Representative Chain-of-Reasoning for Aspect Sentiment Quad Prediction

Anonymous ACL submission

Abstract

Aspect Sentiment Quad Prediction (ASQP) is a crucial sentiment analysis task that has attracted increasing attention. The most recent studies focus on generating complete sentiment quadruples through end-to-end generative models. However, these methods heavily depend on labeled data quality and quantity, performing poorly in low-resource scenarios and less suitable for real-world applications. To address these issues, we propose a novel ***R**epresentative **C**hain-of-**R**easoning* framework (RCR), with the aim of providing representative knowledge for large language models (LLMs) and fully activating their reasoning capabilities for ASQP. Specifically, we develop a Chain Prompting (ChaPT) module to decompose the ASQP task into three subtasks using the step-by-step reasoning mechanism. Then, a Representative Demonstration Retriever (RepDR) is introduced to provide ChaPT with representative demonstrations, balancing diversity and similarity, and enhancing the reasoning capabilities of LLMs at each step. Experimental results confirm the superiority of RCR in both zero-shot and few-shot scenarios, significantly surpassing existing counterparts.

1 Introduction

Given a review text, Aspect Sentiment Quad Prediction (ASQP) aims to predict a comprehensive sentiment view in the form of quadruples (Zhang et al., 2022a, 2023b), each consisting of aspect category, aspect term, opinion term, and sentiment polarity, denoted as (c, a, o, s). For example, given the review sentence, "*The food is great and the environment is even better.*", the ASQP task requires predicting two sentiment quadruples: (*food quality, food, great, positive*) and (*ambiance general, environment, better, positive*). ASQP is a challenging task due to the complexity of sentence structure and the diversity of sentiment expressions, making it difficult to recognize all sentiment quadruples.

Recently, the end-to-end generative models have been extensively applied to solve the ASQP task by generating sentiment quadruples directly from the review text and achieved promising results (Peper and Wang, 2022). A successful application is to construct sequences in natural language format as generation targets, including annotated sentences (Zhang et al., 2021b), paraphrased sentences (Liu et al., 2021; Zhang et al., 2021a; Hu et al., 2022a), and sentiment element sequences (Zhang et al., 2021c). Furthermore, sentiment clues within sentences have been utilized to promote the quadruple generation (Bao et al., 2022). For example, Mao et al. (2022) introduced a parallel generation framework to capture more sentiment information through beam search. Gou et al. (2023) enhanced the model's expressive capability by increasing the output views by adjusting the generation order of quadruples. Despite their potential, a notable issue is that these models are less suitable in low-resource scenarios (Hu et al., 2022a; Gou et al., 2023). That is because generative models heavily rely on the scale and quality of the labeled dataset while annotating datasets is costly and time-consuming in practical applications.

With the rise of In-Context Learning (ICL), addressing the ASQP task by generative models in zero-shot and few-shot scenarios becomes feasible (Wang et al., 2023c; Zhang et al., 2023a,c). Sun et al. (2023) propose a multi-LLM negotiation strategy, demonstrating LLMs' ability to solve sentiment analysis problems involving complex contexts (*e.g.,* clauses and irony) under zero-shot conditions. Moreover, the Three-Hop Reasoning (THOR) CoT framework (Fei et al., 2023) achieves state-of-the-art results in implicit sentiment analysis tasks. However, existing research lacks a discussion on applying LLMs to ASQP tasks, and the reasoning capabilities of LLMs are underutilized.

To this end, we propose a Representative Chain-of-Reasoning framework (RCR) that aims to provide representative knowledge for LLMs and fully activate their reasoning capabilities for the ASQP task. Inspired by the Chain-of-Thought (CoT) prompting (Wei et al., 2022; Zhou et al., 2022; Zhang et al., 2022b), we first introduce a Chain Prompt (ChaPT) module to decompose the one-step ASQP task into three sub-steps, where each step progressively infers aspect-opinion pairs, category-aspect-opinion triplets, and complete quadruples. Hence, a complete sentiment view is obtained through step-by-step reasoning, effectively reducing the ASQP task's complexity. Additionally, considering LLM's reasoning capability is influenced greatly by the quality of demonstrations (Lee et al., 2022; Min et al., 2022; Wang et al., 2023b), we develop a Representative Demonstration Retriever (RepDR) module to provide ChaPT with representative demonstrations, balancing diversity and similarity, and thus enhancing their reasoning capabilities at each step. Specifically, we first paraphrase the sentiment quadruples into natural sentences (Zhang et al., 2021a) and calculate their semantic similarities using SBERT (Reimers and Gurevych, 2019). Based on semantic similarities, the triplet that contains an anchor sentence, a positive sentence, and a negative sentence is picked for further fine-tuning this SBERT model. Hence, this fine-tuned SBERT model is good at retrieving representative demonstrations that possess semantic information of different attributes (Wang et al., 2022a; Shi et al., 2023; Qin et al., 2023).

