Zero-Shot Cross-Domain Aspect-Based Sentiment Analysis via Domain-Contextualized Chain-of-Thought Reasoning

Chuming Shen[1], Wei Wei*[1], Dong Wang[2], Zhonghao Wang[3]
[1]Huazhong University of Science and Technology
[2]State Grid Fujian Electric Power Co.
[3]Hubei Data Group Co., Ltd.
scm@hust.edu.cn, weiw@hust.edu.cn, 627843795@qq.com, zhonghaowang@hbscjt.com.cn

Abstract

Cross-domain aspect-based sentiment analysis (ABSA) aims to learn domain-specific knowledge from a source domain to perform various ABSA tasks in a target domain. Recent works mainly focus on leveraging domain adaptation techniques to transfer the domain-agnostic features from the labeled source domain to the unlabeled target domain, yet manually collecting target domain data remains impractical owing to the facts like privacy concerns in banking or insurance. To alleviate this issue, we propose ZeroABSA, a unified zero-shot framework for cross-domain ABSA that effectively eliminates dependency on target domain annotations. Specifically, ZeroABSA consists of two novel components, namely, (1) a *LLM-driven augmentation module* synthesizing domain-adaptive target data through iteratively evaluating the metrics (e..g, vocabulary richness, semantic coherence, and sentiment/-domain consistency) of augmented exemplars for 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 advances the practicality of cross-domain ABSA in real-world scenarios where 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 collected unlabeled data from the target domain to minimize the domain gap (Blitzer et al., 2007; Zhuang et al., 2015; Dai et al., 2020; Chen et al., 2022). However, this approach faces a significant challenge: unlabeled target domain data is often scarce in practice, especially in domains with strict data security, like banking or insurance.

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

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4558–4573
November 4-9, 2025 ©2025 Association for Computational Linguistics

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.

We define **zero-shot cross-domain ABSA** as a challenging setting where the model is trained without any data from the target domain. In this work, our zero-shot setting aims to overcome the critical limitation of data scarcity and privacy concerns in low-resource domains by completely eliminating the need for target domain data. To overcome these limitations, we propose a novel zero-shot cross-domain ABSA framework that achieves cross-domain invariant feature learning and annotation-free knowledge transfer, comprising: 1) **Zero-Shot Data Augmentation:** We utilize large language models (LLMs) to synthesize target domain data via target domain weak supervision (e.g., domain names) with the existing labeled source domain data. By harnessing in-context learning, we generate diverse, semantically coherent simulated target domain examples. A subset is further augmented with reference-free generation to maximize diversity, facilitating learning of domain-invariant representations that generalize to 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 enhance cross-domain adaptation, we propose Domain-Contextualized Chain-of-Thought (DC-CoT), a structured reasoning framework that guides models through a multi-stage reasoning process: it first considers the domain of the data, then generates intermediate reasoning steps, and finally produces the final output. By grounding intermediate reasoning in domain-specific context, DCCoT systematically aligns latent representations with target domain characteristics, enabling robust performance across diverse ABSA tasks.

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.

