Is Compound Aspect-Based Sentiment Analysis Addressed by LLMs?

Yinhao Bai[1], Zhixin Han[1], Yuhua Zhao[1], Hang Gao[2], Zhuowei Zhang[1], Xunzhi Wang[1], Mengting Hu[1]*,
[1] College of Software, Nankai University,
[2] College of Artificial Intelligence, Tianjin University of Science and Technology
{yinhao,zhixinhan,zyh22,zhuoweizhang,xunzhi}@mail.nankai.edu.cn,
mthu@nankai.edu.cn

Abstract

Aspect-based sentiment analysis (ABSA) aims to predict aspect-based elements from the given text, mainly including four elements, i.e., *aspect category*, *sentiment polarity*, *aspect term*, and *opinion term*. Extracting pair, triple, or quad of elements is defined as compound ABSA. Due to its challenges and practical applications, such a compound scenario has become an emerging topic. Recently, large language models (LLMs), e.g. ChatGPT and LLaMA, present impressive abilities in tackling various human instructions. In this work, we are particularly curious whether LLMs still possess superior performance in handling compound ABSA tasks. To assess the performance of LLMs, we design a novel framework, called ChatABSA. Concretely, we design two strategies: *constrained prompts*, to automatically organize the returned predictions; *post-processing*, to better evaluate the capability of LLMs in recognition of implicit information. The overall evaluation involves 5 compound ABSA tasks and 8 publicly available datasets. We compare LLMs with few-shot supervised baselines and fully supervised baselines, including corresponding state-of-the-art (SOTA) models on each task. Experimental results show that ChatABSA exhibits excellent aspect-based sentiment analysis capabilities and overwhelmingly beats few-shot supervised methods under the same few-shot settings. Surprisingly, it can even outperform fully supervised methods in some cases. However, in most cases, it underperforms fully supervised methods, and there is still a huge gap between its performance and the SOTA method. Moreover, we also conduct more analyses to gain a deeper understanding of its sentiment analysis capabilities.

1 Introduction

Sentiment analysis (SA) is the umbrella term for techniques and approaches that analyze people's sentiments, emotions, and opinions in user-generated content (e.g., blogs, social networks, forums, website reviews, e-commerce websites) (Medhat et al., 2014; Wankhade et al., 2022). To deduce specific sentiment polarities regarding certain aspects of products or services from social media texts or reviews, the field of aspect-based sentiment analysis was born (Do et al., 2019; D'Aniello et al., 2022). ABSA aims to predict aspect-based elements: *aspect term*, *aspect category*, *opinion term*, and *sentiment polarity*, including **single ABSA**, such as aspect term (Chen and Qian, 2020) or aspect category detection (Hu et al., 2019), etc., and **compound ABSA**, such as aspect sentiment triplet extraction (ASTE) and aspect sentiment quad prediction (ASQP) (Zhang et al., 2021a; Cai et al., 2021; Mao et al., 2022; Bao et al., 2022; Hu et al., 2022; Peper and Wang, 2022), etc. Compound ABSA involves multiple-element predictions, bringing more challenges. Peng et al. (2020) define ASTE task by corresponding elements with

Inputs-1	*Great Indian food*
Label-1	(*Indian food*, *Great*, *food quality*, *positive*)
Response-1	*The aspect sentiment quad of the sentence "Great Indian food" is:* • *Aspect category: Food* • *Aspect term: Indian food* • *Opinion term: Great* • *Sentiment polarity: Positive*
Inputs-2	*The food tasted very good .*
Label-2	(*food*, *good*, *food quality*, *positive*)
Response-2	*In the sentence "The food tasted very good .", the aspect category would be "taste" which represents the object or entity that is being evaluated. The aspect term would be "food", which represents the specific aspect of the food being evaluated. The opinion term in this sentence is "good", which is the modifier that adds intensity to the expression of the sentiment. Finally, the sentiment polarity would be positive as the word "good" holds a positive connotation.*
Inputs-3	*its alright .*
Label-3	(*NULL*, *alright*, *restaurant general*, *neutral*)
Response-3	*I'm sorry, but it is not possible to determine the aspect sentiment quad for the given sentence as it does not contain any aspect terms, opinion terms or sentiment polarity.*

Figure 1: Predicted error cases are depicted under the zero-shot settings. **Response** denotes the response from LLMs, and the highlighted parts are the predicted errors. **Label** is shown in the order of (at, ot, ac, sp).

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7836–7861

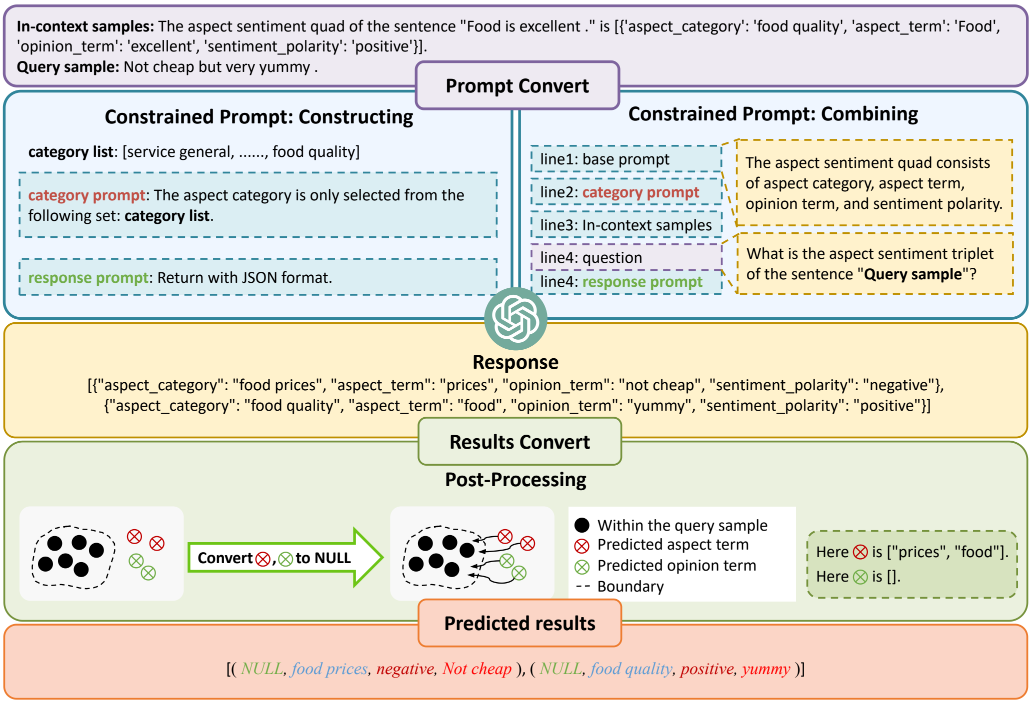

Figure 2: An overview of ChatABSA. We present the details via an example in the ASQP task. The "*in-context samples*" refer to N labeled instances. Here in this example, $N = 1$. The *category list* is a predefined set of aspect categories. Please refer to Figure 3 for the detailed prompts of each other subtask.

(What, How, Why) questions. Cai et al. (2021) define ASQP task by considering implicit expressions of the real world.

Recent advancements in large language models, such as ChatGPT and LLaMA, have drawn significant attention from both the scientific community and the general public. Several studies have demonstrated its universal ability (Susnjak, 2022; King, 2023; Zhang et al., 2022a; Guo et al., 2023; Wei et al., 2023) and well-behaved sentiment analysis (Wang et al., 2023b) capabilities across single ABSA subtasks. However, it remains unclear whether LLMs can still maintain superior performance in more complex compound ABSA tasks. Therefore, we conduct an evaluation of LLMs' performance on five more complex subtasks of compound ABSA.

In these compound extraction subtasks, LLMs' predictions are often unstable, leading to potential issues such as requiring significant human effort to interpret and having out-of-distribution (OOD) in its response. As depicted in Figure 1, we evaluate the ASQP task and observe many failed cases of LLMs. In the first case, the aspect category is predicted to be *"Food"* which is not in the predefined category set. The second case also has an OOD category. In the third case, the *aspect term* is not explicitly mentioned, thereby is `null`. This presents that implicit information is prevalent in these tasks. The above cases all demonstrate that various complexities exist during evaluation. It is very time-consuming to manually identify the results of the LLMs' prediction. Therefore, we can infer that a simple direct evaluation of ABSA using LLMs leads to unstable predictions and cannot fully harness the capabilities of LLMs, which will lead to an unfair assessment. (see §A.5).

To better evaluate the performance of LLMs for compound ABSA tasks, we design ChatABSA, a unified framework that can universally transform five intricate subtasks into the prompting format. Specifically, to limit OOD predictions and format responses, we design *constrained prompts* to build restrictions for subtasks, which can be regarded as conditions for generation. To address weaknesses in predicting implicit information, we design *post-processing* to make use of LLMs' powerful reasoning capabilities. We conduct extensive experiments on five compound ABSA tasks. The main findings are as follows:

- We present an extensive evaluation of LLMs for compound ABSA tasks. The ChatABSA framework makes greater use of LLMs' reasoning ability. Several valuable empirical conclusions are derived, which may provide valuable guidance for future research.

- For all compound ABSA tasks, the evaluation results show that ChatABSA overwhelmingly beats the existing few-shot supervised models.
- ChatABSA can outperform fully supervised methods in some cases. However, in most cases, it underperforms fully supervised methods, and there is still a huge gap between itself and the SOTA method.
- With an in-depth analysis, we found that implicit elements are still challenging and struggling for ChatABSA.

2 ChatABSA

2.1 Formulation and Overview

In a given sentence, there are four types of aspect-level elements: aspect term (at), aspect category (ac), opinion term (ot), and sentiment polarity (sp). In ABSA, the elements at the aspect level to be predicted vary from different subtasks: **Aspect Opinion Pair Extraction (AOPE)** aims to extract aspect terms and their corresponding opinion terms as pairs $\{(at, ot)\}$; **Aspect Category Sentiment Analysis (ACSA)** aims to extract aspect category and their corresponding sentiment polarity as pairs $\{(ac, sp)\}$; **Aspect Sentiment Triplet Extraction (ASTE)** aims to discover more complicated aspect-level triplets $\{(at, ot, sp)\}$; **Target Aspect Sentiment Detection (TASD)** is the task to detect all $\{(at, ac, sp)\}$ triplets for a given sentence; **Aspect Sentiment Quad Prediction (ASQP)** is to predict all aspect-level quadruplets $\{(at, ot, ac, sp)\}$. In TASD and ASQP tasks, if aspect term at (or opinion term ot) is implicit, at (or ot) should be represented by `null`.

Following the prompt engineering of LLMs, we have N (a.k.a. the number of shots) in-context samples with their corresponding ground-truth labels, denoted as $\mathcal{S}$. Given a query sample $\boldsymbol{q}$, ChatABSA aims to detect the compound aspect sentiment elements with the help of $\mathcal{S}$. An example is shown in Figure 2. Firstly, an in-context sample ($N = 1$) with its corresponding ground-truth label $\mathcal{S}$ and a query sample $\boldsymbol{q}$ to be evaluated are first input into the ChatABSA framework. To control LLMs' response format, we design constrained prompts to convert $\mathcal{S}$ and $\boldsymbol{q}$ to templated input. Then, by post-processing, OOD responses are converted to `null`, yielding the predicted output.

