Zero-Shot Cross-Domain Aspect-Based Sentiment Analysis: A Hybrid Augmentation Framework with Domain-Contextualized Chain-of-Thought Reasoning

Anonymous ACL submission

Abstract

Cross-domain aspect-based sentiment analysis (ABSA) aims at learning specific knowledge from a source domain to perform various ABSA tasks on a target domain. Recent works mainly focus on how to use domain adaptation techniques to transfer the domain-agnostic features from the labeled source domain to the unlabeled target domain. However, it would be unwise to manually collect a large number of unlabeled data from the target domain, where such data may not be available owing to the facts like data security concerns in banking or insurance. To alleviate this issue, we propose ZeroABSA, a unified zero-shot learning framework for cross-domain ABSA that effectively eliminates dependency on target-domain annotations. Specifically, ZeroABSA consists of two novel components, namely, (1) A *hybrid data augmentation module* leverages large language models (LLMs) to synthesize high-quality, domain-adaptive target-domain data, by evaluating the generated samples across vocabulary richness, semantic coherence and sentiment/domain consistency, followed by iterative refinement; (2) A *domain-contextualized chain-of-thought (COT)* strategy trains models on augmented data while explicitly modeling domain-invariant reasoning to bridge the well-known cross-domain gap. Extensive evaluations across four diverse domains demonstrate that ZeroABSA surpasses the state-of-the-arts, which effectively advances the practicality of cross-domain ABSA in real-world scenarios where labeled target-domain data is unavailable.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) is a widely-discussed fine-grained sentiment analysis task (Pontiki et al., 2016), aims at identifying sentiment targets within sentences to form the structured pairs like <aspect, polarity>, where the polarity "positive" is a specific sentiment towards a target aspect "food" in sentence "The food at this restaurant is good." This end-to-end formulation has evolved into three principal subtasks: (1) Aspect Term Extraction (ATE), isolating domain-spcific aspect terms from sentences (Liu et al., 2015); (2) Aspect Sentiment Classification (ASC), predicting the sentiment polarities for given terms (Zhang et al., 2016; Wang et al., 2020); and (3) Aspect Sentiment Triplet Extraction (ASTE), extending initial ABSA to a triplet (e.g., "<food, good, positive>"), capturing richer contextual sementics (Peng et al., 2020; Chen et al., 2021; Liang et al., 2023). However, these paradigms still restricted to domain-specific data scarcity in low-resource domains.

Therefore, many researchers tackle various ABSA tasks beyond a specific domain. They primarily focus on cross-domain sentiment correlations by aligning latent feature distributions across domains, which is known as cross-domain ABSA (Wang and Pan, 2018; Li et al., 2019; Zhou et al., 2021), leveraging the availability of a tremendous amount of sentiments expressed across different domains. The principle of such methods is to employ unsupervised domain adaptation (UDA) techniques to learn domain-invariant features for various cross-domain ABSA tasks, which, however, always heavily relies on numerous collected unlabeled data from the target domain to minimize the domain gap for training (Blitzer et al., 2007; Zhuang et al., 2015; Dai et al., 2020; Chen et al., 2022). Nevertheless, it may face a significant challenge, that is, the inadequacy of unlabeled data in target domain, as such data are usually scarce in practice due to facts like data security concerns in the banking or insurance domain.

Recent advancements have explored the use of pre-trained language models for data augmentation in cross-domain ABSA tasks (Yu et al., 2021; Yang et al., 2022; Yu et al., 2023). For instance, Yu et al. combines domain-adaptive pseudo-labeling with language modeling to improve the effectiveness of cross-domain data augmentation. However, these approaches still depend on unlabeled target domain data to generate pseudo-labeled data. Furthermore, the common approach of training first on labeled source domain data and then on generated target domain data (Deng et al., 2023) can lead to inconsistencies. The generated target domain data often differ significantly from the source domain data, causing difficulties in maintaining domain-specific awareness during inference. This can result in models struggling to bridge the gap between source and target domains effectively, ultimately impacting performance.