In summary, the main contributions of this work are as follows:

- We introduce ChaPT, a prompting framework based on the chain-of-reasoning concept, which mitigates task complexity through task decomposition and step-by-step reasoning, fully leveraging the reasoning capabilities of LLMs.
- We use the RepDR module to retrieve demonstrations, providing more representative prior information for model reasoning. To the best of our knowledge, this work is the first to propose retrieving both diversity and similarity samples as demonstrations.
- Experimental results show that our proposed model demonstrates superiority in both zero-shot and few-shot scenarios and greatly surpasses existing counterparts.

2 Methodology

2.1 Problem Definition

The ASQP task is defined as follows: given a sentence X, the model predicts all aspect-based sentiment quadruples (Cai et al., 2021; Zhang et al., 2021a), each formulated as (c, a, o, s) which corresponds to aspect category, aspect term, opinion term, and sentiment polarity, respectively. The aspect category c is part of a predefined category set U_c. The aspect term a is the target of opinion. The opinion term o is the subjective statement. Moreover, the sentiment polarity s belongs to the predefined sentiment set $U_s \in \{positive, neutral, negative\}$. Notably, aspect a is generally within the text scope of sentence X, and if aspect a is not explicitly mentioned, it is represented by the specific tag NULL.

Xie et al. (2021) believe that ICL infers conditional probabilities of the predictive target from the prompt, formulated as $p(y \mid prompt) = \int_{prompt} p(y \mid prompt) d(prompt)$, where y represents the prediction target and d(prompt) represents the prompt set. ICL infers the maximum probability of generating y by integrating y over the $prompt$. Therefore, we can use ICL to model the ASQP task as

$$\hat{y} = \operatorname{argmax} p(y \mid X, prompt) \tag{1}$$

where $\hat{y}$ represents all quadruples, and Zhang et al. (2023c) attempt to construct the following standard $prompt$ paradigm as input to the LLMs:

```
Given the sentence X, tag all the (category, aspect, opinion, sentiment) quadruples.
```

2.2 Representative Chain-of-Reasoning

To fully activate the reasoning capabilities of LLMs for the ASQP task, we propose a Representative Chain-of-Reasoning framework (RCR) consisting of two sub-modules: Chain Prompt Framework (ChaPT) and Representative Demonstration Retriever (RepDR). The former is designed to decompose the ASQP task into three subtasks, while the latter is responsible for providing representative demonstrations to enhance LLMs' reasoning capabilities.

2.2.1 Chain Prompt Framework

Inspired by the impressive reasoning capabilities demonstrated by Chain of Thought (CoT) in han-

