMoE achieves state-of-the-art results on the DiaASQ task, our approach still has its own limitations. Firstly, we use contrastive learning in the mixture-of-experts layer and treat the experts' outputs on activated tokens as positive and negative sample pairs, which increases training time. Secondly, the effectiveness of our proposed Graph-Centric Dialogue Structuring strategy has not yet been validated on other tasks, and although it does not require an additional graph encoder, it increases the context length, leading to higher memory usage. Lastly, we have preliminarily explored the contribution of non-opinion utterances to the DiaASQ task, but how to more effectively distinguish whether utterances contain opinions or their opinions refer to any specific target or aspect remains to be further investigated.

References

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, Online. Association for Computational Linguistics.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Guimin Chen, Yuanhe Tian, and Yan Song. 2020a. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 272–279, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985, Dublin, Ireland. Association for Computational Linguistics.

Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang. 2020b. Towards interpretable clinical diagnosis with Bayesian network ensembles stacked on entity-aware CNNs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3143–3153, Online. Association for Computational Linguistics.

Shuang Chen, Yang Liu, and Erik Cambria. 2023. Knowledge-enhanced aspect-based sentiment analysis with external commonsense graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 13567–13579. Association for Computational Linguistics.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Bidirectional generative framework for cross-domain aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12272–12285, Toronto, Canada. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. In *European Conference on Artificial Intelligence*.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Shigang Hu, Akshi Kumar, Fadi Al-Turjman, Shivam Gupta, Simran Seth, and Shubham. 2020. Reviewer credibility and sentiment analysis based user profile modelling for online product recommendation. *IEEE Access*, 8:26172–26189.

Peijie Huang, Xisheng Xiao, Yuhong Xu, and Jiawei Chen. 2024. Dmin: A discourse-specific multi-granularity integration network for conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 16326–16338. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023a. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13449–13467, Toronto, Canada. Association for Computational Linguistics.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *AAAI Conference on Artificial Intelligence*.

Wei Li, Luyao Zhu, Rui Mao, and E. Cambria. 2023b. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *AAAI Conference on Artificial Intelligence*.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2018a. A unified model for opinion target extraction and target sentiment prediction. In *AAAI Conference on Artificial Intelligence*.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018b. Aspect term extraction with history attention and selective transformation. In *International Joint Conference on Artificial Intelligence*.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2022. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 10594–10609. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *SIGIR '24 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114. Association for Computing Machinery, Inc.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024a. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *ArXiv*, abs/2402.12851.

Xianlong Luo, Meng Yang, and Yihao Wang. 2024b. Overcome noise and bias: Segmentation-aided multi-granularity denoising and debiasing for enhanced quarduples extraction in dialogue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 839–856, Miami, Florida, USA. Association for Computational Linguistics.

Yanying Mao, Qun Liu, and Yu Zhang. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, 36(4):102048.

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.

Erinc Merdivan, Deepika Singh, Sten Hanke, and Andreas Holzinger. 2019. Dialogue systems for intelligent human computer interactions. *Electronic Notes in Theoretical Computer Science*, 343:57–71. The proceedings of AmI, the 2018 European Conference on Ambient Intelligence.

Sandeep Sricharan Mukku, Manan Soni, Chetan Aggarwal, Jitenkumar Rana, Promod Yenigalla, Rashmi Patange, and Shyam Mohan. 2023. InsightNet : Structured insight mining from customer feedback. In *Proceedings of the 2023 Conference on Empirical*

Methods in Natural Language Processing: Industry Track, pages 552–566, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *CoRR*, abs/2303.08774.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2019. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *AAAI Conference on Artificial Intelligence*.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *International Joint Conference on Artificial Intelligence*.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 720–736, Mexico City, Mexico. Association for Computational Linguistics.

Tobias Schröder, Terrence Stewart, and Paul Thagard. 2013. Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive science*, 38.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7007–7014.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7143–7155, Bangkok, Thailand. Association for Computational Linguistics.

Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, and Jinho D. Choi. 2024. Automating PTSD diagnostics in clinical interviews: Leveraging large language models for trauma assessments. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 644–663, Kyoto, Japan. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction tuning for few-shot aspect-based sentiment analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Christos Vlachos, Themos Stafylakis, and Ion Androutsopoulos. 2024. Comparing data augmentation methods for end-to-end task-oriented dialog systems. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7216–7240, Bangkok, Thailand. Association for Computational Linguistics.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. 2024. InstructGraph: Boosting large language models via graph-centric instruction tuning and preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13492–13510, Bangkok, Thailand. Association for Computational Linguistics.