To overcome these limitations, we propose a novel zero-shot cross-domain ABSA framework that enables domain-invariant feature learning and knowledge transfer without requiring target domain annotations. Our method consists of three steps: 1) **Zero-Shot Data Augmentation:** We utilize large language models (LLMs) to generate target domain data by leveraging weak supervision signals (such as names, descriptions) from the target domain along with the existing labeled source domain data. By harnessing the in-context learning abilities of LLMs, we create a diverse and rich set of simulated data for the target domain. Additionally, a portion of this data is generated without any reference to further enrich the diversity, enabling the model to learn domain-invariant features that are adaptable to the target domain characteristics. 2) **Evaluation of Generated Data:** To ensure the quality and fluency of the generated data, we first calculate its vocabulary richness using Shannon entropy. Additionally, we evaluate the data by calculating Domain Consistency, Sentiment Consistency, and Sentence Fluency using a ranking model. Based on these metrics, we select the highest-quality data and combine it with existing domain data for model training. 3) **Domain-Contextualized Chain-of-Thought:** To further bridge the gap between the source and target domains and enhance the model's sensitivity to target domain characteristics, we introduce the Domain-Contextualized Chain-of-Thought Reasoning. During model inference, this prompt guides the model through a structured reasoning process: it first considers the domain of the data, then generates intermediate reasoning steps, and finally produces the final output. This structured process helps the model incorporate domain-specific knowledge and context, ensuring it is more attuned to the target domain and capable of performing various ABSA tasks effectively.

The main contributions of our work can be summarized as follows:

- To our knowledge, we are the first to tackle cross-domain ABSA in a zero-shot setting, where no target domain data is available. This approach is particularly significant for scenarios with strict data privacy and security requirements, where collecting target domain data is not feasible.
- We introduce an innovative framework that integrates hybrid data augmentation with Domain-Contextualized Chain-of-Thought Reasoning. This framework enhances domain-invariant feature learning and bridges the gap between source and target domains by using LLMs to generate high-quality target data and ensuring domain-specific sensitivity during inference.
- Extensive experimental results validate the effectiveness of our method, showing that it outperforms existing approaches in zero-shot settings for cross-domain ABSA tasks, thereby demonstrating the robustness of our approach.

2 Related Work

2.1 Cross-Domain ABSA

Cross-domain ABSA has become a highly discussed topic in recent years. Early studies employed common techniques from Unsupervised Domain Adaptation (UDA), using specific syntactic rules of the target domain to minimize the loss caused by domain transfer (Jakob and Gurevych, 2010; Ding et al., 2017; Wang and Pan, 2019). Additionally, many studies have used domain discriminators to learn generalizable knowledge across different domains (Li et al., 2019; Zhang et al., 2023). Recently, with the rising popularity of the pre-training model paradigm, some works have utilized pre-trained models to generate additional data (Wei and Zou, 2019; Yu et al., 2021; Li et al., 2022; Yu et al., 2023). Although these methods are effective, they almost all require corpus data or other external resources from the target domain, which can pose certain challenges in real-world applications.

2.2 Data Augmentation

Data augmentation is a technique used to increase the amount of training data by applying various transformations to existing data or generating new data, thereby enhancing the model's generalization ability and performance (Feng et al., 2021; Mumuni and Mumuni, 2022). In the field of NLP

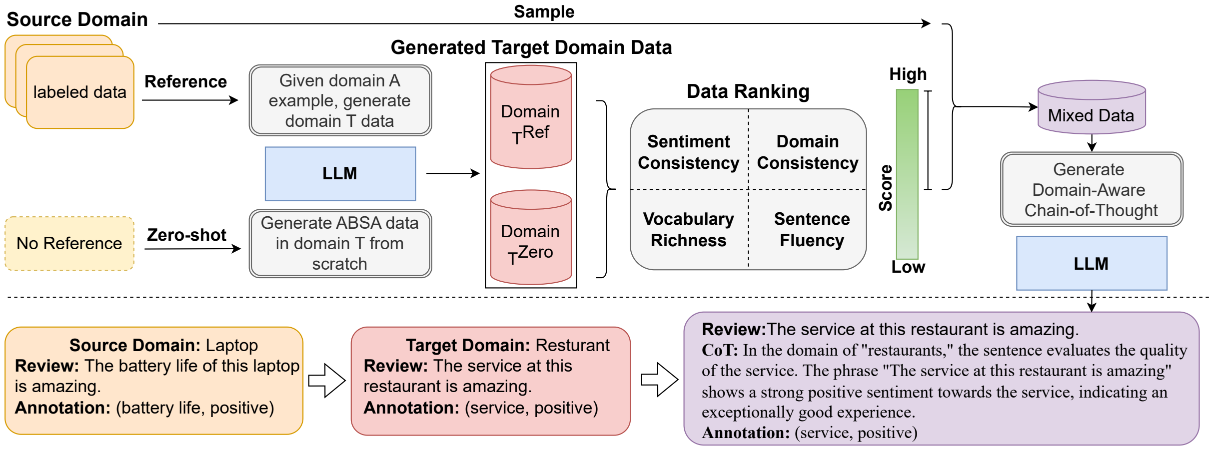

Figure 1: Overall Framework of Data Augmentation in Our Work. Orange represents source domain data, red represents generated target domain data, purple represents generated data filtered by ranking, and red represents the final mixed data with Chain-of-Thought Explanation.