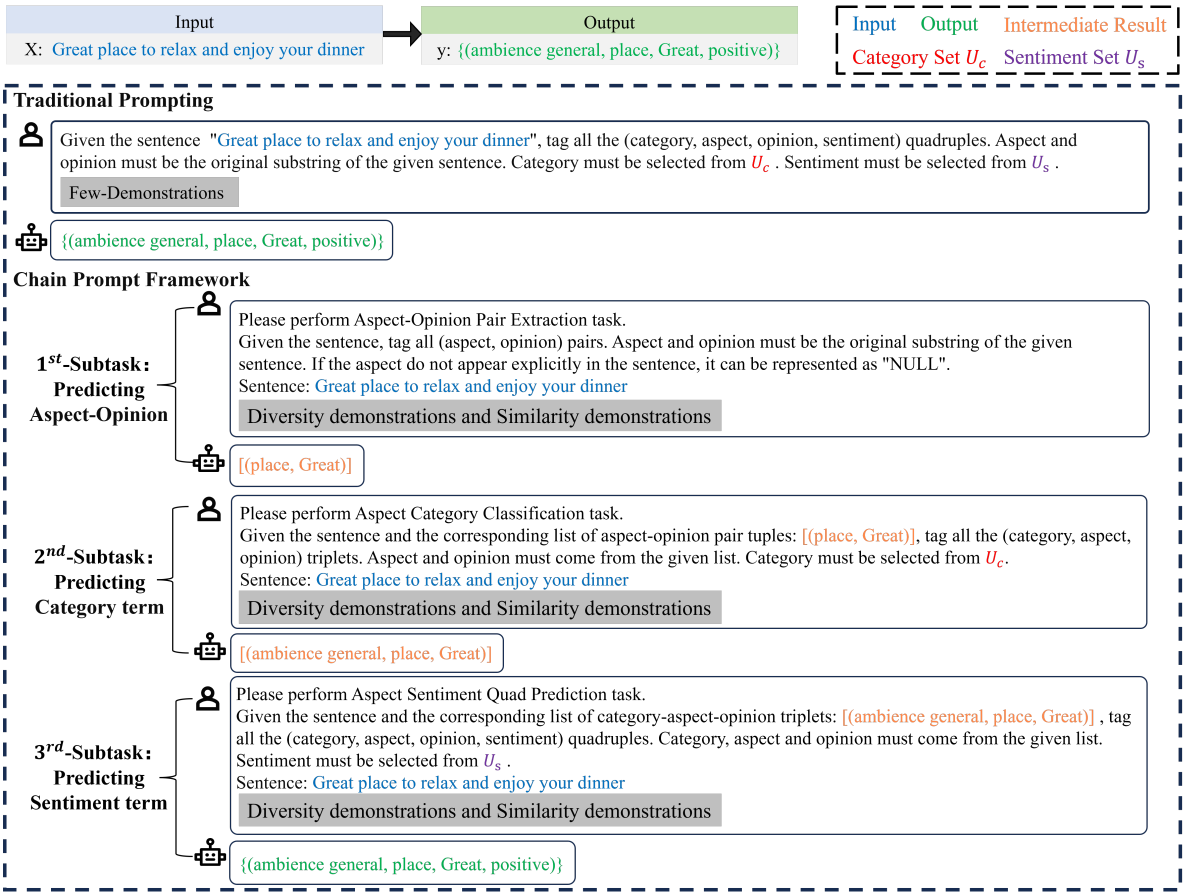


Figure 1: An illustration of our ChaPT framework for Aspect Sentiment Quad Prediction task.

dling complex tasks(Wei et al., 2022), we propose the Chain Prompt framework(ChaPT), shown in Figure 1, to address the ASQP task by decomposing a one-step ASQP solution into three subtasks. The details are as follows.

Subtask 1. Aspect-Opinion Pair Extraction
Empirical studies find that extracting a single aspect or opinion alone would ignore their pairwise relationships, leading to pairing errors (Chen et al., 2020; Zhao et al., 2020). Therefore, instead of finer-grained aspect term or opinion term extraction subtask, we first consider predicting all aspect-opinion pairs appearing in the sentence. Mathematically, this subtask is formulated as:

$$\hat{Z}_1 = \operatorname{argmax} p(y \mid X, prompt_1) \quad (2)$$

where $\hat{Z}_1$ denotes predicted aspect-opinion pairs, the template of $prompt_1$ is defined as:

> Given the sentence X, tag all the (aspect, opinion) pairs.

Subtask 2. Aspect Category Classification
Based on X and the intermediate results $\hat{Z}_1$, we classify category c from the predefined set U_c and obtain the category-aspect-opinion triplets Z_2. This process is represented as:

$$\hat{Z}_2 = \operatorname{argmax} p(y \mid X, Z_1, prompt_2) \quad (3)$$

the template of $prompt_2$ is as follows.

> Given the sentence X and the corresponding list of aspect-opinion pair tuples $\hat{z}_1$, tag all the (category, aspect, opinion) triplets.

Subtask 3. Aspect Sentiment Quad Prediction
Based on the intermediate results $\hat{Z}_2$, we finally predict the complete quadruples y. The final step is denoted as:

$$\hat{y} = \operatorname{argmax} p(y \mid X, Z_2, prompt_3) \quad (4)$$

and the template of $prompt_3$ is as follows.

> Given the sentence X and the corresponding list of category-aspect-opinion triplets $\hat{z}_2$, tag all the (category, aspect, opinion, sentiment) quadruples.

2.2.2 Representative Demonstration Retriever

In few-shot scenarios, we propose the Representative Demonstration Retriever (RepDR), a demon-

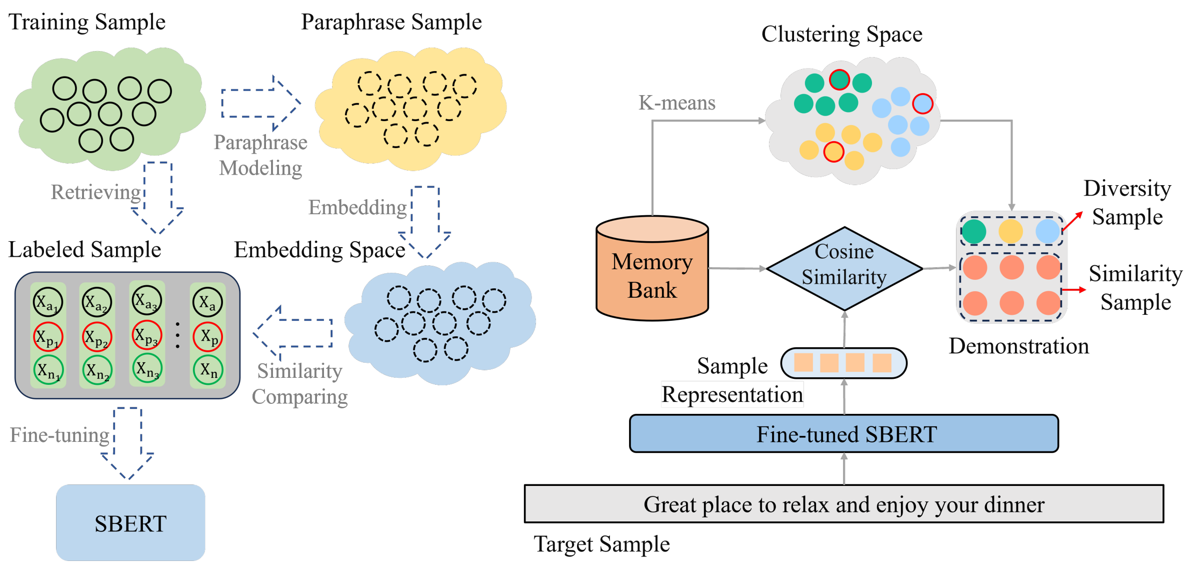


Figure 2: The proposed RepDR module consists of two stages. The first stage generates training samples with pre-trained SBERT, and fine-tunes SBERT using these samples. The second stage generates text embeddings using the fine-tuned model and utilizes clustering and cosine similarity to produce representative demonstrations.

stration retriever that balances diversity and similarity. First, we explain the generation of training sample, then describe training the model with the labeled sample, and finally show using the trained model to retrieve representative demonstrations. As shown in Figure 2.