2.1.1 Constrained Prompt

To deal with the instability of LLMs' predictions, we manually construct the category prompt $\boldsymbol{p_c}$, the response prompt $\boldsymbol{p_r}$, and the base prompt $\boldsymbol{p_b}$ to let LLMs better understand the nature of the compound ABSA tasks. Then these prompts are combined jointly. Specifically, we demonstrate the prompt templates for each task in Figure 3. This constrained prompt can facilitate automated evaluation of the results. Without the constrained prompt, the model's responses would be inconsistent, affecting the evaluation of the model. §A.5 shows the effectiveness of our prompt strategy.

2.1.2 Post-Processing

As shown in Figure 2, we can observe that LLMs have a powerful reasoning capability. In the query sample *"Not cheap but very yummy"*, we know that *"cheap"* and *"yummy"* describe *"price"* and *"food"*, respectively. LLMs correctly predicted *"price"* (though it made a mistake in the singular-plural form) and *"food"*. However, the query sample $\boldsymbol{q}$ doesn't explicitly mention these two aspect terms. In the ASQP task, the aspect term(s) in the final quadruple results should be directly extracted from the original sentence (rather than inferred from facts). This means that if "price" and "food" do not appear in the original sentence, then they should not appear in the final quadruple results, despite our ability to deduce that the aspect terms are "price" and "food". If the original sentence lacks aspect term(s) (though sometimes we can infer the factual aspect term(s)), then the aspect term(s) in the final quadruple results should be $null$. The handling of opinion term(s) follows a similar logic. Simply put, when the aspect term (at) and opinion term (ot) predicted by LLMs do not appear in the sentence, we set them to $null$. This might lead to inaccurate results. Therefore, for the aspect term and the opinion term, we handle their predictions by post-processing. The formulations are as follows:

$$Quad = \begin{cases} (at, ot, ac, sp), & at, ot \in \boldsymbol{q} \\ (null, ot, ac, sp), & at \notin \boldsymbol{q} \\ (at, null, ac, sp), & ot \notin \boldsymbol{q} \\ (null, null, ac, sp), & at, ot \notin \boldsymbol{q} \end{cases} \tag{1}$$

where these formulations judge whether the predicted elements are explicitly consistent with the span of query $\boldsymbol{q}$. at, ot, ac, and sp are the quadruple results in LLMs responses. This post-processing helps to reveal implicit information.

Task	Response / Preds
AOPE	The aspect-opinion pair consists of *aspect term* and *opinion term*. The aspect-opinion pair of the sentence "{***Example Sentence***}" is {***Example Labels***} What is the aspect-opinion pair of the sentence "{***Sentence***}"? Return with JSON format.
ACSA	The category-sentiment pair consists of *aspect category* and *sentiment polarity*. The *aspect category* is only selected from the following set: {***Aspect Category Lists***} The category-sentiment pair of the sentence "{***Example Sentence***}" is {***Example Labels***} What is the category-sentiment pair of the sentence "{***Sentence***}"? Return with JSON format.
ASTE	The aspect sentiment triplet consists of *aspect term*, *opinion term*, and *sentiment polarity*. The aspect sentiment triplet of the sentence "{***Example Sentence***}" is {***Example Labels***} What is the aspect sentiment triplet of the sentence "{***Sentence***}"? Return with JSON format.
TASD	The aspect sentiment triplet consists of *aspect category*, *aspect term*, and *sentiment polarity*. The *aspect category* is only selected from the following set: {***Aspect Category Lists***} The aspect sentiment triplet of the sentence "{***Example Sentence***}" is {***Example Labels***} What is the aspect sentiment triplet of the sentence "{***Sentence***}"? Return with JSON format.
ASQP	The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, and *sentiment polarity*. The *aspect category* is only selected from the following set: {***Aspect Category Lists***} The aspect sentiment quad of the sentence "{***Example Sentence***}" is {***Example Labels***} What is the aspect sentiment quad of the sentence "{***Sentence***}"? Return with JSON format.

Figure 3: The prompts used for each task. The first line of each task is the base prompt $\boldsymbol{p_b}$. The second line of ACSA, TASD, and ASQP is the category prompt $\boldsymbol{p_c}$. The sentence *"Return with JSON format."* is the response prompt $\boldsymbol{p_r}$. Please refer to Figure 8 for the case study demonstration.

Methods	Rest15	Rest16	Laptop14	Rest14
CMLA+CGCN[1]	55.76	62.70	53.03	63.17
HAST+TOWE*	58.12	63.84	53.41	62.39
JERE-MHS*	59.64	67.65	52.34	66.02
SDRN*	65.75	73.67	66.18	73.30
SpanMlt*	64.68	71.78	68.66	75.60
GTS[1]	68.29	74.31	64.61	74.65
STER[1]	69.3	75.89	67.64	74.96
GAS*	67.93	75.42	69.55	75.15
ESGCN[1]	68.34	75.2	68.69	76.22
SynFue+LAGCN[1]	68.91	76.59	68.88	76.62
QDSL[1]	71.22	77.28	70.2	**78.05**
AOPSS[1]	**72.66**	**78.13**	**70.84**	77.41
IT-MTL(fs-0)	0.00	0.00	0.00	0.00
IT-MTL(fs-1)	5.10	5.89	7.19	8.19
IT-MTL(fs-5)	16.25	15.31	10.08	17.52
IT-MTL(fs-10)	**23.19**	**21.39**	**14.21**	**27.69**
ChatABSA(fs-0)	42.12	47.63	30.50	42.24
ChatABSA(fs-1)	47.67	44.82	35.74	54.24
ChatABSA(fs-5)	50.34	52.72	39.00	53.70
ChatABSA(fs-10)	**52.81**	**54.80**	**43.17**	**55.16**

Table 1: Evaluation results on AOPE in terms of F1 (%) score. The results of baseline methods, marked with * and [1], are obtained from (Zhang et al., 2021b) and (Wang et al., 2023a), respectively. The best results of each part are marked in bold.

Methods	Rest15	Rest16	Laptop15	Laptop16
Cartesian-BERT*	58.42	68.94	32.83	39.54
Pipeline-BERT*	49.35	56.21	43.02	39.42
AddOneDim-BERT*	61.67	69.79	48.94	47.23
Hier-GCN-BERT*	64.23	74.55	62.13	54.15
AAGCN-BERT[1]	**71.75**	**80.77**	**72.39**	**69.68**
IT-MTL(fs-0)	0.00	0.00	0.00	0.00
IT-MTL(fs-1)	12.32	4.13	4.38	0.34
IT-MTL(fs-5)	22.46	19.53	14.71	7.42
IT-MTL(fs-10)	**26.86**	**20.33**	**17.24**	**11.65**
ChatABSA(fs-0)	58.56	64.58	38.34	37.05
ChatABSA(fs-1)	63.47	66.58	42.04	38.85
ChatABSA(fs-5)	64.07	67.79	46.25	39.97
ChatABSA(fs-10)	**66.52**	**71.43**	**48.80**	**41.44**

Table 2: Evaluation results on ACSA in terms of F1 (%) score. The results of baseline methods, marked with * and [1], are obtained from (Cai et al., 2020) and (Liang et al., 2021), respectively. The best results of each part are marked in bold.

3 Experimental Results

We conduct experiments using both ChatGPT and Llama-3. Since the experimental results and conclusions of both models are highly consistent, we primarily focus on analyzing the results of ChatGPT, with ChatABSA defaulting to ChatGPT.

3.1 Aspect Opinion Pair Extraction

The evaluation results of the AOPE task are presented in Table 1. Compared to IT-MTL (Varia et al., 2023), **ChatABSA exhibits more powerful information extraction ability under zero-shot and few-shot settings.** We observe that ChatABSA outperforms IT-MTL by average F1 score improvements of +40.62%, +39.03%, +34.15%, and +29.87% under zero-shot, one-shot, five-shot, and ten-shot, respectively. It is worth noting that compared to IT-MTL(fs-10), ChatABSA(fs-0) also gets absolute F1 score improvements by 18.93%, 26.24%, 16.29%, 14.55% in `Rest15`, `Rest16`, `Laptop14`, `Rest14`, respectively.

However, **ChatABSA lags far behind the fully supervised baselines.** It cannot outperform any method within the fully supervised comparison baselines. ChatABSA (fs-10) underperforms compared to the worst fully supervised baseline CMLA+CGCN. In addition, it has a huge gap compared to the best one AOPSS.

Lastly, we perform the qualitative analysis with four samples (see §B.1) and the element-level analysis (see §C.1) for AOPE. We also evaluate other LLMs (see §D.1 and §3.6) to reveal their capabilities in compound ABSA tasks.

3.2 Aspect Category Sentiment Analysis

The evaluation results of ACSA are shown in Table 2. **ChatABSA still overwhelmingly outperforms IT-MTL under the same few-shot settings.** Compared to IT-MTL(fs-10), ChatABSA(fs-0) gets F1 score improvements by 31.70%, 44.25%, 21.10%, 25.40% in `Rest15`, `Rest16`, `Laptop14`, `Rest14`, respectively.

Different from other compound ABSA tasks, ACSA aims to detect *aspect category* and *sentiment polarity*, which do not explicitly exist in the sentence. It can be observed that, even though some BERT-based methods have been fine-tuned on full training data, ChatABSA demonstrates a more compelling semantic comprehension ability than them. Compared to two fully supervised baselines Cartesian-BERT and Pipeline-BERT, ChatABSA(fs-10) surpasses them by 7.12% and 10.05% in the average F1 score of four datasets. However, **ChatABSA(fs-10) consistently underperforms compared to the best method, AAGCN-BERT.**

Finally, we perform the qualitative analysis with four samples (see §B.2) and the element-level analysis (see §C.2) for ACSA.

3.3 Aspect Sentiment Triplet Extraction

Table 3 presents the evaluation results on ASTE. **ChatABSA consistently beats IT-MTL in the few-shot settings.** Compared to IT-MTL(fs-10), ChatABSA(fs-0) also gets absolute F1 score improvements by 20.08%, 21.47%, 14.58%, 10.84% in `Rest15`, `Rest16`, `Laptop14`, `Rest14`, respectively.

Then, ChatABSA demonstrates notable performance compared to some of the fully supervised methods. Even without in-context samples, ChatABSA(fs-0) acquires absolute F1 score improvements by 5.46% and 6.02% in `Rest15` and `Rest16`, respectively, comparing to CMLA+. ChatABSA(fs-10) also slightly outperforms Pipeline in `Rest16` and `Rest14` datasets.

Methods	`Rest15`	`Rest16`	`Laptop14`	`Rest14`
CMLA+*	37.01	41.72	33.16	42.79
Li-unified-R*	47.82	44.31	42.34	51.00
Pipeline*	52.32	54.21	42.87	51.46
Jet+Bert*	57.53	63.83	51.04	58.14
MvP[1]	**65.89**	**73.48**	**63.33**	**74.05**
IT-MTL(fs-0)	0.00	0.00	0.00	0.00
IT-MTL(fs-1)	9.50	8.25	6.74	7.98
IT-MTL(fs-5)	14.78	16.78	7.46	17.56
IT-MTL(fs-10)	**22.39**	**26.27**	**13.05**	**30.15**
ChatABSA(fs-0)	42.47	47.74	27.63	40.99
ChatABSA(fs-1)	46.72	52.28	29.10	51.23
ChatABSA(fs-5)	47.94	52.36	36.19	53.95
ChatABSA(fs-10)	**48.11**	**56.12**	**42.78**	**54.06**

Table 3: Evaluation results on ASTE in terms of F1 (%) score. The results of baseline methods, marked with * and [1], are obtained from (Zhang et al., 2021b) and (Gou et al., 2023), respectively. The best results of each part are marked in bold.