(Natural Language Processing), early data augmentation techniques typically involved synonym replacement, random insertion, random swap, and random deletion. Recently, with the rising popularity of the pre-training model paradigm, some works have utilized pre-trained models to generate additional data for data augmentation (Kumar et al., 2020; Yu et al., 2023). Although these methods have shown remarkable results, they all face the issue of relying on labeled data or pure corpus information from specific domains. Moreover, Existing cross-domain ABSA data augmentation methods typically rely on MLM for word replacement, which often results in generated data that lacks diversity and fluency. Moreover, it's noteworthy that while zero-shot data augmentation has seen some exploration in the field of computer vision (CV) (Fahes et al., 2023), its application in NLP remains relatively underexplored.

2.3 Large Language Model

Since OpenAI released ChatGPT, an increasing number of studies have examined the performance of LLMs on various downstream NLP tasks (OpenAI et al., 2024; Zhao et al., 2023; Wei et al., 2022a). Due to their pre-training on extensive corpora, LLMs have demonstrated excellent generalization and strong transfer learning capabilities across diverse tasks. These models not only generate high-quality natural language text but also perform well on new tasks and domains without specialized training. For example, in sentiment analysis, question answering systems, and text summarization, LLMs have achieved significant results. Moreover, their ability to adapt to structured prediction tasks, such as named entity recognition and syntactic parsing, further highlights their versatility. One key factor contributing to these successes is the emergent capabilities of LLMs, such as in-context learning and Chain-of-Thought reasoning (Wei et al., 2022b). These capabilities enable the models to solve complex reasoning tasks through contextual inference and step-by-step thinking (Wei et al., 2022c). This makes it possible to utilize LLMs for various NLP tasks. Furthermore, the ability of LLMs to generalize across domains has opened up exciting opportunities for applying them to previously unexplored tasks.

With the popularity of LLMs, an increasing number of studies have utilized the strong generalization capabilities of these models for data augmentation to achieve domain adaptation (Sahu et al., 2022). Compared to previous generative models, LLMs trained on more extensive corpora can generate more fluent and diverse data. Although LLMs may lack domain-specific knowledge of the target domain, they excel at capturing broad patterns across different domains (Wei et al., 2022a). Given labeled source domain examples, an LLM can approximate the characteristics of the target domain solely through natural language descriptions of the target domain. Previous studies have demonstrated that LLMs can still generate reasonably good data for data augmentation (Whitehouse et al., 2023), even in unfamiliar domains. However, despite these advancements, few works focus on using LLMs to achieve domain transfer for ABSA tasks, especially in a zero-shot setting.

3 Methodology

3.1 Problem Definition and Notations

Based on the previous work on defining the ABSA task, given a sentence $X = \{w_1, w_2, \ldots, w_n\}$ with n words, the goal of the ABSA task is to extract several tuples $Y = \{(a_i, p_i)\}_{i=1}^{|Y|}$, where a represents aspect terms, which are subsets of words in the sentence S. For each aspect a, the corresponding sentiment polarity p belongs to $P = \{\text{Positive}, \text{Negative}, \text{Neutral}\}$.

Our work focuses on achieving domain adaptation for the ABSA task in a zero-shot setting. In this setting, there are labeled source domain datasets, but no data from the target domain is available before testing. Let $D^S = \{(X_i^S, Y_i^S)\}_{i=1}^{|D^S|}$ represent the labeled data from the source domains. The task is to extract tuples Y from the target domain D^T given labeled data D^S from any source domain.

3.2 Overall Framework

Our method comprises three stages: **Zero-shot Data Augmentation**, **Evaluation of Generated Data**, and **Domain-Contextualized Chain-of-Thought**. In the first stage, we utilize the names and the description of the target domain to generate target data. Leveraging pre-trained large models, we generate a series of simulated data for the target domain. In the second stage, we employ a rank model to score the generated data based on its fluency and relevance. Combined with the vocabulary richness of the data, we conduct a comprehensive ranking, selecting high-scoring data to mix with the existing data. In the third stage, we propose a Domain-Contextualized Chain-of-Thought approach. This involves providing explanatory steps for data generation and using this comprehensive data for model training. By reflecting on specific domains during inference and outputting step-by-step reasoning, the model can become more attuned to the target domain, despite being trained on data from various domains and sources. We present the overall framework of data augmentation in our work in Figure 1.