Generating Training Sample Since we utilize clustering and similarity comparison for demonstration retrieval, it is vital to train a model that precisely captures the similarity among sentence pairs. We choose SBERT (Reimers and Gurevych, 2019), a BERT-based text embedding model, as our target model. By fine-tuning SBERT with generated triplet data, we enhance its ability to capture semantic similarity between sentences. Triplet data consists of an anchor sentence, a positive sentence, and a negative sentence without additional labels. Inspired by paraphrase generation (Zhang et al., 2021a; Gou et al., 2023; Hu et al., 2022b), we propose modeling paraphrases for the training set by linearizing sentiment quadruples (c, a, o, p) into natural sentences I, as shown in Figure 3.

Paraphrase modeling allows us to focus on the quadruples and ignore unnecessary details in sentences. We compute their embeddings using a pre-trained SBERT model for the resulting paraphrase set U_I. Then, we compare the semantic similarity between these paraphrases by employing Cosine Similarity. We retrieve the sentence most similar to the anchor sentence X_a as the positive sentence X_p and the least similar sentence as the negative sentence X_n, thus constructing triplet training samples (X_a, X_p, X_n). Notably, the sentences referred to here are the original sentences, not the paraphrase sentences, to ensure the model focuses more attention on quadruples in the original sentence.

Sentence-1	Our teenage kids love it , too .
Quadruplet-1 ⇩ **Paraphrase-1**	(restaurant general, NULL, love , positive) ⇩ restaurant general is positive because it is love
Sentence-2	The only thing more wonderful than the food (which is exceptional) is the service .
Quadruplet-2 ⇩ **Paraphrase-2**	(food quality, food, exceptional, positive), (service general, service, wonderful, positive) ⇩ food quality is positive because food is exceptional and service general is positive because service is wonderful

Figure 3: Two examples of paraphrase modeling. Notably, if the aspect is not explicitly mentioned, it is represented by the implicit pronoun "it". If a sentence contains multiple sentiment quadruples, the paraphrases are concatenated using "and".

Training Model We fine-tune the SBERT model using the typical triplet network. Given triplet data I, we utilize SBERT (selected mpnet (Song et al., 2020)) to encode X_a, X_p, and X_n, obtaining embeddings O_a, O_p, and O_n. Fine-tuning aims to minimize the distance between O_a and O_p while maximizing the distance between O_a and O_n. We use triplet loss as the loss function, as shown below in Equation 5:

$$\mathcal{L}(O_a, O_p, O_n) \\ = \max(d(O_a, O_p) - d(O_a, O_n) + \alpha, 0) \quad (5)$$

where $d(O_a, O_p) = \|O_a - O_p\|_2$ represents the Euclidean Distance between embeddings. The hyperparameter α specifies the expected difference between $d(O_a, O_p)$ and $d(O_a, O_n)$.

Retrieving Demonstration Firstly, we use the fine-tuned SBERT to encode all training set sentences into text embeddings and store them in a Memory bank (Wu et al., 2018) to avoid redundant computations. Secondly, we measure the similarity of demonstrations utilizing Cosine Similarity, comparing target samples with the training set to extract the top-k most similar samples. Finally, we propose a diversified demonstration retrieval scheme based on K-means clustering (Arthur et al., 2007). We evaluate the optimal number of clusters by calculating the Silhouette Score and find that the optimal number for both datasets is 3(see Appendix A). Based on this result, we perform K-means clustering and select the samples closest to the cluster centers as diversity samples that highlight key characteristics.

3 Experiments

3.1 Datasets

We conducted experiments on two public restaurant datasets, `Rest15` and `Rest16`, from the SemEval task (Pontiki et al., 2015, 2016). These datasets, with multiple annotations (Peng et al., 2020; Wan et al., 2020), were aligned by Zhang et al. (2021a) and ultimately served as the standard datasets for the ASQP task. Each sample contains one or more sentiment quadruples. The statistics are shown in Table 1.

	Rest15				Rest16			
	SEN	POS	NUE	NEG	SEN	POS	NUE	NEG
Train	834	1005	34	315	1264	1369	62	558
Dev	209	252	14	81	316	341	23	143
Test	537	453	37	305	544	583	40	176

Table 1: Dataset statistics for Rest15 and Rest16. SEN, POS, NUE and NEG represent the number of sentences, positive, neutral, and negative quadruples, respectively.