Even though, **it still meets a huge gap with the SOTA method MvP**.

Moreover, we perform the qualitative analysis with four examples (see §B.3) and the element-level analysis (see §C.3) for ASTE.

3.4 Target Aspect Sentiment Detection

3.4.1 Results Analysis

The evaluation results of TASD are presented in Table 4. **ChatABSA can consistently beat the few-shot supervised method IT-MTL.** Concretely, ChatABSA obtains the average F1 score improvements across the two datasets by 40.25%, 31.57%, 32.00%, and 31.24% in zero-shot, one-shot, five-shot, and ten-shot, respectively.

Unfortunately, **ChatABSA significantly performs worse than the fully supervised methods.** Even ChatABSA(fs-10) lags behind the least one, i.e. TAS-LPM-CRF. A huge gap exists between ChatABSA(fs-10) and the SOTA MvP. A possible reason is that the TASD task introduces implicit information, indicating that *aspect term* may not explicitly exist in the text but is expressed in an obscure manner (see §3.4.2).

Moreover, we perform the qualitative analysis with four examples (see §3.4.3) and the element-level analysis (see §C.4) for TASD.

3.4.2 Implicit Information Prediction

We demonstrate the ability of ChatABSA to predict implicit information in Figure 4 by separately evaluating EA and IA. It can be found that ChatABSA's performance under various shots of in-context samples still has a big gap with GAS in recognizing both EA and IA. In addition, we can see that the

Methods	Rest15	Rest16
TAS-LPM-CRF*	54.76	64.66
TAS-SW-CRF*	57.51	65.89
TAS-SW-TO*	58.09	65.44
GAS*	61.47	69.42
MvP[1]	**64.53**	**72.76**
IT-MTL(fs-0)	0.00	0.00
IT-MTL(fs-1)	8.63	6.76
IT-MTL(fs-5)	12.75	11.30
IT-MTL(fs-10)	**15.08**	**15.37**
ChatABSA(fs-0)	39.21	41.28
ChatABSA(fs-1)	37.23	41.92
ChatABSA(fs-5)	43.00	45.04
ChatABSA(fs-10)	**45.93**	**47.00**

Table 4: Evaluation results on TASD in terms of F1 (%). The results of baseline methods, marked with * and [1], are obtained from (Zhang et al., 2021b) and (Gou et al., 2023), respectively. The best results of each part are marked in bold.

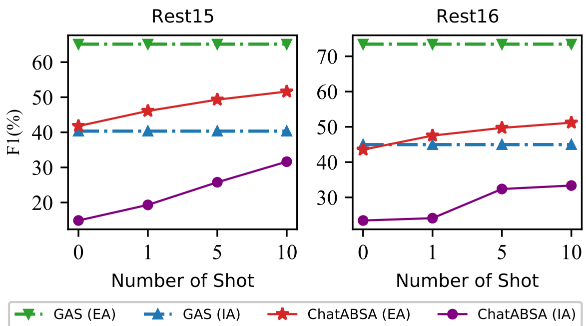

Figure 4: Implicit information prediction in the TASD task. EA and IA denote explicit and implicit aspects, respectively.

evaluation results on EA significantly outperform those on IA. This points out that even for large language models, detecting the implicit information is still challenging. As the number of shots increases, the F1 score on IA is also significantly improved. Thus, it is expected that ChatABSA's performance can be further improved by leveraging more samples. Yet the number of samples is limited by the length of the prompts.

3.4.3 Case Study for TASD

Qualitative analysis is conducted through four test examples with implicit and explicit information, as shown in Figure 12. In the TASD task, the analysis of two different types of test examples is shown in Figure 12. One of the test examples is a sample with an explicit aspect term. Analyzing its explicit aspect term (EA) in the first column under the few-shot setting, it becomes apparent that ChatABSA can accurately determine the aspect term *"service"*. Regarding the two examples in the second column, when involving implicit information, ChatABSA fails to predict precisely under zero-shot settings. It is worth noting that, for IA under the few-shot settings, ChatABSA predicts the implicit expression *"restaurant"*. The triplet is accurately predicted by our post-processing operation. This demonstrates not only the powerful reasoning capability of LLMs but also the success of our post-processing strategy.

3.5 Aspect Sentiment Quad Prediction

3.5.1 Results Analysis

The evaluation results of ASQP are demonstrated in Table 5. **Compared to IT-MTL, despite ChatABSA gains consistent improvements, it still meets challenges in the complex task ASQP.** Firstly, it can be seen that ChatABSA's performance on `Laptop` dataset is relatively worse than other datasets. This shows that it is also struggling with difficult datasets. Secondly, we can observe that in `Rest15` and `Rest16` datasets, the best shot numbers are five and one, respectively. This shows that with the number of shots growing, ChatABSA shows fluctuation rather than gradual improvement. In complex extraction tasks, the semantics of the prompt stays challenging to comprehend for LLMs. We further assume these challenges are imposed by implicit information, which is discussed in §3.5.2.

Then, even with a few samples as prompt, ChatABSA obtains competitive results compared to some of the fully supervised methods, such as ChatABSA(fs-10) and TAS-BERT on both the `Rest15` and `Restaurant` datasets. This shows the superiority of LLMs to some extent. However, compared to MvP, **ChatABSA still meets a huge gap.** Based on the fluctuating results of ChatABSA using 0 to 10 shots, such a gap is difficult to fill by introducing more in-context samples. Thus, we draw an empirical conclusion that **in some cases, small models are still essential even in the recent trends of LLMs emerging and dominating**.

Lastly, we perform the qualitative analysis with four examples (see §B.4), the element-level analysis (see §C.5), and the ablation study (see §A.5) for ASQP. We also evaluate other LLMs (see §D.2 and §3.6) to reveal their capabilities in compound ABSA tasks.

3.5.2 Implicit Information Prediction

To further explore ChatABSA's performance in implicit information prediction, we assess its capabilities on four datasets. We focus on both explicit and

Methods	Rest15			Rest16			Restaurant			Laptop		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
TAS-BERT*	41.86	26.50	32.46	49.37	40.70	44.77	26.29	46.29	33.53	47.15	19.12	27.31
Extract-Classify*	35.64	37.25	36.42	38.40	50.93	43.77	38.54	52.96	44.61	45.56	29.48	35.80
GAS*	45.31	46.70	45.98	54.54	57.62	56.04	57.09	57.51	57.30	43.45	43.29	43.37
Paraphrase*	46.16	47.72	46.93	56.63	59.30	57.93	59.85	59.88	59.87	43.44	42.56	43.00
MvP[1]	-	-	**51.04**	-	-	**60.39**	-	-	**61.54**	-	-	**43.92**
IT-MTL(fs-0)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IT-MTL(fs-1)	14.49	10.15	11.93	15.19	10.35	12.31	13.09	8.51	10.31	13.91	9.79	11.49
IT-MTL(fs-5)	15.41	11.20	12.96	15.64	10.89	12.83	12.03	15.73	12.63	12.58	9.10	10.56
IT-MTL(fs-10)	16.10	14.84	**15.35**	18.73	17.23	**17.84**	13.38	18.99	**15.29**	12.31	10.11	**11.00**
ChatABSA(fs-0)	31.11	24.03	27.11	33.43	27.91	30.42	32.23	24.34	27.74	8.43	6.46	7.31
ChatABSA(fs-1)	26.01	30.69	28.13	32.59	35.21	**33.84**	30.06	34.75	32.19	8.66	10.31	9.39
ChatABSA(fs-5)	30.96	35.93	**33.26**	29.20	35.21	31.92	27.37	31.66	29.34	11.66	15.04	13.13
ChatABSA(fs-10)	29.89	34.76	32.14	30.52	36.59	33.26	31.20	36.43	**33.60**	13.21	17.74	**15.54**

Table 5: Evaluation results on ASQP in terms of precision (Pre, %), recall (Rec, %), and F1 score (F1, %). The results of baseline methods, marked with * and [1], are obtained from (Hu et al., 2022) and (Gou et al., 2023), respectively. The best results of each part are marked in bold.

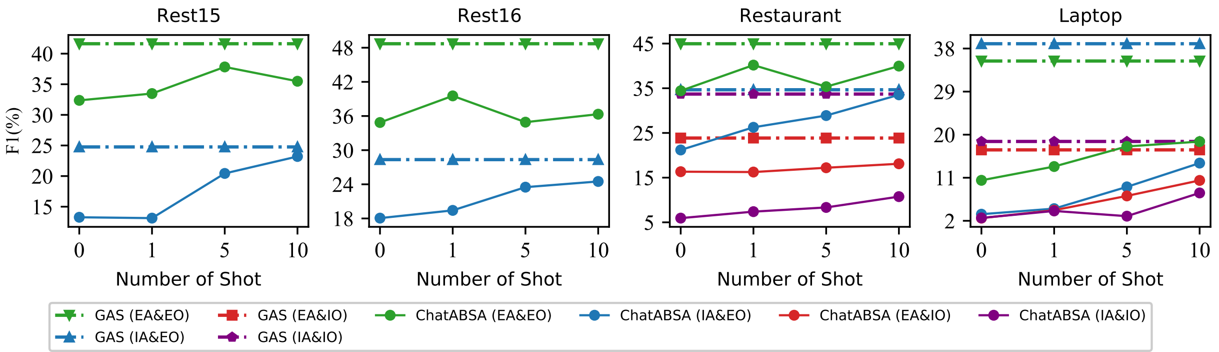

Figure 5: Implicit information prediction in the ASQP task. EA, IA, EO, IO denote explicit aspect, implicit aspect, explicit opinion, implicit opinion, respectively.

implicit information prediction within the ASQP task. The comparisons are shown in Figure 5. The testing set is divided into four subsets that contain various combinations of explicit and implicit aspect/opinion terms, for separate evaluation. These subsets are labeled as EA&EO, IA&EO, EA&IO, and IA&IO.

We can find that the performance of ChatABSA, both in explicit and implicit elements, becomes stronger as the number of shots increases. The F1 scores for implicit information increase more significantly than those for explicit information. In addition, comparing four subsets, it is found that ChatABSA's implicit aspect term (while the opinion term is explicit) predictions at 10-shot are all able to be approximately close to the fully supervised method GAS on the `Rest15`, `Rest16`, and `Restaurant` datasets. Yet, the IA&IO predictions on the `Restaurant` and `Laptop` datasets are the worst and still have a large gap in comparison with GAS.