Qianlong Wang, Keyang Ding, Bin Liang, Min Yang, and Ruifeng Xu. 2023. Reducing spurious correlations in aspect-based sentiment analysis with explanation from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2930–2941, Singapore. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 606–615.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. 2021. Learn from syntax: Improving pair-wise aspect and opinion terms extractionwith rich syntactic knowledge. In *International Joint Conference on Artificial Intelligence*.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. 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 4755–4766, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework 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)*, pages 2416–2429, Online. Association for Computational Linguistics.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. DualGATs: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408, Toronto, Canada. Association for Computational Linguistics.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024. MIntrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021b. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. GreaseLM: Graph REASoning enhanced language models. In *International Conference on Learning Representations*.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Related Work

Aspect-Based Sentiment Analysis (ABSA), a subfield of sentiment analysis (Liu, 2012; Pontiki et al., 2014; Wang et al., 2016), initially focused on extracting single elements (e.g. target, aspect terms, categories, and opinion terms) (Li et al., 2018a,b; Peng et al., 2019) and subsequent research shifting towards multi-pair extraction (e.g. aspect-opinion pair extraction, aspect sentiment term extraction, and aspect sentiment quadruple extraction) (Wu et al., 2021; Chen et al., 2022; Mao et al., 2022). Early research primarily targeted short, unstructured plain texts, and ABSA has now become a pivotal research area in the field of affective computing (Li et al., 2022; Chen et al., 2023).

Conversational Aspect-based Sentiment Quadruple Analysis (DiaASQ) is a new sub-task of ABSA with complex textual content and structures. Li et al. (2023a) design the multi-view interaction layer and fuse rotary position embedding (RoPE) to model the dialogue utterance interactions. Li et al. (2024) introduce a token-level heterogeneous graph to model the complexities of speaker roles and reply relationships, enhancing the understanding of dialogue features. Luo et al. (2024b) propose segmentation-aided order bias mitigation model to simultaneously address both the one-to-many training challenge and the order bias.

Discourse Structure intuitively enhances the model's ability to encode unstructured human conversations more effectively, enabling it to focus on key utterances and achieve more accurate dialogue quadruple extraction and sentiment prediction. Deep sequential models are regarded as practical approaches for conversational discourse parsing (Shi and Huang, 2019; Liu and Chen, 2021). More recently, Peng et al. (2022) introduce a global-to-local hierarchical graph network to model hierarchical discourse structures in dialogues. Li et al. (2023b) employ relational graph convolutional networks (RGCN) as the base graph network to encode the discourse structure as the symbolic knowledge. Zhang et al. (2023) propose DisGAT to integrate discourse structural information, which is built upon graph attention networks (GAT). **Non-opinion Utterances** The meaning and purpose of an utterance are influenced by specific contexts or dialogue history (Schröder et al., 2013). In the DiaASQ task, opinions are often closely linked to sentiment polarity. If an utterance does not contain an opinion or the opinion expressed fails to refer to any specific target or aspect, then it is also impossible to determine a clear sentiment or extract a complete quadruple from that utterance. In an earlier study on dialogue, Godfrey et al. (1992) introduce 42 types of dialogue acts, including statements that primarily convey factual information,

Dataset		Dialogue		Items			Pairs			Quadruples	
		Dia.	Utt.	Tgt.	Asp.	Opi.	T-A	T-O	A-O	Intra.	Cross.
EN	train	800	5,947	6,613	5,109	5,523	4,699	5,931	3,989	3,442	972
	valid	100	748	822	644	719	603	750	509	423	132
	test	100	757	829	681	592	592	751	496	422	123
ZH	train	800	5,947	6,652	5,220	5,622	4,823	6,062	4,297	3,594	1,013
	valid	100	748	823	662	764	621	758	538	440	137
	test	100	757	833	690	705	597	767	523	433	125

Table 4: The statistics of experimental datasets. 'Dia.' and 'Utt.' refer to dialogue and utterance, respectively. 'Tgt', 'Asp', and 'Opi' refer to target, aspect, and opinion terms, respectively. 'Intra' and 'Cross' refer to the intra-/cross utterance quadruples.