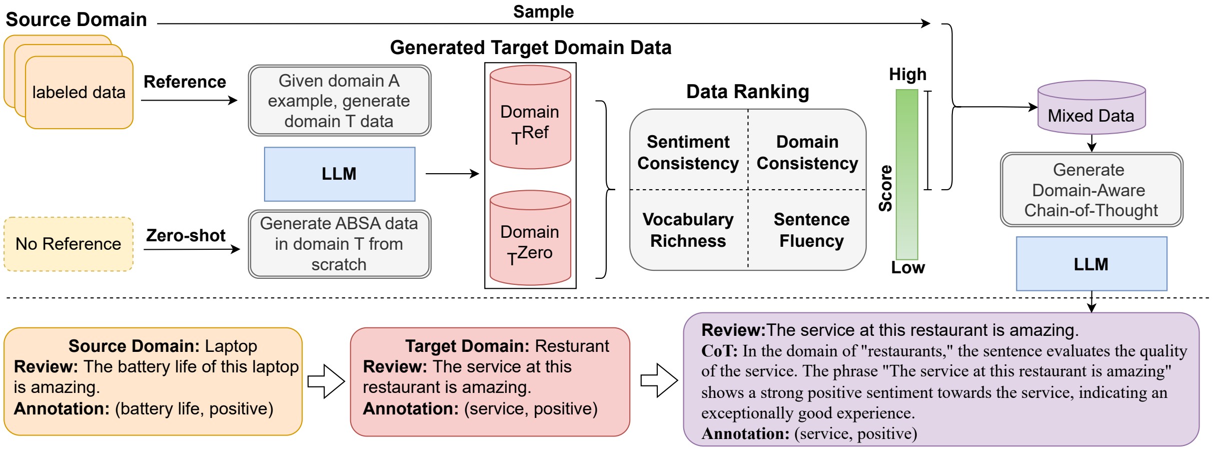
2 Related Work

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 (Natural Language Processing), early data augmen-



Source Domain: Laptop
Review: The battery life of this laptop is amazing.
Annotation: (battery life, positive)

Target Domain: Resturant
Review: The service at this restaurant is amazing.
Annotation: (service, positive)

Review: The service at this restaurant is amazing.
CoT: In the domain of "restaurants," the sentence evaluates the quality of the service. The phrase "The service at this restaurant is amazing" shows a strong positive sentiment towards the service, indicating an exceptionally good experience.
Annotation: (service, positive)

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.

tation 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. The examples we used is presented in Appendix G.

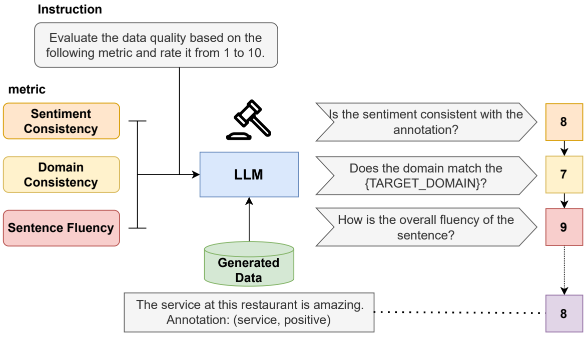

Instruction
Evaluate the data quality based on the following metric and rate it from 1 to 10.
metric
Sentiment Consistency
Domain Consistency
Sentence Fluency
LLM
Generated Data
Is the sentiment consistent with the annotation?
Does the domain match the {TARGET_DOMAIN}?
How is the overall fluency of the sentence?
8
7
9
The service at this restaurant is amazing. Annotation: (service, positive)
8

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

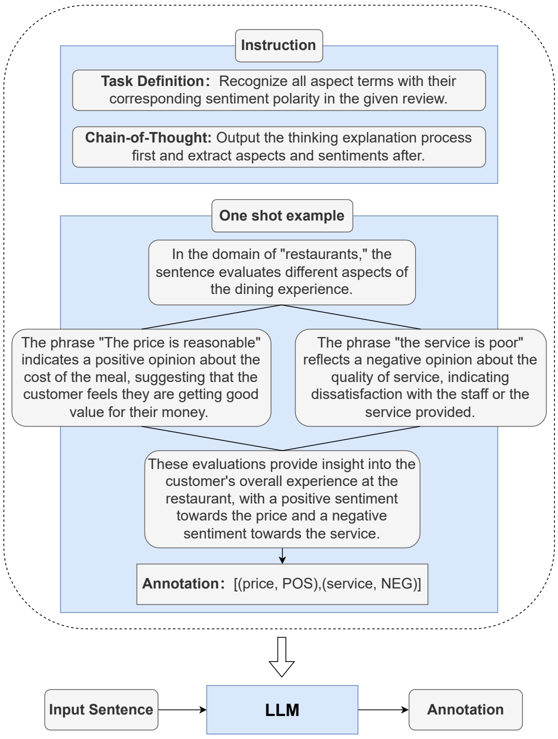

Instruction
Task Definition: Recognize all aspect terms with their corresponding sentiment polarity in the given review.
Chain-of-Thought: Output the thinking explanation process first and extract aspects and sentiments after.
One shot example
In the domain of "restaurants," the sentence evaluates different aspects of the dining experience.
The phrase "The price is reasonable" indicates a positive opinion about the cost of the meal, suggesting that the customer feels they are getting good value for their money.
The phrase "the service is poor" reflects a negative opinion about the quality of service, indicating dissatisfaction with the staff or the service provided.
These evaluations provide insight into the customer's overall experience at the restaurant, with a positive sentiment towards the price and a negative sentiment towards the service.
Annotation: [(price, POS),(service, NEG)]
Input Sentence
LLM
Annotation

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

generated data, our work calculate the Shannon 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. Analysis of the generated data can be found in Appendix E and Appendix B.