3.3 Zero-shot Data Augmentation

In this stage, our primary objective is to generate a rich dataset for the target domain D^T in a zero-shot setting. Inspired by previous work in the field of image classification in computer vision (Fahes et al., 2023), we use only a general description in natural language of the target domain to generate target domain data. To ensure the generated data closely resembles real reviews, we leverage the in-context learning capabilities of LLMs. For every source domain data, we manually construct k examples and ultlize the LLM's extensive corpus to replace them with structurally similar simulated target domain data. Previous work has implemented similar approaches (Yu et al., 2023), primarily relying on BERT-based models for replacements and necessitating additional target domain vocabulary. By using LLMs, we effectively reduce dependency on specific vocabulary. Moreover, due to the autoregressive nature of LLMs, they can dynamically adjust vocabulary and sentence structure during generation, resulting in more natural and diverse target domain data. Through in-context learning with source domain data, the model can generate text that aligns with the style and context of the target domain. To further enhance the diversity of the generated data, we also prompt LLMs to perform reference-free data generation.

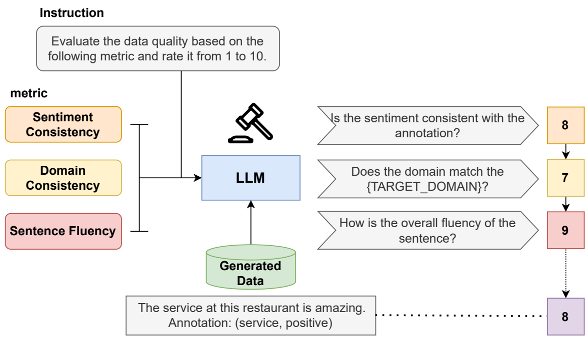

Figure 2: The main process of the Rank Model in our work. We use a large model as the Rank Model, scoring each piece of generated data from the source domain based on three metrics.

3.4 Evaluation of Generated Data

For existing LLMs, although they excel at data generation, the generated data can sometimes exhibit hallucinations (i.e., content that is inaccurate or not factually correct). Even target domain data generated from source domain data can vary in quality, lacking fluency in expression, which are crucial for the model's understanding and generation of natural language. To ensure that the generated data effectively supports model training with high quality, we introduce data ranking and filtering steps.

In our observations, LLMs tend to replace keywords from the source domain with a single vocabulary. To ensure the vocabulary richness of the generated data, our work calculate the Shannon

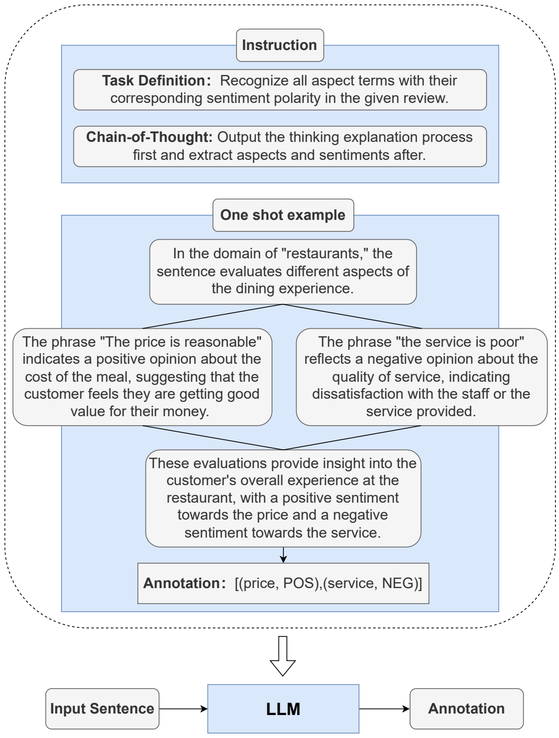

Figure 3: The main process of the Domain-Contextualized Chain-of-Thought.

entropy of the sentences as one of the ranking metrics:

$$H(X) = -\sum_{w \in X} p(w) \log_2 p(w) \tag{1}$$

To measure the fluency and task relevance of the generated data, we adopted the **LLM-as-Judge** framework commonly used in LLM benchmarks (Zheng et al., 2023), utilizing an LLM as the rank model. After obtaining generated data from the previous stage, we first need to remove examples that do not meet the required format. Then, we use the rank model to score the data quality.

We selected Sentiment Consistency, Domain Consistency, and Sentence Fluency as the scoring metrics. For each sentence X, the model outputs scores from 1 to 10 for each metric. We use their average S_{Avg} as final score of the rank model, denoted as S^{Avg}. The main process of the rank model is illustrated in Figure 2.

Finally, we combine the Shannon entropy and the rank model's score to compute the final score S. This score ensures that the generated data is both diverse in vocabulary and high in quality. The final score is calculated as follows:

$$S = \alpha \cdot \frac{H - H_{\min}}{H_{\max} - H_{\min}} + \beta \cdot \frac{S - S_{\min}}{S_{\max} - S_{\min}}, \tag{2}$$

where H_{min} and H_{max} are the minimum and maximum Shannon entropy values in the dataset respectively. S_{min} and S_{max} are the minimum and maximum scores from the rank model in the dataset respectively. α and β are the weights for the two metrics.