For a few-shot experiment, the in-context examples are randomly sampled from the whole training set. They may be insufficient for four types, namely EA&EO, IA&EO, EA&IO, and IA&IO. In some cases, the examples may not encompass the specific type of information for the actual query. This may lead to inferior performance of ChatABSA on all four subsets. However, this does not imply that our evaluation method is inappropriate. Here we only take into account the naive few-shot scenario, following Varia et al. (2023). Continuously finding perfectly matched in-context examples for a query is potentially effective for ChatABSA, which guides a promising direction for future research. In addition, **ChatABSA naturally has difficulty to predict implicit elements well.** A possible reason relies on the inherent ambiguity of natural languages. Even though LLMs have learned from a tremendous corpus in the pre-training stage, implicit information requires special knowledge and linguistic background to understand. Demonstrat-

Tasks	Datasets	zero-shot	one-shot	five-shot	ten-shot
ACSA	Rest15	60.76	68.59	70.90	75.89
	Rest16	64.59	69.32	71.16	76.63
	Laptop15	50.31	53.73	55.45	58.61
	Laptop16	44.95	47.97	47.01	53.39
AOPE	Rest14	46.80	53.53	52.79	56.44
	Rest15	39.13	37.58	45.01	47.27
	Rest16	43.45	47.43	47.99	51.32
	Laptop14	30.55	37.60	39.08	38.36
ASTE	Rest14	42.13	49.51	45.90	47.11
	Rest15	35.79	39.74	45.69	44.65
	Rest16	40.18	42.79	46.95	50.13
	Laptop14	25.86	21.40	28.45	29.04
TASD	Rest15	26.91	29.66	34.17	36.00
	Rest16	27.97	34.25	38.18	40.75
ASQP	Rest15	18.50	22.96	23.28	23.50
	Rest16	20.30	15.16	23.98	27.23
	Restaurant	20.63	30.10	33.94	37.30
	Laptop	11.28	10.44	13.24	12.59

Table 6: Evaluation results of Llama-3 on all tasks in terms of F1 score (%). The best results of each part are marked in bold.

ing more in-context samples will better promote its potential for understanding implicit information.

3.6 Results of Llama-3

In fact, our ChatABSA framework is applicable to many other LLMs besides ChatGPT. Therefore, we also apply ChatABSA to Meta-Llama-3-70B-Instruct (AI@Meta, 2024) and conduct comprehensive experiments across all tasks, datasets, and experimental settings to demonstrate the universality of ChatABSA. Experimental results are shown in Table 6. It can be observed that the experimental results of Llama-3 exhibit similar trends to those of ChatGPT, and comparable conclusions can also be drawn from these results.

4 Related Works

4.1 Aspect-Based Sentiment Analysis

Recently, aspect-based sentiment analysis has received extensive attention, including single ABSA tasks, and compound ABSA tasks. Early works focus on single ABSA tasks (Zhang et al., 2022b; Hu et al., 2021; Seoh et al., 2021), such as extracting aspect terms (Chen and Qian, 2020), detecting aspect categories (Bu et al., 2021), and predicting the sentiment polarity for an aspect term (Huang and Carley, 2018) or category (Hu et al., 2019). Recent studies in ABSA aim to produce more comprehensive results by learning compound ABSA tasks. The compound ABSA tasks aim to produce more comprehensive results by simultaneously predicting multiple aspect-level elements: Peng et al. (2020) define ASTE task by corresponding elements with (What, How, Why) questions. Cai et al. (2021) define ASQP task based on ASTE task by considering implicit expressions of the real applications.

In this work, we focus on evaluations of the following five compound ABSA subtasks: Aspect Sentiment Quad Prediction (ASQP) (Zhang et al., 2021a; Cai et al., 2021), Aspect Sentiment Triplet Extraction (ASTE) (Peng et al., 2020), Target Aspect Sentiment Detection (TASD) (Wan et al., 2020), Aspect Opinion Pair Extraction (AOPE) (Zhao et al., 2020; Chen et al., 2020), and Aspect Category Sentiment Analysis (ACSA) (Cai et al., 2020; Liang et al., 2021). Due to compound ABSA tasks having more complexity, we examine whether LLMs can solve them using our designed framework ChatABSA to reliably evaluate its capability on compound ABSA tasks.

4.2 Large Language Model

Thanks to the Transformer architecture, LLMs exhibit amazing emergent abilities by simple instructions and begin to come into people's ordinary life. They usually have a large number of model parameters and are trained on extremely large amounts of raw data, some of LLMs as follows: GPT-3 (Brown et al., 2020), LaMDA (Thoppilan et al., 2022), MT-NLG (Smith et al., 2022), PaLM (Chowdhery et al., 2022), and GPT-4 (OpenAI, 2023).

One of the best-known examples of LLMs is OpenAI's ChatGPT, which has exploded the field of artificial intelligence (AI) and attracted an unprecedented wave of enthusiasm. Its influence can be seen in various fields, including online testing (Susnjak, 2022) and medicine (King, 2023), both of which are experiencing significant growth. Additionally, LLMs have also been utilized in the web domain for various applications including but not limited to automated customer service, and content generation (Biswas, 2023). Its ability to understand and process natural language enables it to help manage and organize web content, and support web development tasks (Fajkovic and Rundberg, 2023), meanwhile, help in enhancing user engagement (Paul et al., 2023).

Recently, Wei et al. (2023) and Wang et al. (2023b) find that LLMs have a strong performance on information extraction and sentiment analysis, respectively. We are particularly curious whether it still maintains such powerful performance for compound ABSA. Inspired by prompt engineering (Dong et al., 2023; Wei et al., 2022), we try to explore its ability to compound ABSA and design ChatABSA to pack a unified framework for more reliable evaluation.

5 Conclusion

In this work, we explore the boundaries of LLMs' capabilities in compound ABSA by comparing fully supervised methods and few-shot supervised methods. Because LLMs' predictions are often unstable, we have designed a more rational framework called ChatABSA. This framework aims to better evaluate LLMs' performance on compound ABSA. ChatABSA exhibits excellent aspect-based sentiment analysis capabilities and overwhelmingly beats few-shot supervised methods under the same few-shot settings. Surprisingly, it can even outperform fully supervised methods in some cases. However, in most cases, it underperforms fully supervised methods, and there is still a huge gap between its performance and the SOTA method. Furthermore, it is still challenging and struggling for LLMs to predict implicit elements. In summary, although LLMs possess strong language comprehension and can accurately follow instructions for specific tasks, it does not perform well on compound ABSA tasks. We hope that our research will inspire future research in LLMs and aspect-based sentiment analysis.

Limitations

We design a new framework ChatABSA to evaluate the performance of LLMs in compound aspect-based sentiment analysis. Despite extensive evaluation under the zero-shot and few-shot settings, our work still has limitations that may guide future work.

Firstly, limitations of model selection. Due to resource constraints, the evaluation of aspect-level extraction capability in language models is limited. As a result, our assessment solely focuses on the gpt-3.5-turbo variant of ChatGPT and the Meta-Llama-3-70B-Instruct of the Llama-3 family. However, the field of language models is rapidly advancing, and there are numerous other notable models such as GPT-4. As a result, a comprehensive aspect-based sentiment analysis capability of more language models will be necessary in the future.

Secondly, limitations to automatic evaluation. Due to limited resources, we use simple prompt engineering and conduct few-shot prompting under a low-resource setting, with no more than 10-shot prompting. However, this approach may not accurately reflect the optimal performance of LLMs on the corresponding downstream tasks.

References

AI@Meta. 2024. Llama 3 model card.

Xiaoyi Bao, Z Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 4044–4050.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, pages 34–45.

Som Biswas. 2023. The function of chat gpt in social media: According to chat gpt. *Available at SSRN 4405389*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2069–2079.

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th international conference on computational linguistics*, pages 833–843.

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 (ACL-IJCNLP), 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.

Zhuang Chen and Tieyun Qian. 2020. Enhancing aspect term extraction with soft prototypes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2107–2117, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Giuseppe D'Aniello, Matteo Gaeta, and Ilaria La Rocca. 2022. Knowmis-absa: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Artificial Intelligence Review*, 55(7):5543–5574.

Edvin Fajkovic and Erik Rundberg. 2023. The impact of ai-generated code on web development: A comparative study of chatgpt and github copilot.

Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518.

Lei Gao, Yulong Wang, Tongcun Liu, Jingyu Wang, Lei Zhang, and Jianxin Liao. 2021. Question-driven span labeling model for aspect–opinion pair extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12875–12883.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

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

Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. 2021. Multi-label few-shot learning for aspect category detection. In *Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. Association for Computational Linguistics (ACL).

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. CAN: Constrained attention networks for multi-aspect sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4601–4610.

Binxuan Huang and Kathleen Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1091–1096.

Michael R King. 2023. The future of ai in medicine: A perspective from a chatbot. *Annals of Biomedical Engineering*, pages 291–295.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6714–6721.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4194–4200.

Bin Liang, Hang Su, Rongdi Yin, Lin Gui, Min Yang, Qin Zhao, Xiaoqi Yu, and Ruifeng Xu. 2021. Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 208–218.

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)*, pages 2215–2225.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. *Preprint*, arXiv:2309.10954.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Justin Paul, Akiko Ueno, and Charles Dennis. 2023. Chatgpt and consumers: Benefits, pitfalls and future research agenda.

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

Joseph J Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. *arXiv preprint arXiv:2211.07743*.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation (SemEval 2016)*, pages 19–30.

Maria Pontiki, Dimitrios Galanis, Harris 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.

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 (JMLR)*, 21(140):1–67.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6311–6322.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *Preprint*, arXiv:2107.02137.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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. *Preprint*, arXiv:2210.06629.

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*, pages 9122–9129.

Chengwei Wang, Tao Peng, Yue Zhang, Lin Yue, and Lu Liu. 2023a. Aopss: A joint learning framework for aspect-opinion pair extraction as semantic segmentation. In *Web and Big Data*, pages 101–113, Cham. Springer Nature Switzerland.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI conference on artificial intelligence*.

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

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. 2021a. Learn from syntax: Improving pair-wise aspect and opinion terms extractionwith rich syntactic knowledge. *arXiv preprint arXiv:2105.02520*.

Shengqiong Wu, Hao Fei, Yafeng Ren, Bobo Li, Fei Li, and Donghong Ji. 2021b. High-order pair-wise aspect and opinion terms extraction with edge-enhanced syntactic graph convolution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2396–2406.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *arXiv preprint arXiv:2010.04640*.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. *arXiv preprint arXiv:2010.02609*.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022a. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

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 (EMNLP)*, pages 9209–9219.

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 (ACL-IJCNLP)*, pages 504–510.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022b. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge & Data Engineering*, pages 1–20.

Yue Zhang, Tao Peng, Ridong Han, Jiayu Han, Lin Yue, and Lu Liu. 2022c. Synchronously tracking entities and relations in a syntax-aware parallel architecture for aspect-opinion pair extraction. *Applied Intelligence*, 52(13):15210–15225.

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.

Yan Zhou, Wei Jiang, Po Song, Yipeng Su, Tao Guo, Jizhong Han, and Songlin Hu. 2020. Graph convolutional networks for target-oriented opinion words extraction with adversarial training. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

A Experimental Settings

A.1 Datasets

To evaluate the potential of LLMs on compound ABSA, we select eight publicly available datasets. They all originate from the challenges of SemEval. The statistics of datasets are shown in Table 7.

AOPE To evaluate ChatABSA, four datasets including *Laptop* and *Restaurant* domains are selected from Semeval 2014 Task 4, Semeval 2015 Task 12, and Semeval 2016 Task 5. `Rest16` is provided by (Fan et al., 2019), where the at and ot pairs are annotated. In addition, other datasets are provided by (Wang et al., 2016, 2017).

ACSA In the ACSA task, four benchmark datasets are selected. These datasets from Semeval 2015 (Pontiki et al., 2015) (`Rest15` and `Laptop15`) and Semeval 2016 (Pontiki et al., 2016) (`Rest16` and `Laptop16`) consisting of two domains and each domain includes two datasets, i.e., two Restaurant domain datasets (`Rest15` and `Rest16`) and two Laptop domain datasets (`Laptop15` and `Laptop16`).

ASTE In the ASTE task, the four public datasets are based on (Fan et al., 2019), which have already annotated opinion terms. Peng et al. (2020) also label sentiment to form tuple (at, ot, sp).