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 that it can recognize and understand the context of the target domain, thus 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. Unlike prior approaches such as Kim et al. that allow the model to perform arbitrary-direction reasoning, our Domain-Contextualized CoT explicitly models domain-aware reasoning steps (Figure 3, e.g., "Considering the rest domain, the aspect 'deliver' likely relates to..."). This design ensures that the model adapts its reasoning process to domain-specific nuances, whereas previous methods lack explicit mechanisms for domain adaptation.

To ensure that the model strictly follows the

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

Domain-Contextualized CoT 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. To ensure that the model strictly follows the Domain-Contextualized CoT format, we provide a one-shot example as guidance in the prompt. This one-shot example is provided in Appendix D.1. An illustration of this example is provided 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 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 and Qwen2.5-7B-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. Further experimental details can be found in Appendix D.

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:

- **BERT-UDA** (Gong et al., 2020) An instance-based domain adaption method.
- **BERT-CDRG** (Yu et al., 2021) An method that generates pseudo-labels for target domain texts.
- **BGCA** (Yu et al., 2023) A model that leverages a bidirectional generative framework for data augmentation. We select the *label-to-text* version of the model proposed in this work.
- **DA^{2}LM** (Yu et al., 2023) An approach based on Domain-Adaptive Language Modeling. We select the *GPT* version of the model proposed in this work for comparison.
- **RSDA** (Wang et al., 2024) A method that refines generated labeled data and synthesizes diverse labeled data.

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

Methods	S→R	L→R	D→R	R→S	L→S	D→S	R→L	S→L	R→D	S→D
					Target Domain Needed					
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
DA^{2}LM[*]	58.64	60.39	58.98	40.44	36.84	35.75	42.91	36.97	41.28	40.28
RSDA[*]	56.36	62.78	59.79	44.84	45.27	48.66	46.85	36.59	36.22	37.19
					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
Qwen-base	51.59	46.18	49.22	37.01	34.50	35.95	43.51	37.35	**44.67**	40.28
GPT-4o	55.91	**49.85**	54.37	29.33	26.91	30.09	31.87	34.02	37.32	35.26
ZeroABSA(L)	**60.45**	48.97	**57.49**	46.27	43.83	**51.22**	36.80	38.09	34.08	40.89
ZeroABSA(Q)	52.14	46.75	50.81	**49.62**	**45.92**	46.18	37.92	39.21	35.07	**41.56**

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 † and $*$ denote results from Yu et al. and Wang et al..

domain data.

- **LLaMA-base and Qwen-base** LLaMA-3-8b-instruct and Qwen2.5-7B-Instruct models only fine-tuned on the labeled source domain, employing the same prompts and training format, except for the chain-of-thought component.
- **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.

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. However, we observe that the particularly large gains on the service domain arise because, in zero-shot mode, the untrained generator tends to extract a greater number of {aspect, polarity} pairs—boosting recall at the expense of precision, and the service test set itself has relatively sparse annotated aspects. As a result, extra extractions (which on denser domains would count as false positives) still improve overall F_1.

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

Methods	**S→R**	**L→R**	**D→R**	**R→S**	**L→S**	**D→S**	**R→L**	**S→L**	**R→D**	**S→D**
					Target Domain Needed					
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
DA^{2}LM*	65.78	68.72	63.86	43.41	41.06	38.20	54.55	44.96	44.29	43.24
RSDA*	63.69	69.53	66.74	49.82	51.48	54.45	58.15	47.47	38.25	39.12
					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
Qwen-base	61.22	52.54	53.53	39.84	39.38	38.97	49.09	43.76	49.15	41.38
GPT-4o	69.22	64.90	**66.69**	47.61	45.34	48.30	51.31	**54.76**	40.48	38.78
ZeroABSA(L)	65.98	53.30	63.82	51.99	50.26	55.43	41.45	44.99	36.78	42.07
ZeroABSA(Q)	**70.64**	**65.72**	63.12	**52.45**	**50.80**	**57.80**	43.50	45.62	**50.09**	**45.66**

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 † and ∗ denote results from Yu et al. and Wang et al..