3.2 Implementation Details

We utilized two OpenAI models, including ChatGPT (Open, 2022) (gpt-3.5-turbo3) and the newly released GTP-4 (Achiam et al., 2023) (gpt-4o), as the backbone for the ChaPT framework (Section 2.2.1) to evaluate its effectiveness under zero-shot conditions. The temperature for all models was set to 0 to ensure stable predictions.

Moreover, for few-shot scenarios, we employed all-mpnet-base-v1 (Song et al., 2020) as pre-trained SBERT (Section 2.2.2), using a typical triplet network for fine-tuning. During fine-tuning, we used a batch size of 64, a learning rate of 2e-5, and 5 training epochs. The hyperparameter α of the model was set to 5. Additional implementation details of generative models in low-resource scenarios are provided in Appendix B.

During demonstration retrieval, we used the model at the best checkpoint to re-encode sentences for text similarity comparison and clustering, obtaining demonstrations with similarity and diversity. We only considered three k-shot settings: 1-shot, 5-shot, and 10-shot. For each setting, we maintained a constant number of diversity demonstrations and adjusted the number of similarity demonstrations to achieve k-shot. For example, in the 1-shot scenario, we retrieved 3 diversity samples and the 1 most similar sample.

3.3 Baselines

We employ generative-based models and ICL-based large language models as our comparative baselines. For generative methods, we select the following four models:

- **GAS** (Zhang et al., 2021b) The first attempt to use generative methods to handle aspect-based sentiment analysis, we modify it to use sentiment quadruple sequences as target sequences.

- **Paraphrase** (Zhang et al., 2021a) A paraphrasing modeling framework, using paraphrased sentences as training targets to generate sentiment quadruples end-to-end.

- **DLO/ILO** (Hu et al., 2022a) Selecting the appropriate quadruple generation order as a data augmentation method for paraphrase generation.

- **MVP** (Gou et al., 2023) Enhances the model predictive capability by increasing output views of different quadruple generation orders.

	Rest15				Rest16			
	0-shot	1-shot	5-shot	10-shot	0-shot	1-shot	5-shot	10-shot
Generative-based Baselines								
GAS† (Zhang et al., 2021b)	-	4.43	10.65	13.82	-	2.24	16.04	19.03
Paraphrase† (Zhang et al., 2021a)	-	7.78	11.53	18.60	-	2.36	12.85	16.34
DLO† (Hu et al., 2022a)	-	6.79	13.07	18.92	-	1.90	17.68	28.95
ILO† (Hu et al., 2022a)	-	7.25	14.85	20.99	-	2.41	15.71	21.32
MvP† (Gou et al., 2023)	-	9.33	18.54	22.82	-	3.62	21.51	29.24
Prompt-based Baselines								
w/ GPT-3.5								
LMMs for SA† (Zhang et al., 2023c)	7.77	27.83	26.86	25.74	10.06	28.45	38.63	37.13
THOR† (Fei et al., 2023)	10.21	22.13	27.24	23.51	14.11	28.62	37.26	36.04
RCR(Ours)	13.82	28.46	31.44	32.29	19.00	30.60	41.51	42.09
w/ GPT-4								
LMMs for SA† (Zhang et al., 2023c)	<u>32.40</u>	<u>35.63</u>	<u>37.65</u>	<u>36.72</u>	35.87	<u>40.56</u>	<u>42.07</u>	<u>40.32</u>
THOR† (Fei et al., 2023)	30.57	35.01	36.22	35.81	<u>36.37</u>	38.02	41.58	39.97
RCR(Ours)	**33.01**	**39.78**	**42.26**	**44.33**	**38.07**	**40.97**	**48.08**	**51.23**

Table 2: Report the model's experimental results under zero-shot and few-shot settings. F1 score is used as the evaluation metric. The best and second-best results are indicated in bold and underlined, respectively. The baseline methods, marked with†, follow the few-shot settings of this work (Zhang et al., 2023c), where k-shot represents sampling k examples for each aspect category.

For ICL methods, we selected the following research approaches:

- **LMMs for SA** (Xu et al., 2024) A comprehensive study of sentiment analysis using LLMs, including Flan-T5, FLan-UL2, T5, and GPT-3.5.
- **THOR** (Fei et al., 2023) A Three-hop reasoning (THOR) CoT framework for addressing implicit sentiment analysis issues. In our setup, it serves as one of the benchmarks for ICL methods by modifying the prompt.

Experimental results for these supervised methods are derived from the base pre-trained models (BERT or T5) to ensure a fair comparison.