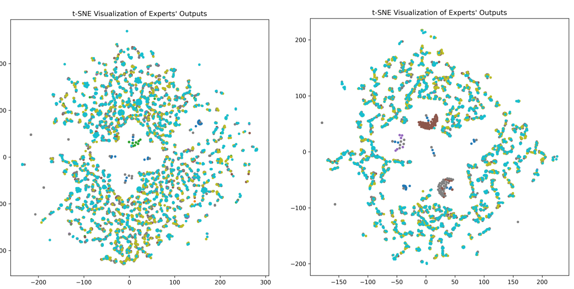
which are defined as *statement-non-opinion*. Given the uncertainty in defining the boundary for identifying out-of-scope utterances, Larson et al. (2019) define them as those that do not belong to any of the existing intent classes and Zhang et al. (2024) adopt this definition in a recent study about intent recognition. Inspired by the aforementioned work, we believe that considering non-opinion utterances better aligns with real-world scenarios and practical applications. In this paper, we define *statement-non-opinion* utterances as those that **do not contain extractable opinions or their opinions do not refer to any specific target or aspect**.

B Dataset Statistics

The statistics of DiaASQ dataset are reported in Table 4. The dataset is divided into train/test/dev sets in an 8:1:1 ratio. Also, there is an average of one sentimental expression in each utterance.

C In-depth Analysis

C.1 Experts Representation Visualization



(a) w/o $\mathcal{L}_{Con}$ (b) TaCoMoE

Figure 6: t-SNE visualization of representations learned by each expert. Each color represents the output of a specific expert, each point represents a token's 2D projection after t-SNE dimensionality reduction, and the distribution of points reflects the division of labor among experts.

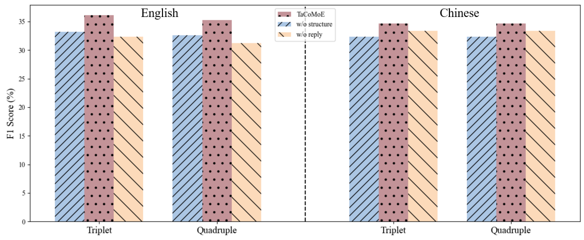


Figure 7: Triplet and quadruple extraction scores on cross-utterance instances. The term 'w/o reply' and 'w/o structure' denotes the TaCoMoE$_{\text{w/o}\mathcal{T}^{Reply}}$, TaCoMoE$_{\text{w/o Structure}}$.

We qualitatively visualize the learned representations of the experts with t-SNE (van der Maaten and Hinton, 2008). Figure 6 shows the visualization of the samples from different tasks. Compared with not using contrastive objective, the distribution of each expert representation learned by our TaCoMoE is more tight and united. It indicates that, under TaCoMoE, the outputs of the same expert are closer, enhancing the expert's focus on specific tasks. The outputs of different experts are farther apart, helping the model allocate resources more effectively in multi-task learning, promoting clear division of labor, and reducing interference between tasks.

C.2 Experiment Result in Cross-utterance

To further analyze our proposed Graph-Centric Dialogue Structuring strategy, we compare the performance of TaCoMoE, TaCoMoE$_{\text{w/o}\mathcal{T}^{Reply}}$, and TaCoMoE$_{\text{w/o Structure}}$ on cross-utterance quadruples as demonstrated in Figure 7.

Cross-utterance quadruple refers to the elements of the quadruples potentially coming from different utterances. The comparison results show that removing either the task or the reply relationships leads to a noticeable decrease in the model's per-

Data	Methods	Entity (F1)			Pair (F1)			Triplet			Quadruple		
		T	A	O	T-A	T-O	A-O	P	R	F	P	R	F
ZH	w Linear Textual	89.54	78.84	61.42	48.79	49.83	48.70	38.34	39.84	39.08	36.39	37.82	37.09
	w GCDS	91.18	81.48	64.63	55.85	52.48	52.55	45.87	42.49	44.12	42.58	39.44	40.95
EN	w Linear Textual	90.06	77.51	57.64	51.50	49.56	50.95	40.99	42.00	41.49	38.58	39.53	39.05
	w GCDS	91.04	77.02	63.13	54.53	52.86	53.71	44.09	44.27	44.18	41.99	42.16	42.08

Table 5: Performance (%) evaluation metrics for entity, pair, triplet, and quadruple extraction in both ZH (Chinese) and EN (English) datasets.

formance on extracting cross-utterance quadruples. As such, TaCoMoE, enhanced with the GCDS strategy, shows a marginal but discernible improvement in the extraction of cross-utterance quadruples on both Chinese and English datasets. Combining the experimental results mentioned above with those presented in Section 5.2 underscores the superiority and robustness of the proposed GCDS strategy.

C.3 Compared with Linear Textual Description

In this section, we compare our proposed Graph-Centric Dialogue Structuring strategy with a simple linear textual description. The results are shown in the Table 5. The results show that the proposed GCDS strategy outperforms the simple linear prompts in single-element extraction, pairwise tuple extraction, triplet, and quadruplet extraction tasks, demonstrating the effectiveness of the strategy.