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. More experiment results and analysis could be found in Appendix A, Appendix B and Appendix C.

4.5 Ablation Study

We conducted an ablation study to assess the contribution of individual components in our zero-shot cross-domain ABSA method. We employ the LLaMA-3-8b-instruct model for the task of End-to-End ABSA. The source and target domains are "restaurant" and "service", respectively. To assess the impact of each component, we evaluated our method after removing the data ranking, data augmentation, and CoT steps, respectively. The detailed results can be found in Table 4. The model parameters in the ablation study remain unchanged compared to previous experiments; only specific steps and components have been removed.

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. We also conducted further ablation studies on various aspects of the data augmentation and data ranking pipeline, with detailed results available in Appendix A and Appendix D.

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.

5 Conclusion

In this paper, we introduced ZeroABSA, a novel zero-shot framework for cross-domain ABSA that completely eliminates the need for target domain data. Our framework effectively bridges the domain gap by integrating a large language model-driven augmentation module with a Domain-Contextualized Chain-of-Thought (DCCoT) strategy. The LLM module synthesizes high-quality, domain-adaptive data, while DCCoT guides the model to learn domain-invariant features. Our comprehensive experiments demonstrated that ZeroABSA significantly outperforms existing methods, validating its effectiveness and robustness in zero-shot settings.

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.

7 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No.62276110, No.62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors thank the Shanghai AI Laboratory for supporting GPU resources. The authors would also like to thank all the anonymous reviewers for their comments on improving the quality of this paper.

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.

Jieyong Kim, Ryang Heo, Yongsik Seo, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2024. Self-consistent reasoning-based aspect-sentiment quad prediction with extract-then-assign strategy. In *Findings of the Association for Computational Linguistics ACL 2024*, page 7295–7303. 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.

Haining Wang, Kang He, Bobo Li, Lei Chen, Fei Li, Xu Han, Chong Teng, and Donghong Ji. 2024. Refining and synthesis: A simple yet effective data augmentation framework for cross-domain aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10318–10329, Bangkok, Thailand. 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 Additional Experiments

A.1 More Comparison with BERT-UDA

We conducted experiments comparing our model with BERT-UDA trained on augmented data. We compare our methods on the End-to-End ABSA task. The results across various domain transfers (source-to-target) are presented in Table 11. The results demonstrate that our model 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.

$\gamma \downarrow, \alpha \rightarrow$	0.1	0.3	0.5	0.7	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 5: Hyperparameter sensitivity analysis of LLaMA-3-8B-instruct on End-to-End ABSA tasks.

A.2 Hyperparameter Sensitivity Analysis

We conducted a sensitivity analysis to evaluate the effect of different hyperparameters on performance. We used LLaMA-3-8B-instruct as our base model for the End-to-End ABSA task. The source domain was set to "device", and the target domain was set to "restaurant". The results are summarized in Table 5.

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

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

A.3 Multi-Domain Transfer

We conducted preliminary experiments that explored the performance of a multi-domain transfer approach. For each experiment, we tested the performance of LLaMA-3-8B-instruct with three domains combined as the source and the remaining one as the target. The final dataset was the same size as the one used in the main experiment, with each of the three source domains contributing one-third of the data. The results of these experiments, shown in Table 6, indicate that the multi-domain approach yielded an overall F1- score slightly above the average of the individual domain scores.

A.4 Ablation Study without In-Context Examples

In our experiments, for models that cannot be fine-tuned on specific domains, such as GPT-4o, we assess their cross-domain generalization ability by using examples from the source domain as in-context examples. In order to compare the performance of different settings without in-context examples, we evaluate GPT-4o and LLaMA3 in a setting where they are only prompted with the desired output format, without being provided any in-context examples as guidance. Table 7 reports the results across four domains: Restaurant, Laptop, Device, and Service. For comparison, we include the original models, as well as those that are fine-tuned with source domain data or provided with few-shot examples.