TASD Experiments are performed on two restaurant domain datasets, namely `Rest15` from SemEval-2015 Task 12 (Pontiki et al., 2015) and `Rest16` (Pontiki et al., 2016) from SemEval-2016 Task 5.

ASQP In the ASQP task, there are four publicly available datasets: `Rest15`, `Rest16`, `Restaurant` and `Laptop`. `Rest15` and `Rest16` are annotated by Zhang et al. (2021a) based on Semval tasks (Pontiki et al., 2015, 2016); Cai et al. (2021) propose `Restaurant` and `Laptop`. `Restaurant` is based on SemEval 2016 Restaurant (Pontiki et al., 2016) and the extension of SemEval 2016 Restaurant (Fan et al., 2019; Xu et al., 2020). The Laptop dataset is annotated by Cai et al. (2021) based on Amazon 2017 and 2018.

Task	Datasets	Train		Test		Dev	
		#S	#E	#S	#E	#S	#E
AOPE	Rest14	1462	2383	500	864	163	260
	Rest15	678	969	325	436	76	107
	Rest16	971	1357	328	457	108	155
	Laptop14	1035	1485	343	482	116	149
ACSA	Rest15	1102	1451	572	761	0	0
	Rest16	1680	2216	580	735	0	0
	Laptop15	1397	1970	644	947	0	0
	Laptop16	2037	2903	572	797	0	0
ASTE	Rest14	1266	2338	492	994	310	577
	Rest15	605	1013	322	485	148	249
	Rest16	857	1394	326	514	210	339
	Laptop14	906	1460	328	543	219	346
TASD	Rest15	1120	1654	582	845	10	13
	Rest16	1708	2507	587	859	29	44
ASQP	Rest15	834	1354	537	795	209	347
	Rest16	1264	1989	544	799	316	507
	Restaurant	2934	4172	816	1161	326	440
	Laptop	1530	2484	583	916	171	261

Table 7: Data statistics. #S and #E denote the number of sentences and tuples, respectively.

A.2 Compared Methods

We choose the following two types of comparison baselines for each task: 1) **fully supervised methods**, which are trained using the full set of training data; 2) **few-shot supervised methods**, which are trained with a few training data. The second type is designed to convert the elements to be predicted into a target sequence by a pre-defined template. For few-shot supervised methods and ChatABSA, we set the number of shots to few-shot (fs) 0, 1, 5, and 10.

AOPE For fully supervised baselines, we choose the current state-of-the-art, including AOPSS (Wang et al., 2023a) and some other strong baselines CMLA+CGCN (Wang et al., 2017; Zhou et al., 2020), HAST+TOWE (Li et al., 2018; Fan et al., 2019), JERE-MHS (Bekoulis et al., 2018), SDRN (Chen et al., 2020), SpanMlt (Zhao et al., 2020), GTS (Wu et al., 2020), STER (Zhang et al., 2022c), GAS (Zhang et al., 2021b), ESGCN (Wu et al., 2021b), SynFue+LAGCN (Wu et al., 2021a), QDSL (Gao et al., 2021). For few-shot supervised baselines, we select IT-MTL (Varia et al., 2023), which is the first to address and formulate the few-shot ABSA problem. Varia et al. (2023) fine-tune a T5 model (Raffel et al., 2020) incorporated with instructional prompts in a multi-task learning fashion covering all the subtasks, including the entire quadruple prediction task.

ACSA For fully supervised baselines, we select the current SOTA method AAGCN-BERT (Liang et al., 2021) and some other BERT-based methods, including Cartesian-BERT (Cai et al., 2020), Pipeline-BERT (Cai et al., 2020), ADDOneDim-BERT (Cai et al., 2020), Hier-GCN-BERT (Cai et al., 2020). The few-shot supervised baseline method is IT-MTL (Varia et al., 2023).

ASTE For fully supervised baselines, we choose CMLA+ (Wang et al., 2017), Li-unified-R (Li et al., 2019), Pipeline (Peng et al., 2020), Jet+BERT (Xu et al., 2020) and the current SOTA model MvP (Gou et al., 2023). For few-shot supervised baselines, we choose IT-MTL (Varia et al., 2023).

TASD For fully supervised baselines, we select the current SOTA model MvP (Gou et al., 2023) and some other methods TAS-LPM-CRF, TAS-SW-CRF, TAS-SW-CRF (Wan et al., 2020), and GAS (Zhang et al., 2021b). The few-shot supervised baseline is IT-MTL (Varia et al., 2023).

ASQP For fully supervised baselines, we adopt the current SOTA model MvP (Gou et al., 2023) and some other baselines, TAS-BERT (Wan et al., 2020), Extract-Classify (Cai et al., 2021), Paraphrase (Zhang et al., 2021a), and GAS (Zhang et al., 2021b). The few-shot supervised baseline is IT-MTL (Varia et al., 2023).

A.3 Evaluation Metrics

We adopt accuracy, recall, and F1 scores for the ASQP task. For AOPE, ACSA, ASTE, and TASD tasks, we employ the F1 score. For above all tasks, an aspect-sentiment tuple is regarded as correct if and only if exactly the same as the corresponding ground-truth label. For few-shot experiments, to eliminate the impact of sampling examples on the results, all the experimental results are the average of 3 runs, following Milios et al. (2023).

A.4 Usage of ChatGPT

The ChatGPT used for evaluation is a variant of GPT3.5, specifically using the gpt-3.5-turbo version and setting the temperature to 0. For ChatGPT response generation, whether it's zero-shot or few-shot, we do not need to manually observe and record the results, instead, obtain its predicted results through automated code searching.

A.5 Ablation Study

The above subsections have provided an extensive evaluation of ChatABSA on 4 datasets. However, the effects of its individual components remain unclear. Thus, a systematic ablation study based on one-shot is performed in the ASQP task. We specifically design the following variations:

- **-PP** means removal of post-processing.
- **-ACP** means the removal of the aspect category prompt from the constrained prompt.
- **-RP+Table** means that the response prompt has been changed from a JSON response to a Table response ("*Respond in the form of a table with four columns and a header of (aspect category, aspect term, sentiment polarity, opinion term)*").
- **-PP+NP** is to use a constrained prompt ("*If no aspect term (or opinion term) is presented in the given sentence, aspect term (or opinion term) will be* `null`.") instead of post-processing.

Datasets	Model	Pre	Rec	F1
Rest15	**Our**	**26.01**	**30.69**	**28.13**
	-ACP	10.47	11.65	11.02
	-PP	24.71	29.26	26.77
	-RP+Table	22.46	24.81	23.53
	-PP+NP	24.79	18.64	22.47
Rest16	**Our**	**32.59**	**35.21**	**33.84**
	-ACP	10.42	10.76	10.59
	-PP	29.85	32.58	31.14
	-RP+Table	28.29	34.67	31.16
	-PP+NP	24.26	24.78	24.52
Restaurant	**Our**	**30.06**	**34.75**	**32.19**
	-ACP	11.92	13.46	12.64
	-PP	27.81	32.17	29.80
	-RP+Table	27.93	31.00	29.38
	-PP+NP	23.93	29.43	26.40
Laptop	**Our**	**8.66**	**10.31**	**9.39**
	-ACP	0.94	1.12	1.02
	-PP	7.76	9.24	8.42
	-RP+Table	8.43	6.46	7.31
	-PP+NP	6.22	7.58	6.83

Table 8: Evluation results of ablation study. The minus "-" denotes removing components.

The experimental results are presented in Table 8. It is found that, by removing various parts of ChatABSA, the results descend in all four datasets. It is additional evidence that our designed constrained prompts and post-processing are effective.

First, **-ACP** makes the prediction significantly less effective, which shows that the category prompt successfully influences the prediction result. Similarly, **-PP** also suggests that our post-processing method is particularly good at handling implicit information. Finally, as for the response format, the response in JSON format (**Our**) is always better than in Table format (**-RP+Table**) in all four datasets. Therefore, it can be inferred that the JSON form makes it easier to find the correct tuples than the Table form. We believe that the main reason the JSON format is better than the table format is its relative simplicity; the table format tends to be more complex in comparison.

The ablation study indicates that the ChatABSA framework assists in guiding LLMs to produce desired outcomes and facilitates proper processing and rational evaluation of LLMs' outputs. This ensures a more objective assessment of LLMs, preventing underestimation of their capabilities due to poor prompts or inappropriate handling of their outputs.

B Case Study

B.1 Case Study for AOPE

A quantitative analysis is performed with two test examples. The specific results are shown in Figure 9, where we show examples originating from different datasets, `Rest15` and `Laptop14`, and their results under zero-shot and few-shot settings, respectively.

The first sentence describes an aspect term, "*mens bathroom*" rather than "*bathroom*". However, under zero-shot settings, ChatABSA incorrectly predicts it as "*bathroom*". Similarly, in another test example "*I also wanted Windows 7, which this one has.*", the opinion term extracted by ChatABSA is "*this one has*", which is not the opinion term corresponding to the aspect term "*Windows 7*". Fortunately, they are all predicted accurately in the few-shot setting. This also shows that both aspect term and opinion term cannot be predicted accurately in the zero-shot scenario, and some demonstration examples are needed to accurately predict individual elements.

B.2 Case Study for ACSA

In the ACSA task, two instances are selected from different datasets for qualitative analysis. As shown in Figure 10, without any example of the ACSA task, ChatABSA is more prone to errors: in the first case, incorrectly predicting category as *"food general"*; in the second case, incorrectly predicting category as *"restaurant miscellaneous"*. ChatABSA is able to achieve substantial performance improvements with only one demonstration example, and both test instances have their errors corrected under the one-shot settings.

B.3 Case Study for ASTE

Similarly, two test examples are selected for qualitative analysis and the results are shown in Figure 11. In the first example, ChatABSA fails to identify the triplet in the zero-shot. It suggests that there is no aspect sentiment triplet in this sentence. In the other example, the sentiment polarity in the sentence cannot be accurately identified. However, with a demonstration example, it can accurately identify the aspect sentiment triplet. We believe that this happens due to LLMs' deficiency in understanding the nature of the ASTE task under zero-shot. It is possible to better understand the nature of the ASTE task through the demonstration example.

Element	Shot	`Rest15`	`Rest16`	`Laptop14`	`Rest14`
Aspect Term	0-shot	66.88	68.97	58.15	72.89
	1-shot	71.61	70.54	59.75	76.29
	5-shot	72.99	74.00	64.48	77.93
	10-shot	73.23	75.36	65.92	77.99
Opinion Term	0-shot	53.05	56.98	40.40	51.77
	1-shot	57.60	52.44	47.10	62.68
	5-shot	60.77	61.35	47.49	61.08
	10-shot	62.69	62.87	47.29	62.23

Table 9: Analysis at element-level for AOPE in terms of F1 (%) score.

B.4 Case Study for ASQP

Similar to the TASD task, four prediction examples are selected from the perspective of implicit or explicit information. EA&EO is an explicit aspect term and an explicit opinion term; IA&EO is an implicit aspect term and an explicit opinion term; EA&IO is an explicit aspect term and an implicit opinion term; IA&IO is an implicit aspect term and an implicit opinion term.