D Instruction

In this section, we provide examples of instruction templates for conducting few-shot learning with ChatGPT-4. The detailed instructions are detailed in Figure 8.

For the quadruple extraction task, we first assign a specific role to the dialogue model and inform it of the particular task at hand along with its definition. Following this, we establish several rules to standardize the model's output, making it more aligned with real-label outputs and easier to evaluate using metrics. Specifically, for the few-shot learning with ChatGPT-4, we designed two versions: one that considers non-opinion detection and one that does not. For the version that includes non-opinion detection, we added utterances labeled as 'statement-non-opinion' along with normal containing quadruple utterances to the examples. For the latter version, we only included utterances with quadruple.

Train	Instruction
English	
w NOD	Now you are an expert in conversational sentiment quadruple extraction. Given a conversation that contains the input utterance and its context and the corresponding replying structure, you first need to understand the replying structure and and extract all target-aspect-opinion triples, then identify the sentiment polarity associated with the opinion. Note that: 1) If the corresponding opinion of the target item cannot be found in the conversation, you should output 'statement-non-opinion'. 2) Each element must appear in the conversation. 3) You only need to identify the discussed quadruples from the input utterance. 4) If the corresponding aspect of the target item cannot be found in the conversation, you can use 'Not mentioned' as a substitute. 5) Formulate your output into (target, aspect, opinion, sentiment), ..., ensuring each element is clearly identified and the sentiment must be one of Positive, Neutral or Negative.\n###Context:\n<root>speaker0: So I still bought 12X , although the cost - effective is not high , but I have no choice .\n###Input:<root>speaker0: So I still bought 12X , although the cost - effective is not high , but I have no choice .\n###Replying structure:\nGraph[name=\"dialogue-replying-structure\"]{\n entity_list = ['<root>']\n triple_list = ['(<root> -> <root>)[relation=\"reply\"]']\n}\n###Answer:\nHere are a few examples you can refer to:\n###Input:<u4>speaker0: I sometimes feel that the pictures I shoot are very good , maybe the screen is not very good , and the pictures don't look very good ,\n###Answer:(10 Extreme, screen, not very good, Negative)\n###Input:<u8>speaker0: 13Pro consumption is really so fast [Hum] is not just mine consume power that fast , okay ?\n###Answer:(13Pro, consumption, fast, Negative), (13Pro, consume power, fast, Negative)\n###Input:<u4>speaker2: [Longing] Let 's see how long my 10Pro can be used~\n###Answer:statement-non-opinion\n###Input:<u5>speaker4: Samsung 's battery life , dddd\n###Answer:statement-non-opinion\n
w/o NOD	Now you are an expert in conversational sentiment quadruple extraction. Given a conversation that contains the input utterance and its context and the corresponding replying structure, you first need to understand the replying structure and extract all target-aspect-opinion triples, then identify the sentiment polarity associated with the opinion. Note that: 1) If the corresponding opinion of the target item cannot be found in the conversation, you should output 'statement-non-opinion'. 2) Each element must appear in the conversation. 3) You only need to identify the discussed quadruples from the input utterance. 