From Table 7, we observe that while the models (LLaMA-no 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 in-context 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 the setting without in-context examples. The setting without in-context examples (ICE) corresponds to directly evaluating the model's raw performance in each domain. For comparison, the performance with ICE is computed as the average of the model's results when using source domain in-context examples. For example, the result of LLaMA w/ ICE on the Device domain is the average of LLaMA-base's performance in the R→D and S→D settings, as shown in Table 2.

Method	Rest	Laptop	Device	Service
LLaMA	52.86	31.09	35.82	36.84
GPT-4o	52.99	26.08	32.75	25.94
LLaMA w/ ICE	54.85	32.59	**42.63**	**40.17**
GPT-4o w/ ICE	53.38	32.95	36.29	28.78
Ours	**55.64**	**47.11**	37.85	37.49

Table 7: Performance comparison across domains in the setting without in-context examples.

Table 8: Manual error analysis on 200 randomly sampled LLM-generated instances.

Error Category	Frequency	Representative Example
Hallucinated content or domain irrelevance	12%	“The engine performance was outstanding” in a restaurant review.
Sentiment mismatch	4%	Labeling “The slow service was positive.”
Fluency issues	2%	Fragmented syntax: “Food was good. But service.”

B Qualitative Error Analysis of LLM-Generated Data

Although we employ Shannon entropy and an LLM-as-Judge ranking to filter out low-quality synthetic instances, purely quantitative measures may overlook subtle biases or hallucinations. To complement our quantitative analysis, we randomly sampled 200 examples across all domains and manually categorized the most common error types. Table 8 summarizes the prevalence of each error category and provides representative examples. Overall, 12% of instances contained hallucinated or domain-irrelevant content (e.g., “The engine performance was outstanding” in a restaurant review), 4% exhibited sentiment mismatches (e.g., labeling “The slow service was positive”), and 2% suffered from fluency issues (e.g., “Food was good. But service.”). The full set of annotated examples.

C Implementation Complexity and Deployment Efficiency

Our proposed framework indeed consists of several interlinked modules (data augmentation, Domain-Contextualized Chain-of-Thought (CoT), and LoRA-based fine-tuning), which can incur nontrivial computational and latency overheads in real-world deployment. In the **data augmentation** stage, we rely primarily on GPT API calls to generate synthetic training examples, while in the **model training** stage we adopt LoRA to efficiently fine-tune large language models. To quantify the deployment cost of our CoT component, we conducted an *ablation without CoT*, removing both standard and domain-contextualized CoT from inference in Table 4. This simplification yields only a modest performance degradation ($\Delta\mathrm{F}_1 \approx -2.66$ points), indicating that CoT can be omitted when latency or cost constraints are stringent. Moreover, we measured the inference latency of Domain-Contextualized CoT on LLaMA-3-8B: enabling this component increases end-to-end latency by roughly 150% relative to the base model.

Source Domain	Target-Domain Keyword
Restaurant	Online Restaurant Reviews
Device	Online Device Reviews
Laptop	Online Laptop Reviews
Service	Online Service Reviews

Table 9: Mapping from each source domain to its target-domain keyword used in prompts.

D Implementation Details

D.1 One-Shot Example for Thinking Paths Generation

To enforce adherence to the Domain-Contextualized CoT format, we incorporate a single illustrative example in the prompt to guide the model’s output. Its sole purpose is to prompt the model to follow the desired output format. For all intermediate CoT generations, we use the same example as guidance, as shown below:

> Sentence: The price is reasonable although the service is poor .
>
> Target: [(price, POS),(service, NEG)]
>
> Explaination: In the domain of "restaurants," the sentence evaluates different aspects of the dining experience. First, the phrase "The price is reasonable" indicates a positive opinion about the cost of the meal, suggesting that the customer feels they are getting good value for their money. Second, the phrase "the service is poor" reflects a negative opinion about the quality of service, indicating dissatisfaction with the staff or the service provided. These evaluations provide insight into the customer’s overall experience at the restaurant, with a positive sentiment towards the price and a negative sentiment towards the service.

> Generate explaination of the sentence below following the above example.
>
> Sentence:{Sentence}
>
> Target:{Target}
>
> Explaination:...