Results for EA&EO, IA&EO, EA&IO, and IA&IO, as well as test examples, are shown in Figures 13 and 14. For the EA&EO prediction example, we find that ChatABSA incorrectly predicts the sentiment polarity as *"positive"*. It seems that there is an error in the understanding of *"fair"*. However, under few-shot settings, this understanding error is resolved. In the IA&EO scenario, it misunderstands the nature of the ASQP task and incorrectly predicts the aspect term as the opinion term and the opinion term as the aspect term. In the EA&IO scenario, the test example implicitly expresses an opinion, where *"not the place"* potentially means *"can't eat in the first place"*, and ChatABSA incorrectly predicts it as *"not"*, so its inference of implicit opinion is wrong. However, with the showing example, it is able to predict the implicit information accurately. Finally, in the IA&IO scenario, ChatABSA also fails to infer the implicit information in the sentence. However, under the few-shot scenario, it is able to predict the implicit information accurately. This also shows that the accurate prediction of implicit opinion requires in-context samples.

Element	Shot	Rest15	Rest16	Laptop15	Laptop16
Aspect Category	0-shot	65.04	69.49	45.09	41.07
	1-shot	69.45	71.08	47.09	41.35
	5-shot	69.58	72.56	51.46	41.96
	10-shot	72.09	75.94	54.27	45.59
Sentiment Polarity	0-shot	82.88	84.08	77.76	73.51
	1-shot	86.75	88.90	86.65	86.55
	5-shot	88.30	91.39	89.36	87.83
	10-shot	89.83	91.85	89.81	88.85

Table 10: Analysis at element-level for ACSA in terms of F1 (%) score.

Element	Shot	Rest15	Rest16	Laptop14	Rest14
Aspect Term	0-shot	70.81	74.71	58.85	73.58
	1-shot	70.88	71.58	61.24	74.50
	5-shot	72.27	73.31	65.42	76.63
	10-shot	73.38	74.97	66.98	77.94
Opinion Term	0-shot	53.66	57.88	40.93	50.94
	1-shot	61.72	66.16	42.10	62.36
	5-shot	63.91	64.81	51.51	66.19
	10-shot	62.52	67.93	56.05	65.48
Sentiment Polarity	0-shot	85.80	87.22	79.01	87.66
	1-shot	85.22	85.17	80.82	86.22
	5-shot	85.54	88.21	82.31	89.01
	10-shot	86.90	88.09	82.57	88.60

Table 11: Analysis at element-level for ASTE in terms of F1 (%) score.

C Additional Experimental Results

C.1 Analysis at Element-Level for AOPE

To further explore the effect of each different element for ChatABSA, the analysis at element-level for AOPE is performed, and the results of which are presented in Table 9. We can find that the average increase in F1 score from 0-shot to 10-shot is notable, by 6.41% for the aspect term and 8.22% for the opinion term. It can be found that ChatABSA has a large number of incorrect judgments without any displayed examples, but such errors are gradually eliminated as the number of displayed examples grows.

C.2 Analysis at Element-Level for ACSA

ChatABSA has shown impressive performance in the ACSA task, and to further understand its ability to predict each component element, an element-level analysis is performed. As shown in Table 10, It can be found that ChatABSA's performance in predicting the aspect category depends on the number of categories. Concretely, as the number of categories increases, the performance of ChatABSA sharply decline . As for the datasets in the restaurant and laptop domains, the number of categories is 30 for the `Rest` and 198 for the `Laptop`, respectively. We can observe that it has better results

Element	Shot	Rest15	Rest16
Aspect Term	0-shot	65.86	64.52
	1-shot	63.18	65.66
	5-shot	68.53	68.63
	10-shot	70.98	70.56
Aspect Category	0-shot	70.41	71.48
	1-shot	71.50	70.58
	5-shot	70.78	72.34
	10-shot	73.22	73.33
Sentiment Polarity	0-shot	84.84	88.80
	1-shot	86.01	89.64
	5-shot	87.40	90.56
	10-shot	88.63	89.23

Table 12: Analysis at element-level for TASD in terms of F1 (%) score.

Element	Shot	Rest15	Rest16	Restaurant	Laptop
Aspect Term	0-shot	60.99	64.52	59.47	52.41
	1-shot	59.17	61.07	61.06	50.36
	5-shot	65.25	66.35	62.56	57.16
	10-shot	67.74	69.17	67.75	59.45
Opinion Term	0-shot	48.48	54.16	51.49	41.70
	1-shot	51.96	53.10	58.20	46.44
	5-shot	53.71	53.64	52.88	50.64
	10-shot	49.41	54.65	54.88	48.86
Aspect Category	0-shot	67.34	69.15	68.79	33.32
	1-shot	66.95	69.55	67.76	35.94
	5-shot	71.34	67.35	70.56	39.43
	10-shot	71.87	70.19	72.76	45.30
Sentiment Polarity	0-shot	85.89	86.32	84.89	84.41
	1-shot	85.49	86.35	84.00	84.89
	5-shot	87.85	88.71	86.47	86.40
	10-shot	88.59	89.98	87.01	87.09

Table 13: Analysis at element-level for ASQP in terms of F1 (%) score.

in the `Rest` domain than in the `Laptop` domain. In addition, ChatABSA's prediction in sentiment polarity is better than that in aspect category.

C.3 Analysis at Element-Level for ASTE

A similar exploration is also performed in the ASTE task, the results of which are shown in Table 11. To our surprise, ChatABSA's prediction results for the aspect term seem to fluctuate slightly in the domain of restaurant, with an average growth of 2.40% on the `Rest14`, `Rest15` and `Rest16` datasets. It can be conjectured that ChatABSA is naturally able to identify commonly occurring aspect terms, but for some less common ones, it needs some display examples to make accurate predictions. The performance of ChatABSA in predicting the opinion term still fluctuates considerably. On the contrary, its performance fluctuation in predicting sentiment is not very notable. In the `Rest14` dataset, we can find that the prediction of sentiment polarity decreases in different degrees from 0-shot to 1-shot and from 5-shot to 10-shot, respectively. From this, it can be inferred that ChatABSA is naturally able to determine the sentiment polarity in sentences, but some demonstration examples are still needed to accurately extract cues for the sentiment polarity and the opinion terms.

C.4 Analysis at Element-Level for TASD

In the TASD task, the results of the element-level analysis are shown in Table 12. As for the aspect term prediction, the increase from 0-shot to 10-shot is still notable. This indicates that, in the TASD task, ChatABSA's natural perception of the aspect term is not very strong, and its ability to predict the aspect term is gradually enhanced as the number of shots increases. Moreover, the results of the sentiment analysis have shown that the prediction ability remains stable. For the aspect category prediction, ChatABSA demonstrates a strong capability, even surpassing its ability to predict aspect terms. Moreover, as the shot number increases, the prediction accuracy for the aspect category also improves.

C.5 Analysis at Element-Level for ASQP

The element-level results of the ASQP task are shown in Table 13. First, ChatABSA's ability to predict aspect terms from 0-shot to 10-shot is improved in all four datasets. Second, as for the ability to predict opinion terms, it can be found that it does not improve much in the `Restaurant` domain from 0-shot to 10-shot, but it improves significantly in the `Laptop` domain, with an increase of 7.16% in F1. Its ability to predict the aspect category from 0-shot to 10-shot is also improved to various extents. It is worth noting that, in the `Laptop` dataset, the improvement is very significant, with an increase of 11.98% in the F1 score. We conclude that one of the main reasons for this is that the number of predefined categories in the `Laptop` dataset is so large that ChatABSA cannot accurately determine which one is the correct one without sufficient display examples. Finally, ChatABSA's ability to predict the sentiment polarity is also consistently improved in four datasets.

D Results of Other LLMs

D.1 Results of Other LLMs in AOPE

In the AOPE task, we chose two other LLMs (ERNIE-Bot and Llama-2) for evaluation.

Methods	`Rest15`	`Rest16`	`Laptop14`	`Rest14`
ERNIE-Bot(fs-0)	9.82	8.84	7.21	11.49
ERNIE-Bot(fs-1)	27.39	34.15	24.36	33.29
ERNIE-Bot(fs-5)	25.47	**37.13**	**25.82**	31.65
ERNIE-Bot(fs-10)	**30.68**	36.95	23.44	**33.96**
ChatABSA(fs-0)	42.12	47.63	30.50	42.24
ChatABSA(fs-1)	47.67	44.82	35.74	54.24
ChatABSA(fs-5)	50.34	52.72	39.00	53.70
ChatABSA(fs-10)	**52.81**	**54.80**	**43.17**	**55.16**

Table 14: Evaluation results of ChatABSA and ERNIE-Bot on AOPE in terms of F1 (%) score. The best results of each part are marked in bold.

D.1.1 Results of ERNIE-Bot in AOPE

ERNIE-Bot (Sun et al., 2021), developed by Baidu, is an advanced conversational AI model that excels in understanding and generating human-like responses. Specifically, we selected the ERNIE-Bot-turbo version for evaluation, with the experimental results shown in Table 14.

Compared to ChatABSA, ERNIE-Bot exhibits less powerful information extraction ability under zero-shot and few-shot settings. We observe that ERNIE-Bot underperforms ChatABSA by average F1 score deteriorations of -31.28%, -15.82%, -18.92%, and -20.23% under zero-shot, one-shot, five-shot, and ten-shot, respectively. It is worth noting that compared to ChatABSA(fs-0), ERNIE-Bot(fs-10) also gets absolute F1 score deteriorations by -11.44%, -10.68%, -7.06%, -8.28% in `Rest15`, `Rest16`, `Laptop14`, `Rest14`, respectively.

Observing ERNIE-Bot's experimental results, there is a significant leap in the F1 scores from zero-shot to one-shot. ERNIE-Bot exhibits very poor performance in the zero-shot setting. From this, we infer that ERNIE-Bot struggles to understand the nature of AOPE without examples (it cannot comprehend the nature of AOPE from just descriptive text about the AOPE task). This suggests that ERNIE-Bot's natural language understanding capabilities may be far inferior to ChatGPT, which is enlightening for future research.

D.1.2 Results of Llama-2 in AOPE

Llama-2 (Touvron et al., 2023) is a large language model developed by Meta AI. It's designed for processing and generating text, offering advanced capabilities in understanding and responding to a wide array of language tasks. Specifically, we selected the Llama-2-70B-Chat version for evaluation, with the experimental results shown in Table 15. Compared to ChatABSA, Llama-2 exhibits

Methods	Rest15	Rest16	Laptop14	Rest14
Llama-2(fs-0)	8.06	12.66	10.76	16.02
Llama-2(fs-1)	27.68	36.94	20.02	40.02
Llama-2(fs-5)	36.82	39.77	**32.18**	45.86
Llama-2(fs-10)	**37.94**	**41.00**	32.01	**48.11**
ChatABSA(fs-0)	42.12	47.63	30.50	42.24
ChatABSA(fs-1)	47.67	44.82	35.74	54.24
ChatABSA(fs-5)	50.34	52.72	39.00	53.70
ChatABSA(fs-10)	**52.81**	**54.80**	**43.17**	**55.16**

Table 15: Evaluation results of ChatABSA and Llama-2 on AOPE in terms of F1 (%) score. The best results of each part are marked in bold.

less powerful information extraction ability under zero-shot and few-shot settings. Additionally, in terms of zero-shot performance, Llama-2 exhibits a similar trend to ERNIE-Bot and yields comparable experimental results, leading to the same conclusions.

Compared to ChatABSA, Llama-2 exhibits less powerful information extraction ability under zero-shot and few-shot settings. We observe that Llama-2 underperforms ChatABSA by average F1 score deteriorations of -28.75%, -14.45%, -10.28%, and -11.72% under zero-shot, one-shot, five-shot, and ten-shot, respectively. It is worth noting that compared to ChatABSA(fs-0), Llama-2(fs-10) also gets absolute F1 score deteriorations by -4.18%, -6.63%, +1.51%, +5.87% in `Rest15`, `Rest16`, `Laptop14`, `Rest14`, respectively.