4) If the corresponding aspect of the target item cannot be found in the conversation, you can use 'Not mentioned' as a substitute. 5) Formulate your output into (target, aspect, opinion, sentiment), ..., ensuring each element is clearly identified and the sentiment must be one of Positive, Neutral or Negative.\n###Context:\n<root>speaker0: I sincerely advise everyone not to buy black sharks ! Intersection\n###Input:<root>speaker0: I sincerely advise everyone not to buy black sharks ! Intersection\n###Replying structure:\nGraph[name=\"dialogue-replying-structure\"]{\n entity_list = ['<root>']\n triple_list = ['(<root> -> <root>)[relation=\"reply\"]']\n}\n###Answer:\nHere are a few examples you can refer to:\n###Input:<u4>speaker0: I sometimes feel that the pictures I shoot are very good , maybe the screen is not very good , and the pictures don't look very good ,\n###Answer:(10 Extreme, screen, not very good, Negative)\n###Input:<u8>speaker0: 13Pro consumption is really so fast [Hum] is not just mine consume power that fast , okay ?\n###Answer:(13Pro, consumption, fast, Negative), (13Pro, consume power, fast, Negative)\n###Input:<u8>speaker0: I feel that my V30pro can still fight for several years , it 's still the core of 990 [allow sad]\n###Answer:(V30pro, Not mentioned, fight for several years, Positive)\n###Input:<u3>speaker3: Brother , Fold3 really ca n't beat X2\n###Answer:(X2, Not mentioned, can't beat, Positive), (Fold3, Not mentioned, can't beat, Negative)\n
Chinese	
w NOD	你现在是一位对话情感四元组提取的专家。 给定一段包含上下文以及输入语句的对话以及对应的回复结构，你首先需要理解句间依赖关系并提取所有的目标-方面-意见三元组，然后识别与意见相关的情感极性最后组成四元组。请注意以下几点： 1) 如果目标对应的意见在对话找不到，你应当输出‘不含意见’。2) 每个元素必须出现在对话中。3) 你只需要提取输入语句中的四元组并且情感必须是积极、中立或消极中的一个。 4) 如果目标项的对应方面在对话中找不到，可以使用‘未提及’作为替代。5) 将你的输出格式化为 (目标, 方面, 意见, 情感)，...，确保每个四元组和元素被提取出来。\n###上下文：\n<root>说话人0: 所以我还是买了 12 x , 虽然性价比不高 , 但是没得选\n###输入语句：<root>说话人0：所以我还是买了 12 x , 虽然性价比不高 , 但是没得选\n###回复结构：\nGraph[name=\“dialogue-replying-structure\”]{\n entity_list = [‘<root>’]\n triple_list = [‘(<root> -> <root>)[relation=\“reply\”]’]\n}\n###你的回答：这里有几个示例你可以作为参考:\n###输入语句：<u2>说话人2：长焦分为潜望式长焦和普通长焦 [doge] 潜望式长焦可以做更大的光学变焦倍数 [doge]\n###你的回答：不含意见\n###输入语句：<u2>说话人0：P 50 怎么样 ? P 50 没有 5 G 还是不太想入手的\n###你的回答：(P50, 未提及, 不太想入手, 消极)\n###输入语句：<u5>说话人0：原来是蓝厂的 boy [杰瑞]\n###你的回答：不含意见\n###输入语句：<root>说话人0：红米吧 , 系统维护一方面 , 真我的话系统方面不行 。 其他方面红米比较均衡 , 系统功能性方面更比不上红米\n###你的回答：(真我, 系统, 不行, 消极), (红米, 其他, 比较均衡, 中性), (红米, 系统功能性, 比不上, 积极), (真我, 系统功能性, 比不上, 消极)\n
w/o NOD	你现在是一位对话情感四元组提取的专家。 给定一段包含上下文以及输入语句的对话以及对应的回复结构，你首先需要理解句间依赖关系并提取所有的目标-方面-意见三元组，然后识别与意见相关的情感极性最后组成四元组。请注意以下几点： 1) 如果目标对应的意见在对话找不到，你应当输出‘不含意见’。2) 每个元素必须出现在对话中。3) 你只需要提取输入语句中的四元组并且情感必须是积极、中立或消极中的一个。 4) 如果目标项的对应方面在对话中找不到，可以使用‘未提及’作为替代。5) 将你的输出格式化为 (目标, 方面, 意见, 情感)，...，确保每个四元组和元素被提取出来。\n###上下文：\n<root>说话人0: 所以我还是买了 12 x , 虽然性价比不高 , 但是没得选\n###输入语句：<root>说话人0：所以我还是买了 12 x , 虽然性价比不高 , 但是没得选\n###回复结构：\nGraph[name=\"dialogue-replying-structure\"]{\n entity_list = ['<root>']\n triple_list = ['(<root> -> <root>)[relation=\"reply\"]']\n}\n###你的回答：这里有几个示例你可以作为参考:\n输入语句：<u7>说话人3：19 年买的 mate 30 也贼好用\n###你的回答：(mate30, 未提及, 贼好用, 积极)\n###输入语句：<u2>说话人0：P 50 怎么样 ? P 50 没有 5 G 还是不太想入手的\n###你的回答：(P50, 未提及, 不太想入手, 消极)\n###输入语句：<u9>说话人3：一个 3 月一个 8 月 , 等得起就等呗 , 而且 mix 拍照不咋地\n###你的回答：(mix, 拍照, 不咋地, 消极)\n###输入语句：<root>说话人0：红米吧 , 系统维护一方面 , 真我的话系统方面不行 。 其他方面红米比较均衡 , 系统功能性方面更比不上红米\n###你的回答：(真我, 系统, 不行, 消极), (红米, 其他, 比较均衡, 中性), (红米, 系统功能性, 比不上, 积极), (真我, 系统功能性, 比不上, 消极)\n

Figure 8: Instructions for conducting few-shot learning with ChatGPT4 in quadruple extraction task.