Based on the aforementioned data ranking, we select the top $\gamma\%$ of the target domain data generated from each source domain as the training data, ultimately mixing the data generated from n source domains. To ensure diversity and authenticity of the data, we also mix the generated data with the source domain data for model training.

3.5 Domain-Contextualized Chain-of-Thought

To address the issue of performance instability caused by training on multi-source domain generated data, we propose the Domain-Contextualized Chain-of-Thought Reasoning. This method guides the model to perform step-by-step reasoning during inference, ensuring it can recognize and understand the characteristics and context of the target domain, thereby enhancing its performance in the target domain.

Specifically, at the start of the inference, the model first identifies the domain to which the current data belongs. This step enables the model to adjust its subsequent reasoning process and generation strategy accordingly. Then, based on the domain information, the model generates intermediate steps through a pre-designed chain of thought. These steps involve reflecting on and understanding domain-specific features, ensuring that the model fully considers the context and characteristics of the target domain during generation. Finally, after going through the chain of thought process, the model produces the final output. This process not only ensures the accuracy and fluency of the generated content but also enhances the model's sensitivity and adaptability to the target domain.

To ensure that the model strictly follows the Domain-Contextualized Chain-of-Thought process, we first utilize LLMs to generate the thinking process for the training data. This allows the model to internalize domain-specific reasoning patterns and learn the prior probability distribution of the generation process, reinforcing its ability to follow structured logical steps. Additionally, we include a one-shot example in the prompt to further enhance the model's performance in the ABSA task, demonstrating the effectiveness of our approach in

Table 1: Statistics of the datasets.

Dataset	Total	Positive	Negative	Neutral
Device$_{\text{train}}$	1411	908	503	0
Device$_{\text{test}}$	697	481	216	0
Laptop$_{\text{train}}$	2303	988	861	454
Laptop$_{\text{test}}$	634	339	130	165
Rest$_{\text{train}}$	4314	2610	1037	667
Rest$_{\text{test}}$	2289	1524	501	264
Service$_{\text{train}}$	1844	1034	698	112
Service$_{\text{test}}$	887	506	320	61

sentiment-oriented domain adaptation. The inference process is illustrated in Figure 3.

4 Experiments

4.1 Datasets and Metrics

In our experiments, to validate the effectiveness of our method for cross-domain ABSA tasks, we follow previous work and evaluate on four datasets: Laptop (L), Restaurant (R), Device (D), and Service (S). The statistics for these four datasets are shown in Table 1.

Among these datasets, Laptop and Restaurant are from SemEval (Pontiki et al., 2014, 2015, 2016). They are two of the most common English datasets in ABSA tasks. Device comes from the work of Hu and Liu, and includes reviews of digital cameras, cellular phones, MP3 players, and DVD players. Service is from the work of Toprak et al. and mainly contains reviews of online services such as PayPal, eGroups, and eTrade. We applied the most commonly used metrics in ABSA tasks, Accuracy and Macro-F1. For the extraction of $(aspect, polarity)$ tuples, a tuple is considered correct only if both components are entirely accurate.

4.2 Experimental Settings

In our experiments, we used gpt-4o-mini as the model for generating target domain data and as the rank model. In the stage of data generation, about 20% of our data is generated in reference-free settings. The remaining data is generated with reference to the labeled source domain data in a few-shot setting with $k = 3$. For model training, we adopted LLaMA-3-8b-instruct as our base model (Grattafiori et al., 2024). We fine-tuned the model for downstream tasks using LoRA, setting the LoRA rank and LoRA alpha to 32. We optimized the parameters using the Adam algorithm with a learning rate of 1e-4. The model was trained for 10 epochs on 8 NVIDIA RTX 4090 GPUs with 24GB of memory each. For the hyperparameter settings in the data evaluation phase, based on extensive experimentation, we set $\alpha = 0.5$, $\beta = 0.5$ and $\gamma = 0.25$. After the model outputs its results, given that the model is case-insensitive, we restored the original casing of each word in the output to ensure complete matching. All data presented in this study are averaged over five runs.

4.3 Baselines

To demonstrate the effectiveness of our method in zero-shot settings, we compared our method with the following competitive cross-domain adaptation methods. Since most previous works are unable to perform in a zero-shot setting, for a fair comparison, we compared our method with some baselines under non-zero-shot settings, and our approach still shows competitive results.