4 Results and Discussions

4.1 Zero-shot and Few-shot Results

The experimental results are shown in Table 2. Notably, in the zero-shot scenario, the T5-based (Raffel et al., 2020) generative model struggled with ASQP tasks and failed to generate effective results. However, the performance gradually improved as the number of samples increased, highlighting the importance of high-quality labeled data for generative models. Compared to the best generative model baseline MVP, the ICL-based LLMs (LLMs for SA) showed significant performance improvements in both zero-shot and few-shot scenarios. For

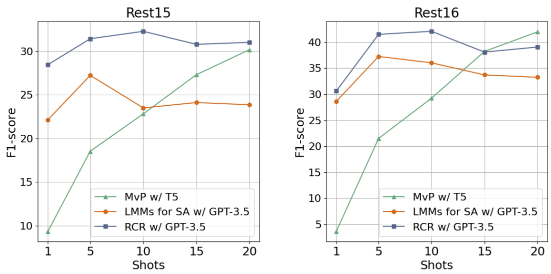


Figure 4: The evaluation curve of the model with varying sample sizes.

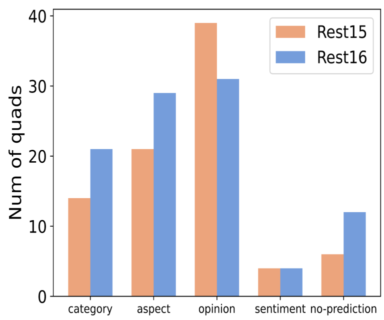
GPT-3.5, under few-shot conditions, the average F1 score gained on the Rest15 and Rest16 datasets was 9.91% and 16.61%, respectively. This demonstrates the great potential of LLMs in ASQP tasks. Furthermore, our proposed Representative Chain-of-Reasoning (RCR) framework achieved the best performance with both GPT-3.5 and GPT-4 compared to the original ICL baselines. Specifically, with GPT-3.5, the average F1 score gained on the Rest15 and Rest16 datasets were 4.45% and 4.73%, respectively. With GPT-4, the F1 scores improved by an average of 4.24% and 4.88%. This indicates that the RCR framework provides sufficient prior information for LLMs, fully leveraging their reasoning capabilities in ASQP tasks.

4.2 Ablation Study

We conducted ablation experiments to further validate our RCR framework's effectiveness. In a

Methods	Rest15			Rest16		
	Pre	Rec	F1	Pre	Rec	F1
RCR	**27.99**	**35.85**	**31.44**	**36.82**	**47.56**	**41.51**
RCR w/o [RepDR]	23.39	34.21	27.78	24.53	32.42	27.92
RCR w/o [ChaPT]	21.83	27.30	24.26	28.16	36.29	31.71
RCR w/o [RepDR,ChaPT]	21.25	26.68	23.66	25.10	30.16	27.40

Table 3: The results of ablation study.



Example 1(Rest15)

Sentence: Watch the talented belly dancers as you enjoy delicious baba ganoush that s more lemony than smoky .
Gold: (food quality, baba ganoush, enjoy delicious, positive), (ambience general, belly dancers, talented, positive)
Prediction:
(ambience general, belly dancers, talented, positive) ✓
(food quality, baba ganoush, delicious, positive) ✗
(food quality, baba ganoush, lemony, positive) ✗

Example 2(Rest16)

Sentence: The regular menu here is slightly above average that is not worth the snotty attitude that you receive .
Gold: (food quality, regular menu, above average, neutral)
(food quality, regular menu, not worth the snotty attitude, neutral)
(service general, NULL, snotty, negative)
Prediction:
(food quality, menu, slightly above average, positive) ✗
(service general, attitude, snotty, negative) ✗
(service general, NULL, snotty, negative) ✓

Figure 5: Statistics of error types and two examples of prediction errors. Notably, no-prediction indicates the samples where LLMs made no predictions.

5-shot scenario, we analyzed the impact on the results by removing individual modules, with results shown in Table 3. ChaPT decomposes the ASQP task into subtasks, reducing the complexity of LLM reasoning. RepDR is responsible for providing more accurate prior semantic knowledge to LLMs through demonstration retrieval. The results indicate that removing any module significantly reduces RCR performance, demonstrating the effectiveness of ChaPT and RepDR in stimulating the reasoning capabilities of LLMs.

Furthermore, We observed performance differences across the Rest15 and Rest16 datasets when removing specific modules. For instance, removing the RepDR module resulted in a 13.59% decrease in F1 score for Rest16, but only a 3.66% decrease for Rest15. This indicates that different datasets have varying dependencies on the ChaPT and RepDR modules, reflecting the distinct knowledge support these two components provide to LLMs.

4.3 Influence of Different Sample Sizes

Our preliminary research reveals that the LLMs' reasoning capabilities for ASQP tasks improve significantly with an increased sample size. However, this raises the question of whether this improvement is always directly proportional to the number of samples. To explore this issue, we further increased the sample size, as shown in Figure 4. We found that the T5-based MvP model's performance steadily improved with more samples, indicating that the generative-based models rely on sufficient high-quality labeled data. Surprisingly, for ICL-based methods, performance tends to decline after reaching a certain sample size threshold. Our analysis suggests two main reasons for this decline. First, a large number of examples provides excessive prior semantic information, causing LLMs to become confused and lose focus on core aspects. Second, lower-ranked samples are poorer in quality and contain more redundancy. Notably, compared to previous ICL methods, the RCR framework mitigates this performance degradation, indicating that RepDR retrieves higher-quality demonstrations and introduces fewer errors.