Observing Llama-2's experimental results, there is a significant leap in the F1 scores from zero-shot to one-shot. Llama-2 exhibits very poor performance in the zero-shot setting. From this, we infer that Llama-2 struggles to understand the nature of AOPE without examples (it cannot comprehend the nature of AOPE from just descriptive text about the AOPE task). This suggests that Llama-2's natural language understanding capabilities may be far inferior to ChatGPT, which is enlightening for future research.

D.2 Results of Other LLMs in ASQP

In the ASQP task, we chose two other LLMs (ERNIE-Bot-turbo and Llama-2-70B-Chat) for evaluation. The experimental results can be seen in Table 16 and 17. The two LLMs demonstrate the same experimental conclusions on the ASQP task as they do on the AOPE task.

D.2.1 Results of ERNIE-Bot in ASQP

Specifically, we selected the ERNIE-Bot-turbo version for evaluation, with the experimental results shown in Table 16.

Methods	Rest15	Rest16	Restaurant	Laptop
ERNIE-Bot(fs-0)	6.06	8.87	8.77	**2.90**
ERNIE-Bot(fs-1)	9.72	5.12	14.10	1.23
ERNIE-Bot(fs-5)	**12.50**	**11.13**	**20.72**	1.87
ERNIE-Bot(fs-10)	11.07	10.88	17.65	1.76
ChatABSA(fs-0)	27.11	30.42	27.74	7.31
ChatABSA(fs-1)	28.13	**33.84**	32.19	9.39
ChatABSA(fs-5)	**33.26**	31.92	29.34	13.13
ChatABSA(fs-10)	32.14	33.26	**33.60**	**15.54**

Table 16: Evaluation results of ChatABSA and ERNIE-Bot on ASQP in terms of F1 (%) score. The best results of each part are marked in bold.

Methods	Rest15	Rest16	Restaurant	Laptop
Llama-2(fs-0)	8.90	12.58	14.19	3.83
Llama-2(fs-1)	**19.59**	13.00	**31.57**	4.73
Llama-2(fs-5)	18.05	21.70	28.52	6.78
Llama-2(fs-10)	18.14	**22.08**	28.93	**6.99**
ChatABSA(fs-0)	27.11	30.42	27.74	7.31
ChatABSA(fs-1)	28.13	**33.84**	32.19	9.39
ChatABSA(fs-5)	**33.26**	31.92	29.34	13.13
ChatABSA(fs-10)	32.14	33.26	**33.60**	**15.54**

Table 17: Evaluation results of ChatABSA and Llama-2 on ASQP in terms of F1 (%) score. The best results of each part are marked in bold.

Compared to ChatABSA, ERNIE-Bot exhibits less powerful information extraction ability under zero-shot and few-shot settings. We observe that ERNIE-Bot underperforms ChatABSA by average F1 score deteriorations of -16.50%, -18.35%, -15.36%, and -18.30% under zero-shot, one-shot, five-shot, and ten-shot, respectively. It is worth noting that compared to ChatABSA(fs-0), ERNIE-Bot(fs-10) also gets absolute F1 score deteriorations by -16.04%, -19.54%, -10.09%, -5.55% in `Rest15`, `Rest16`, `Restaurant`, `Laptop`, respectively.

Observing ERNIE-Bot's experimental results, there is a significant leap in the F1 scores from zero-shot to one-shot. ERNIE-Bot exhibits very poor performance in the zero-shot setting. From this, we infer that ERNIE-Bot struggles to understand the nature of ASQP without examples (it cannot comprehend the nature of ASQP from just descriptive text about the ASQP task). This suggests that ERNIE-Bot's natural language understanding capabilities may be far inferior to ChatGPT, which is enlightening for future research.

D.2.2 Results of Llama-2 in ASQP

Specifically, we selected the Llama-2-70B-Chat version for evaluation, with the experimental results shown in Table 17.

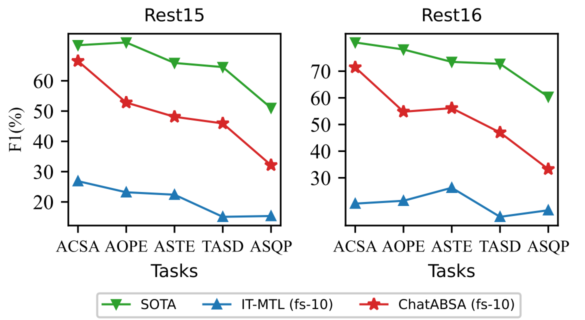

Figure 6: Performance trends of supervised, IT-MTL, and ChatABSA methods across increasing task complexities. "SOTA" refers to the best-performing method among the fully supervised approaches.

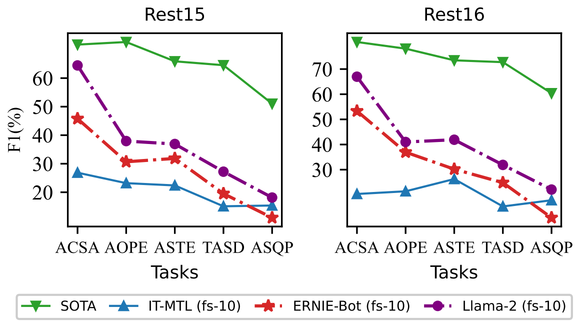

Figure 7: Performance trends of supervised, IT-MTL, ERNIE-Bot, and Llama-2 methods across increasing task complexities. "SOTA" refers to the best-performing method among the fully supervised approaches.

Compared to ChatABSA, Llama-2 exhibits less powerful information extraction ability under zero-shot and few-shot settings. We observe that Llama-2 underperforms ChatABSA by average F1 score deteriorations of -13.27%, -8.67%, -8.15%, and -9.60% under zero-shot, one-shot, five-shot, and ten-shot, respectively. It is worth noting that compared to ChatABSA(fs-0), Llama-2(fs-10) also gets absolute F1 score deteriorations by -8.97%, -8.34%, +1.19%, -0.32% in `Rest15`, `Rest16`, `Restaurant`, `Laptop`, respectively.

Observing Llama-2's experimental results, there is a significant leap in the F1 scores from zero-shot to one-shot. Llama-2 exhibits very poor performance in the zero-shot setting. From this, we infer that Llama-2 struggles to understand the nature of ASQP without examples (it cannot comprehend the nature of ASQP from just descriptive text about the ASQP task). This suggests that Llama-2's natural language understanding capabilities may be far inferior to ChatGPT, which is enlightening for future research.

E Comparing Methods across Tasks

The five tasks vary in difficulty and are ranked in order of increasing complexity as follows: ACSA, AOPE, ASTE, TASD, ASQP. We observed that the performance of all methods decreases to varying degrees as task difficulty increases, when using fully supervised methods, IT-MTL, and ChatABSA. However, the extent of the performance decline significantly differs among these three approaches.

Figure 6 details the performance variations of these three methods on the Rest15 and Rest16 datasets as task difficulty increases. By comparing the curves in the figure, we observe that although the performance of all models declines with increasing task complexity, the fully supervised small models and IT-MTL exhibit a smaller decrease in performance when facing complex tasks compared to ChatABSA. This observation suggests that fully supervised small models and few-shot methods have a relative advantage in handling complex tasks.

Additionally, while LLMs exhibits high versatility, our results indicate that fully supervised methods still excel in certain specific complex tasks, and this advantage becomes more pronounced as task complexity increases.

We observed similar conclusions on other large language models as well, as shown in Figure 7. Specifically, in the ASQP task, the performance of ERNIE-Bot was even inferior to that of IT-MTL.

Example Sentence	*Gross food – Wow -*
Example Labels (at, ac, sp, ot)	*[{"aspect_category": "food quality", "aspect_term": "food", "opinion_term": "gross", "sentiment_polarity": "negative"}]*
Sentence	*Great Indian food*

Task	**Response / Preds**
AOPE	The aspect-opinion pair consists of *aspect term* and *opinion term*. The aspect-opinion pair of the sentence "***Gross food – Wow -***" is *[{"aspect_term": "food", "opinion_term": "gross"}]* What is the aspect-opinion pair of the sentence "***Great Indian food***"? Return with JSON format.
ACSA	The category-sentiment pair consists of *aspect category* and *sentiment polarity*. The *aspect category* is only selected from the following set: ['service general', 'ambience general', 'restaurant miscellaneous', 'food quality', 'restaurant prices', 'drinks quality', 'restaurant general', 'food prices', 'drinks prices', 'drinks style_options', 'food style_options', 'location general', 'food general'] The category-sentiment pair of the sentence "***Gross food – Wow -***" is *[{"aspect_category": "food quality", "sentiment_polarity": "negative"}]* What is the category-sentiment pair of the sentence "***Great Indian food***"? Return with JSON format.
ASTE	The aspect sentiment triplet consists of *aspect term*, *opinion term*, and *sentiment polarity*. The aspect sentiment triplet of the sentence "***Gross food – Wow -***" is *[{"aspect_term": "food", "opinion_term": "gross", "sentiment_polarity": "negative"}]* What is the aspect sentiment triplet of the sentence "***Great Indian food***"? Return with JSON format.
TASD	The aspect sentiment triplet consists of *aspect category*, *aspect term*, and *sentiment polarity*. The *aspect category* is only selected from the following set: ['service general', 'ambience general', 'restaurant miscellaneous', 'food quality', 'restaurant prices', 'drinks quality', 'restaurant general', 'food prices', 'drinks prices', 'drinks style_options', 'food style_options', 'location general', 'food general'] The aspect sentiment triplet of the sentence "***Gross food – Wow -***" is *[{"aspect_category": "food quality", "aspect_term": "food", "sentiment_polarity": "negative"}]* What is the aspect sentiment triplet of the sentence "***Great Indian food***"? Return with JSON format.
ASQP	The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, and *sentiment polarity*. The *aspect category* is only selected from the following set: ['service general', 'ambience general', 'restaurant miscellaneous', 'food quality', 'restaurant prices', 'drinks quality', 'restaurant general', 'food prices', 'drinks prices', 'drinks style_options', 'food style_options', 'location general', 'food general'] The aspect sentiment quad of the sentence "***Gross food – Wow -***" is *[{"aspect_category": "food quality", "aspect_term": "food", "opinion_term": "gross", "sentiment_polarity": "negative"}]* What is the aspect sentiment quad of the sentence "***Great Indian food***"? Return with JSON format.

Figure 8: The detailed prompt cases for five tasks.

AOPE(Zero-shot) in *Rest15*

The aspect-opinion pair consists of *aspect term* and *opinion term*. What is the aspect-opinion pair of the sentence "***oh speaking of bathroom , the mens bathroom was disgusting .***"? Return with JSON format.

{"aspect": *"bathroom"*, "opinion": *"disgusting"*}

Preds: (*bathroom*, *disgusting*)

Ground Truth: (*mens bathroom*, *disgusting*) ✗

AOPE(Zero-shot) in *Laptop14*

The aspect-opinion pair consists of *aspect term* and *opinion term*. What is the aspect-opinion pair of the sentence "***I also wanted Windows 7 , which this one has .***"? Return with JSON format.