The baselines that require target domain data for comparison are as follows:

- **Hier-Joint** (Ding et al., 2017) A method using LSTM for domain adaptation that incorporates syntactic rule-based auxiliary tasks to enhance cross-domain ABSA.
- **RNSCN** (Wang and Pan, 2018) A model that is RNN-based and employs structural correspondence through syntactic structures and autoencoders.
- **AD-SAL** (Li et al., 2019) A method that leverages Selective Adversarial Learning to achieve fine-grained domain adaptation by aligning local semantics.
- **BERT-UDA** (Gong et al., 2020) An unified feature and instance-based domain adaption method.
- **BGCA** (Yu et al., 2023) A model that leverages a bidirectional generative framework for data augmentation in cross-domain ABSA. We select the *label-to-text* version of the model proposed in the work.

The baselines we compared under the zero-shot settings are as follows:

- **BERT-base** Directly fine-tuned version of bert-uncased from Devlin et al. on labeled source domain data.
- **LLaMA-base** A LLaMA-3-8b-instruct version only fine-tuned on the labeled source domain, employing the same prompts and training format as our method, except for the chain-of-thought component.

Methods	S→R	L→R	D→R	R→S	L→S	D→S	R→L	S→L	R→D	S→D
Target Domain Needed										
Hier-Joint†	31.10	33.54	32.87	15.56	13.90	19.04	20.72	22.65	24.53	23.24
RNSCN†	33.21	35.65	34.60	20.04	16.59	20.03	26.63	18.87	33.26	22.00
AD-SAL†	41.03	43.04	41.01	28.01	27.20	26.62	34.13	27.04	35.44	33.56
BERT-UDA†	47.09	45.46	42.68	33.12	27.89	28.03	33.68	34.77	34.93	32.10
BERT-CDRG†	47.92	49.79	47.64	35.14	38.14	37.22	38.68	33.69	27.46	34.08
BGCA†	56.39	**61.69**	**59.12**	43.20	39.76	47.94	**45.52**	36.40	34.16	36.57
Zero-shot										
BERT-base†	44.66	40.38	40.32	19.48	25.78	30.31	31.44	30.47	27.55	33.96
LLaMA-base	59.99	48.56	56.34	32.04	27.54	38.28	**45.52**	**39.73**	**42.12**	38.22
GPT-4o	55.91	49.85	54.37	29.33	26.91	30.09	31.87	34.02	37.32	35.26
ZeroABSA	**60.45**	48.97	57.49	**46.27**	**43.83**	**51.22**	36.80	38.09	34.08	**40.89**

Table 2: Comparison results of different methods for Cross-Domain End-to-End ABSA tasks based on Macro-F1. The best results are highlighted in **bold**, while the second-best results are underlined. The notation † denotes results from Yu et al..

Methods	S→R	L→R	D→R	R→S	L→S	D→S	R→L	S→L	R→D	S→D
Target Domain Needed										
Hier-Joint†	46.39	48.61	42.96	27.18	25.22	29.28	34.11	33.02	34.81	35.00
RNSCN†	48.89	52.19	50.39	30.41	31.21	35.50	47.23	34.03	**46.16**	32.41
AD-SAL†	52.05	56.12	51.55	39.02	38.26	36.11	45.01	35.99	43.76	41.21
BERT-UDA†	56.08	51.91	50.54	34.62	32.49	34.52	46.87	43.98	40.34	38.36
BERT-CDRG†	56.26	60.03	52.71	42.36	47.08	41.85	46.65	39.51	32.60	36.97
BGCA†	63.20	**69.53**	65.33	45.86	44.85	54.07	**57.13**	46.15	37.15	38.24
Zero-shot										
BERT-base†	54.29	46.74	44.63	22.31	30.66	33.33	37.02	36.88	32.03	38.06
LLaMA-base	65.12	51.84	59.07	35.92	30.34	39.58	53.09	44.84	45.43	40.22
GPT-4o	**69.22**	64.90	**66.69**	47.61	45.34	48.30	51.31	**54.76**	40.48	38.78
ZeroABSA	65.98	53.30	63.82	**51.99**	**50.26**	**55.43**	41.45	44.99	36.78	**42.07**

Table 3: Comparison results of different methods for Cross-Domain ATE tasks based on Macro-F1. The best results are highlighted in **bold**, while the second-best results are underlined. The notation † denotes results from Yu et al..

- **GPT-4o** Utilizing one of the most powerful LLMs currently available, GPT-4o, to achieve cross-domain ABSA. Specifically, we selected the gpt-4o-2024-08-06 version and employed three randomly chosen labeled source domain data points as few-shot examples for inference.

We are the first group to investigate zero-shot cross-domain ABSA. Compared to previous work, our approach considers scenarios where target domain data is inaccessible, achieving domain transfer in zero-shot settings. If our method surpasses previous approaches that require target domain data, it demonstrates that our method can still ensure effectiveness even in the absence of target domain corpus.