4.4 Error analysis and Case Study

In order to comprehensively analyze the reasoning errors of our proposed method, we conducted error analysis and case studies. We randomly selected 100 prediction results from each dataset in the 5-shot scenario using GPT-4. The incorrectly predicted quadruples were categorized by error type, as shown in Figure 5. We found that errors were primarily concentrated on the predictions of aspect and opinion terms in both datasets.

The main reason for this phenomenon is that aspect and opinion terms often appear as text spans rather than individual words. LLMs struggle to match these text spans accurately, as illustrated by Example 1. Another significant cause of errors is the presence of multiple quadruples in the text, which confuses the LLMs. This typically occurs in the first subtask of the ChaPT framework, making it difficult to match each aspect-opinion pair precisely. Example 2 shows an incorrect aspect-opinion pair (attitude, snotty) being generated. Furthermore, errors from the previous subtask can propagate and interfere with predicting aspect categories, leading to cumulative errors. In summary, accurately matching text spans and handling sentences with multiple quadruples are challenging issues that LLMs must address in ASQP problems.

5 Related work

5.1 Aspect Sentiment Quad Prediction

Aspect Sentiment Quad Prediction (ASQP) is a crucial sentiment analysis task that has attracted increasing attention. (Zhang et al., 2022a, 2023b; Zhong et al., 2023). Initially, ASQP was mainly handled using pipeline approaches that combined multiple baseline models (Cai et al., 2021). Further studies have shown that generative models achieve promising results (Zhang et al., 2021b; Bao et al., 2022; Peper and Wang, 2022). For example, Zhang et al. (2021a) introduced a paraphrase model, transforming a quadruple prediction task into a text generation task. Mao et al. (2022) constructed a search tree for the optimal generation path. Bao et al. (2022) developed an opinion tree to jointly detect all sentiment elements. Additionally, many efforts have focused on enhancing generative models through data augmentation. Hu et al. (2022a) first considered selecting the appropriate quadruple generation order as a data augmentation method. Gou et al. (2023) proposed an MVP framework to increase output views. Wang et al. (2023a) suggested generating new data containing quadruples through generation models. However, models trained on specific domain datasets often perform poorly when transferred to other domains.

5.2 In-Context Learning

In-context learning (ICL) refers to the ability of large language models (LLMs) to handle complex tasks with only a few annotated examples without additional training or gradient updates (Zhao et al., 2023). Research on ICL focuses on two main areas. On the one hand, it involves investigating prompting frameworks (Long, 2023; Paranjape et al., 2023; Diao et al., 2023; Li et al., 2024). For example, Wei et al. (2022); Wang et al. (2022b) proposed the Chain of Thought (CoT) to enhance reasoning capabilities. Yao et al. (2024) further refined CoT into the Tree of Thoughts (ToT), maintaining the intermediate thoughts in a search tree and evaluating these thoughts. On the other hand, considerable work studies focus on providing better demonstrations (Li et al., 2022; Min et al., 2022; Li et al., 2023; Wang et al., 2023b). Liu et al. (2022) found that samples closely related to the target data in the embedding space perform better. Building on this idea, Wang et al. (2022a) proposed enhancing inputs by retrieving similar examples. Rubin et al. (2022) introduced a demonstration retriever. Moreover, examples representing diversity can also improve ICL performance (Qin et al., 2023; Xu et al., 2024).

Owing to developments in ICL, some studies have addressed sentiment analysis tasks in zero-shot or few-shot scenarios using ICL, achieving effective results (Wang et al., 2023c). For instance, Zhong et al. (2023) observed that the zero-shot performance of LLMs is comparable to fine-tuned BERT. Sun et al. (2023) proposed a multi-LLM negotiation framework for sentiment analysis. Fei et al. (2023) introduced a THOR framework, significantly enhancing implicit sentiment analysis performance. In light of this, we explore the potential of LLMs for the ASQP problem. To our knowledge, this work is the first to discuss the application of LLMs to ASQP task systematically.

6 Conclusion

In this work, we propose a new RCR framework to solve the ASQP task in low-resource scenarios. To reduce complexity, the chain prompting module (ChaPT) is designed to decompose the ASQP task into three subtasks and enable LLMs to conduct step-by-step reasoning. Furthermore, a representative demonstration retriever (RepDR) is developed to provide ChaPT with demonstrations that balance diversity and similarity, maximizing the reasoning ability of LLMs at each step. Detailed experiments demonstrate the effectiveness of our proposed RCR framework in both zero-shot and few-shot scenarios, enabling GPT-4 to achieve state-of-the-art performance on the ASQP task.

Limitations

Despite our proposed method achieves state-of-the-art performance in ASQP tasks under low-resource scenarios, our work still has limitations. Firstly, we observe that the performance of RCR improves with the increasing intelligence of the integrated LLM models. Therefore, it is necessary to explore the effects of integrating LLMs of different scales with RCR. Secondly, our proposed ChaPT framework requires manually designed prompts, leading to instability in LLM reasoning results as the quality of the prompts varies. Exploring better automatic prompt generation strategies could address this issue. Finally, the experiments only validate the improvements of RCR in the ASQP task. Intuitively, the RCR framework can be easily extended to aspect-based Sentiment analysis subtasks similar to ASQP, such as Aspect Sentiment Triplet Extraction (ASTE), Aspect-Category-Sentiment Detection (ACSD), and Aspect Category Opinion Sentiment (ACOS).