{"aspect": *"Windows 7"*, "opinion": *"this one has"*}

Preds: (*Windows 7*, *this one has*)

Ground Truth: (*Windows 7*, *wanted*) ✗

AOPE(Few-shot) in *Rest15*

The aspect-opinion pair consists of *aspect term* and *opinion term*. The aspect-opinion pair of the sentence "*My steak au poivre was one of the worst I 've had .*" is *[{'aspect_term': 'steak au poivre', 'opinion_term': 'worst'}]*
What is the aspect-opinion pair of the sentence "***oh speaking of bathroom , the mens bathroom was disgusting .***"? Return with JSON format.

{"aspect_term": *"mens bathroom"*, "opinion_term": *"disgusting"*}

Preds: (*mens bathroom*, *disgusting*)

Ground Truth: (*mens bathroom*, *disgusting*) ✓

AOPE(Few-shot) in *Laptop14*

The aspect-opinion pair consists of *aspect term* and *opinion term*. The aspect-opinion pair of the sentence "*The programs are great*" is *[{'aspect_term': 'programs', 'opinion_term': 'great'}]*
What is the aspect-opinion pair of the sentence "***I also wanted Windows 7 , which this one has .***"? Return with JSON format.

{"aspect_term": *"windows 7"*, "opinion_term": *"wanted"*}

Preds: (*Windows 7*, *wanted*)

Ground Truth: (*Windows 7*, *wanted*) ✓

Figure 9: Case study for AOPE.

ACSA(Zero-shot) in *Rest15*

The category-sentiment pair consists of *aspect category and sentiment polarity*.
The *aspect category* is only selected from the following set: [restaurant general, service general, food style options, food quality, drinks style options, drinks prices, restaurant prices, ambience general, restaurant miscellaneous, food prices, location general, drinks quality, food general].
What is the category-sentiment pair of the sentence "***So good***"? Return with JSON format.

{"aspect_category": *"food general"*, "sentiment_polarity": *"positive"*}

Preds: (*food general*, *positive*)

Ground Truth: (*restaurant general*, *positive*)

ACSA(Zero-shot) in *Rest16*

The category-sentiment pair consists of *aspect category and sentiment polarity*.
The *aspect category* is only selected from the following set: [restaurant general, service general, food style options, food quality, drinks prices, drinks style options, restaurant prices, restaurant miscellaneous, ambience general, food prices, location general, drinks quality].
What is the category-sentiment pair of the sentence "***i will never return again .***"? Return with JSON format.

{"aspect_category": *"restaurant miscellaneous"*, "sentiment_polarity": *"negative"*}

Preds: (*restaurant miscellaneous*, *negative*)

Ground Truth: (*restaurant general*, *negative*)

ACSA(Few-shot) in *Rest15*

The category-sentiment pair consists of *aspect category and sentiment polarity*.
The *aspect category* is only selected from the following set: [restaurant general, service general, food style options, food quality, drinks style options, drinks prices, restaurant prices, ambience general, restaurant miscellaneous, food prices, location general, drinks quality, food general].
The category-sentiment pair of the sentence "*we love the food .*" is *[{'aspect_category': 'food quality', 'sentiment_polarity': 'positive'}]*
What is the category-sentiment pair of the sentence "***So good***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "sentiment_polarity": *"positive"*}

Preds: (*restaurant general*, *positive*)

Ground Truth: (*restaurant general*, *positive*)

ACSA(Few-shot) in *Rest16*

The category-sentiment pair consists of *aspect category and sentiment polarity*.
The *aspect category* is only selected from the following set: [restaurant general, service general, food style options, food quality, drinks prices, drinks style options, restaurant prices, restaurant miscellaneous, ambience general, food prices, location general, drinks quality].
The category-sentiment pair of the sentence "*the staff is rude and not attentive .*" is *[{'aspect_category': 'service general', 'sentiment_polarity': 'negative'}]*
What is the category-sentiment pair of the sentence "***i will never return again .***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "sentiment_polarity": *"negative"*}

Preds: (*restaurant general*, *negative*)

Ground Truth: (*restaurant general*, *negative*)

Figure 10: Case study for ACSA.

Figure 11: Case study for ASTE.

TASD(Zero-shot) in *Rest15* (EA)

The aspect sentiment triplet consists of *aspect category*, *aspect term*, and *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
What is the aspect sentiment triplet of the sentence "***Have never had a problem with service save a missing rice once .***"? Return with JSON format.

{"aspect_category": *"service general"*, "aspect_term": *"problem"*, "sentiment_polarity": *"negative"*}

Preds: (*problem*, *service general*, *negative*)

Ground Truth: (*service*, *service general*, *negative*)

TASD(Zero-shot) in *Rest15* (IA)

The aspect sentiment triplet consists of *aspect category*, *aspect term*, and *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
What is the aspect sentiment triplet of the sentence "***Love it .***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "aspect_term": *"it"*, "sentiment_polarity": *"positive"*}

Preds: (*it*, *restaurant general*, *positive*)

Ground Truth: (*NULL*, *restaurant general*, *positive*)

TASD(Few-shot) in *Rest15* (EA)

The aspect sentiment triplet consists of *aspect category*, *aspect term*, and *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
The aspect sentiment triplet of the sentence "*the waitstaffs are nice though .*" is *[{'aspect_category': 'service general', 'aspect_term': 'waitstaffs', 'sentiment_polarity': 'positive'}]*
What is the aspect sentiment triplet of the sentence "***Have never had a problem with service save a missing rice once .***"? Return with JSON format.

{"aspect_category": *"service general"*, "aspect_term": *"service"*, "sentiment_polarity": *"negative"*}

Preds: (*service*, *service general*, *negative*)

Ground Truth: (*service*, *service general*, *negative*)

TASD(Few-shot) in *Rest15* (IA)

The aspect sentiment triplet consists of *aspect category*, *aspect term*, and *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
The aspect sentiment triplet of the sentence "*the waitstaffs are nice though .*" is *[{'aspect_category': 'service general', 'aspect_term': 'waitstaffs', 'sentiment_polarity': 'positive'}]*
What is the aspect sentiment triplet of the sentence "***Love it .***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "aspect_term": *"restaurant"*, "sentiment_polarity": *"positive"*}

Preds: (*NULL*, *restaurant general*, *positive*)

Ground Truth: (*NULL*, *restaurant general*, *positive*)

Figure 12: Case study for TASD.

ASQP(Zero-shot) in *Restaurant* (EA&EO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
What is the aspect sentiment quad of the sentence "***fair menu selection .***"? Return with JSON format.

{"aspect_category": *"food style_options"*, "aspect_term": *"menu selection"*, "opinion_term": *"modern"*, "sentiment_polarity": *"positive"*}

Preds: (*menu selection*, *food style_options*, *positive*, *modern*)

Ground Truth: (*menu selection*, *food style_options*, *neutral*, *modern*)

ASQP(Zero-shot) in *Restaurant* (IA&EO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
What is the aspect sentiment quad of the sentence "***yum !***"? Return with JSON format.

{"aspect_category": *"food quality"*, "aspect_term": *"yum"*, "opinion_term": *"NULL"*, "sentiment_polarity": *"positive"*}

Preds: (*yum*, *food quality*, *positive*, *NULL*)

Ground Truth: (*NULL*, *food quality*, *positive*, *yum*)

ASQP(Few-shot) in *Restaurant* (EA&EO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
The aspect sentiment quad of the sentence "*Not only is the cuisine the best around , the service has always been attentive and charming .*" is *[{'aspect_category': 'food quality', 'aspect_term': 'cuisine', 'opinion_term': 'best', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'attentive', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'charming', 'sentiment_polarity': 'positive'}]*
What is the aspect sentiment quad of the sentence "***fair menu selection .***"? Return with JSON format.

{"aspect_category": *"food style_options"*, "aspect_term": *"menu selection"*, "opinion_term": *"modern"*, "sentiment_polarity": *"neutral"*}

Preds: (*menu selection*, *food style_options*, *neutral*, *modern*)

Ground Truth: (*menu selection*, *food style_options*, *neutral*, *modern*)

ASQP(Few-shot) in *Restaurant* (IA&EO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
The aspect sentiment quad of the sentence "*Not only is the cuisine the best around , the service has always been attentive and charming .*" is *[{'aspect_category': 'food quality', 'aspect_term': 'cuisine', 'opinion_term': 'best', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'attentive', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'charming', 'sentiment_polarity': 'positive'}]*
What is the aspect sentiment quad of the sentence "***yum !***"? Return with JSON format.

{"aspect_category": *"food quality"*, "aspect_term": *"NULL"*, "opinion_term": *"yum"*, "sentiment_polarity": *"positive"*}

Preds: (*NULL*, *food quality*, *positive*, *yum*)

Ground Truth: (*NULL*, *food quality*, *positive*, *yum*)

Figure 13: Case study for ASQP.

ASQP(Zero-shot) in *Restaurant* (EA&IO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and* *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
What is the aspect sentiment quad of the sentence "***not the place it once was***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "aspect_term": *"place"*, "opinion_term": *"not"*, "sentiment_polarity": *"negative"*}

Preds: (*place*, *restaurant general*, *negative*, *not*)

Ground Truth: (*place*, *restaurant general*, *negative*, *NULL*)

ASQP(Zero-shot) in *Restaurant* (IA&IO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and* *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
What is the aspect sentiment quad of the sentence "***i can not wait to go back again this coming weekend !***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "aspect_term": *"go back"*, "opinion_term": *"not wait to go back"*, "sentiment_polarity": *"positive"*}

Preds: (*go back*, *restaurant general*, *positive*, *not wait to go back*)

Ground Truth: (*NULL*, *restaurant general*, *positive*, *NULL*)

ASQP(Few-shot) in *Restaurant* (EA&IO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and* *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
The aspect sentiment quad of the sentence "*Not only is the cuisine the best around , the service has always been attentive and charming .*" is *[{'aspect_category': 'food quality', 'aspect_term': 'cuisine', 'opinion_term': 'best', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'attentive', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'charming', 'sentiment_polarity': 'positive'}]*
What is the aspect sentiment quad of the sentence "***not the place it once was***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "aspect_term": *"place"*, "opinion_term": *"NULL"*, "sentiment_polarity": *"negative"*}

Preds: (*place*, *restaurant general*, *negative*, *NULL*)

Ground Truth: (*place*, *restaurant general*, *negative*, *NULL*)

ASQP(Few-shot) in *Restaurant* (IA&IO)

The aspect sentiment quad consists of *aspect category*, *aspect term*, *opinion term*, *and* *sentiment polarity*.
The *aspect category* is only selected from the following set: [service general, ambience general, restaurant miscellaneous, food quality, restaurant prices, drinks quality, restaurant general, food prices, drinks prices, drinks style_options, food style_options, location general, food general].
The aspect sentiment quad of the sentence "*Not only is the cuisine the best around , the service has always been attentive and charming .*" is *[{'aspect_category': 'food quality', 'aspect_term': 'cuisine', 'opinion_term': 'best', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'attentive', 'sentiment_polarity': 'positive'}, {'aspect_category': 'service general', 'aspect_term': 'service', 'opinion_term': 'charming', 'sentiment_polarity': 'positive'}]*
What is the aspect sentiment quad of the sentence "***i can not wait to go back again this coming weekend !***"? Return with JSON format.

{"aspect_category": *"restaurant general"*, "aspect_term": *"NULL"*, "opinion_term": *"NULL"*, "sentiment_polarity": *"positive"*}

Preds: (*NULL*, *restaurant general*, *positive*, *NULL*)

Ground Truth: (*NULL*, *restaurant general*, *positive*, *NULL*)

Figure 14: Case study for ASQP.