4.4 Main Results

We present the results for the End-to-End ABSA and ATE tasks in Table 2 and Table 3, respectively. Overall, our method performs exceptionally well across both the ***target domain needed*** and ***zero-shot*** baseline settings. Notably, even when compared with state-of-the-art methods that require unlabeled target domain data, our method leads in most tasks. For instance, in tasks where the target domain is **service**, our method surpasses the previous state-of-the-art by 3-4%.

Compared to previous methods that require unlabeled target domain data, our approach demonstrates robust effectiveness. Despite the absence of target domain corpus for learning domain-specific features, our data augmentation and prompt techniques enable the model to significantly improve its performance in the target domain. Compared to the baselines in the zero-shot setting, our model significantly outperformed the BERT-base model, indicating that decoder-only models are also suitable for extraction-based tasks like ABSA. When compared with one of the most powerful closed-source LLMs, GPT-4o, our zero-shot approach, using only the 8B base model, surpasses its performance in few-shot settings.

Experimental results demonstrate that GPT-4o performs well across various ATE tasks. Furthermore, even with simple adjustments to prompts and inference methods, and fine-tuning on the LLaMA model, its performance far surpasses that of traditional BERT models. This finding indicates that leveraging advanced LLMs allows our approach to achieve superior results in cross-domain ABSA tasks, even in zero-shot settings, significantly improving performance in the target domain. This clearly underscores the potential and advantages of LLMs in data augmentation and domain adaptation. Despite the significant progress achieved with fine-tuning LLaMA and GPT-4o, our approach further integrates Hybird Data Augmentation and Domain-Contextualized Chain-of-Thought Reasoning, resulting in even more outstanding performance in cross-domain ABSA tasks.

4.5 Further Analysis

Ablation Study We conducted an ablation study to assess the contribution of individual components in our zero-shot cross-domain ABSA method. Table 4 reports the performance of the full model and several variants obtained by removing specific components.

Excluding the data ranking module led to a noticeable drop in performance, which confirms that high-quality generated data is essential for effective knowledge transfer. When both the data augmentation and the Domain-Contextualized Chain-of-Thought components are removed, the model achieves the worst results across all metrics. Omitting the chain-of-thought reasoning caused a decline in F1-score, although its effect on recall was less pronounced.

Model	Recall	Precision	F1-score
w/o Data Rank	32.72	48.57	39.11
w/o DA and CoT	24.35	46.84	32.04
w/o DA	36.02	40.01	37.91
w/o CoT	45.77	41.64	43.61
Full	**48.69**	**44.08**	**46.27**

Table 4: Ablation study results of our method. "w/o" denotes version without the specific component.

Zero-shot Experiment In order to compare the performance of different zero-shot settings, we evaluate fully zero-shot models, including GPT-4o and LLaMA3. Table 5 reports the results across four domains: Restaurant, Laptop, Device, and Service. For comparison, we include models that are used in a purely zero-shot fashion, as well as those that are fine-tuned with source domain data or provided with few-shot examples.

From Table 5, we observe that while the zero-shot models (LLaMA zero-shot and GPT-4o without source domain examples) perform reasonably well across the domains, their performance improves when fine-tuning with source domain data or incorporating few-shot examples. In particular, our method consistently outperforms all baselines. This indicates that our approach, which leverages both data augmentation and domain-contextualized chain-of-thought reasoning, effectively bridges the gap between source and target domains, leading to superior performance in a zero-shot setting.

Method	Rest	Laptop	Device	Service
L-zs	52.86	31.09	35.82	36.84
L-ft	54.85	32.59	**42.63**	**40.17**
G-zs	52.99	26.08	32.75	25.94
G-ex	53.38	32.95	36.29	28.78
Ours	**55.64**	**47.11**	37.85	37.49

Table 5: Zero-shot performance comparison across domains. Method abbreviations: L-zs (LLaMA zero-shot), L-ft (LLaMA fine-tuned), G-zs (GPT4o without examples), G-ex (GPT4o with examples).

5 Conclusion

In this work, we introduce a novel zero-shot cross-domain ABSA method that effectively combines hybrid data augmentation with Domain-Contextualized Chain-of-Thought, enabling domain transfer without requiring any target domain data. We generated high-quality target domain data, which was later evaluated and selected for training. The experimental results validate the effectiveness of our method, offering new insights and approaches for advancing cross-domain ABSA research.

6 Limitations

The proposed method relies on data from only four domains in the SemEval dataset, which may not fully represent the diversity of real-world domains. This limits the generalizability of the approach to other domains with different linguistic features or specific sentiment nuances. Additionally, the method's reliance on LLMs could pose scalability and computational challenges in real-world applications. The use of synthetic data generated by LLMs could unintentionally introduce biases or even violate privacy in sensitive domains, such as finance or healthcare, if not properly managed.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

David Z. Chen, Adam Faulkner, and Sahil Badyal. 2022. Unsupervised data augmentation for aspect based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6746–6751, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12666–12674.