Ethical Statement

All our experiments are based on publicly available datasets and code repositories. We maintain impartiality and honesty in our analysis of the experimental results, and our research and work do not harm any individuals or groups. We will open-source our code for further discussion and exploration. Regarding broader impacts, this work may promote further research using large language models(LLMs) for sentiment analysis tasks in low-resource scenarios, contributing to lightweight and automated opinion mining and sentiment analysis in the real world. Additionally, we recognize the robust capabilities and potential risks of LLMs. Thus, we strictly adhere to ethical standards throughout our research to ensure that our work is not misused or causes any negative impact.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Arthur, Sergei Vassilvitskii, et al. 2007. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035.

Xiaoyi Bao, Zhongqing Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *IJCAI*, volume 2022, pages 4044–4050.

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, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6515–6524.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

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.

Zhibin Gou, Yujiu Yang, et al. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022a. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022b. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900.

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2022. Good examples make a faster learner: Simple demonstration-based learning for low-resource ner. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2687–2700.

Kaijian Li, Shansan Gong, and Kenny Q Zhu. 2022. Few-shot natural language inference generation with pdd: Prompt and dynamic demonstration. *arXiv preprint arXiv:2205.10593*.

Mukai Li, Shansan Gong, Jiangtao Feng, Yiheng Xu, Jun Zhang, Zhiyong Wu, and Lingpeng Kong. 2023. In-context learning with many demonstration examples. *arXiv preprint arXiv:2302.04931*.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2024. Guiding large language models via directional stimulus prompting. *Advances in Neural Information Processing Systems*, 36.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? *DeeLIO 2022*, page 100.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4406–4416.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

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.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, editor = "Goldberg Yoav Zettlemoyer, Luke", Zornitsa Kozareva, and Yue Zhang. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

AI Open. 2022. Introducing chatgpt. open ai.

Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 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.

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. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Chengwei Qin, Aston Zhang, Anirudh Dagar, and Wenming Ye. 2023. In-context learning with iterative demonstration selection. *arXiv preprint arXiv:2310.09881*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. 2023. Why larger language models do in-context learning differently? In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9122–9129.

An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. 2023a. Generative data augmentation for aspect sentiment quad prediction. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 128–140.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022a. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023b. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2023c. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Xin Xu, Yue Liu, Panupong Pasupat, Mehran Kazemi, et al. 2024. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *FinLLM at IJCAi*.

Hao Zhang, Yu-N Cheah, Osamah Mohammed Alyasiri, and Jieyu An. 2023b. A survey on aspect-based sentiment quadruple extraction with implicit aspects and opinions.

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.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023c. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. 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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022a. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. 2021c. Towards navigation by reasoning over spatial configurations. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 42–52.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3239–3248.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Silhouette Scores

Figure 6 shows the results of running K-means and calculating silhouette scores on `Rest15` and

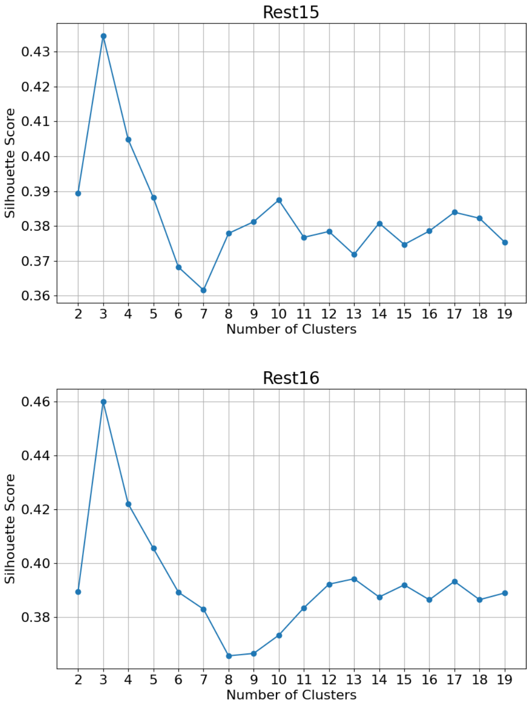


Figure 6: Silhouette Scores for different number of clusters.

`Rest16`. A larger silhouette score indicates better clustering quality. Therefore, from Figure 6, we can determine that the optimal number of clusters for both datasets is 3. To maintain clustering stability, we first standardize, normalize, and reduce dimensionality of the sentence embedding.

B Implementation Details of Generative Models

The few-shot training of generative models follows the settings proposed by Zhang et al. (2023c), where k-shot represents sampling k examples for each aspect category. We set the batch size of all models to 8, the learning rate to 1e-4, and the training epochs to 100. All experiments were conducted using an Nvidia RTX 3090 GPU.