Yong Dai, Jian Liu, Xiancong Ren, and Zenglin Xu. 2020. Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7618–7625. AAAI Press.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. 2023. PØda: Prompt-driven zero-shot domain adaptation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18577–18587.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 968–988. Association for Computational Linguistics.

Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7035–7045, Online. Association for Computational Linguistics.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045, Cambridge, MA. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Junjie Li, Jianfei Yu, and Rui Xia. 2022. Generative cross-domain data augmentation for aspect and opinion co-extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4219–4229, Seattle, United States. Association for Computational Linguistics.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. 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 4590–4600, Hong Kong, China. Association for Computational Linguistics.

Shuo Liang, Wei Wei, Xian-Ling Mao, Yuanyuan Fu, Rui Fang, and Dangyang Chen. 2023. Stage: span tagging and greedy inference scheme for aspect sentiment triplet extraction. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal. Association for Computational Linguistics.

Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.

OpenAI et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

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

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

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Wenya Wang and Sinno Jialin Pan. 2018. Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2171–2181, Melbourne, Australia. Association for Computational Linguistics.

Wenya Wang and Sinno Jialin Pan. 2019. Transferable interactive memory network for domain adaptation in fine-grained opinion extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7192–7199.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. FactMix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5360–5371, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777, Online. Association for Computational Linguistics.

Jianfei Yu, Qiankun Zhao, and Rui Xia. 2023. Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1456–1470, Toronto, Canada. Association for Computational Linguistics.

Kai Zhang, Qi Liu, Hao Qian, Biao Xiang, Qing Cui, Jun Zhou, and Enhong Chen. 2023. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):377–389.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3087–3093. AAAI Press.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yan Zhou, Fuqing Zhu, Pu Song, Jizhong Han, Tao Guo, and Songlin Hu. 2021. An adaptive hybrid framework for cross-domain aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14630–14637.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised representation learning: transfer learning with deep autoencoders. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 4119–4125. AAAI Press.

A Appendix

A.1 More Comparison with BERT-UDA

We conducted experiments comparing our model with BERT-UDA trained on augmented data. The results across various domain transfers (source-to-target) are presented in Table 6:

The results demonstrate that our model significantly outperforms BERT-UDA-based models. This improvement can be attributed to our method's ability to leverage CoT, which enhances performance when working with augmented CoT-based data. In contrast, simply using augmented CoT data with BERT-UDA does not fully leverage the advantages of the CoT structure, resulting in suboptimal performance.

A.2 Hyperparameter Sensitivity Analysis

We conducted a sensitivity analysis to evaluate the effect of different hyperparameters on performance. The results are summarized in Table 7:

From the analysis, we found that the optimal combination of hyperparameters ($\lambda = 25$, $\alpha = 0.5$, $\beta = 0.5$) yields the best performance, with a peak score of 57.49. Further details will be provided in the revised manuscript.

A.3 Results on Multiple-Domain Transfer

We conducted preliminary experiments that explored the performance of a multi-domain transfer approach. The results of these experiments, shown in Table 8, indicate that the multi-domain approach yielded an overall F1- score slightly above the average of the individual domain scores.

Method	S→R	L→R	D→R	R→S	L→S	D→S	R→L	S→L	R→D	S→D
UDA	47.09	45.46	42.68	33.12	27.89	28.03	33.68	34.77	34.93	32.10
UDA-0shot	44.89	43.12	41.34	34.48	29.56	26.31	35.69	33.25	32.92	30.77
ZeroABSA	**60.45**	**48.97**	**57.49**	**46.27**	**43.83**	**51.22**	**36.80**	**38.09**	**34.08**	**40.89**

Table 6: Comparison of our model with BERT-UDA on domain transfer tasks. Our model significantly outperforms BERT-UDA.

γ	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
5	45.2	48.6	50.1	49.5	46.7
15	48.0	51.5	53.2	52.3	50.0
25	50.9	54.3	**57.5**	55.4	53.1
35	49.6	52.0	55.0	54.0	51.2
45	47.8	50.5	52.9	51.5	48.9

Table 7: Hyperparameter sensitivity analysis. The combination of $\lambda = 25$, $\alpha = 0.5$, and $\beta = 0.5$ yielded the best performance with a peak score of 57.49.

Mix Domain	**Recall**	**Precision**	**F1-Score**
LDR→S	43.27	61.72	50.87
SDR→L	41.45	41.51	41.48
SLR→D	48.08	30.51	37.33
SLD→R	45.42	54.24	49.44

Table 8: Preliminary results on multiple-domain transfer. For example, SLR→D denotes transfer from Service, Laptop, and Restaurant domains to the Device domain.