D.2 Target Domain Information Generation

As noted in Section 3.3, ZeroABSA requires no target sentences but does need a natural-language description of the target domain. Concretely, we use the following prompt template to synthesize domain-targeted reviews:

```
Translate a review from the
{source_domain} domain to the
{target_domain} domain. Please
maintain the sentence structure
as much as possible while
replacing the subject of the
description.
```

For example, if the source domain is "Device" and the target is "Rest", we substitute domain-specific entities (e.g. "battery life" → "food quality") while preserving sentiment and syntax. This procedure generates labeled examples that reflect the new domain distribution without any human annotations. The specific configuration of domain-specific keywords is shown in Table 9.

D.3 Example Construction with LLM

To generate structurally similar simulated target-domain data, we first manually construct k seed examples (e.g. 3 restaurant reviews) in our source domain. We then prompt the LLM to replace both the *aspect term* and the *domain-specific context*—while preserving the original sentence structure and sentiment polarity. For instance:

> **Device Review:** "Finally, Amazon's free shipping is really getting good; it took only three working days for the player to reach me!"
>
> **Generated Restaurant Review:** "Finally, the restaurant's free delivery is really getting good; it took only 30 minutes for the food to reach me!"

Unlike prior BERT-based methods (Yu et al., 2023), our LLM dynamically adjusts its vocabulary and syntax. It therefore produces richer, domain-adaptive samples without requiring an explicit list of target-domain terms. In practice, we generate approximately 1,500 new samples per target domain, striking a balance between diversity and quality. The $k = 3$ seed examples serve solely as in-context prompts; we maintain the same three prompts for all generations in a given domain to ensure format consistency. Detailed information can be found in Appendix G.

D.4 Ablation of LLM-as-Judge

While using the same model for both generation and evaluation could introduce bias, we mitigate this risk through two key strategies: iterative refinement and rank-based filtering, which assesses vocabulary richness, fluency, and domain/sentiment consistency. To further validate our approach, we evaluated the generated restaurant dataset using GPT-4o-mini-all and Gemini-2.5-pro-exp-03-25, more advanced models than the one used for generation. The results showed no significant difference compared to previous evaluations, supporting the reliability of our methodology. Results are shown in Table 10.

Table 10: Zero-Shot Evaluation of ZeroABSA (LLaMA) with Different Judges

Judge	S→R	L→R	D→R
gpt-4o-mini	60.45	48.97	57.49
GPT-o3-mini-all	60.29	49.65	57.22
Gemini-2.5-Pro	59.79	49.16	57.60

E Quantifying Domain Alignment

To validate that our contextualized CoT induces domain-specific reasoning, we compare the aspect–polarity distribution of generated data versus both source and real target data using KL divergence:

$$D_{\mathrm{KL}}\big(p_{\mathrm{gen}} \parallel p_{\mathrm{real}}\big) \quad \text{and} \quad D_{\mathrm{KL}}\big(p_{\mathrm{src}} \parallel p_{\mathrm{real}}\big).$$

We find $D_{\mathrm{KL}}(p_{\mathrm{gen}} \parallel p_{\mathrm{real}}) = 17.038$, which is lower than $D_{\mathrm{KL}}(p_{\mathrm{src}} \parallel p_{\mathrm{real}}) = 17.418$ (Δ KL = 0.38). Although absolute values remain high—reflecting the inherent complexity of cross-domain aspect distributions—the relative reduction demonstrates that our hybrid augmentation better bridges the gap to the target distribution. For instance, in the device → restaurant task, generated reviews emphasize "service speed" and "food freshness" (mirroring real target examples) rather

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 (L)	**60.45**	**48.97**	**57.49**	**46.27**	**43.83**	**51.22**	**36.80**	**38.09**	**34.08**	**40.89**

Table 11: Comparison of our model with BERT-UDA on End-to-End ABSA tasks. Our model significantly outperforms BERT-UDA.

than "battery life," confirming that the model internalizes domain-specific CoT patterns during generation.

F Future Work

While our current focus is on extracting (aspect, polarity) pairs for fair comparison with prior cross-domain ABSA work, our LLM-based framework can readily extend to extract (aspect, opinion, polarity) triplets via prompt modifications. In addition, exploring output formats that are better aligned with the nature of extraction tasks may further improve performance—for instance, by employing techniques from controllable text generation. We leave a comprehensive study of this extension to future work.

G Domain-Transfer Examples

- **device**
  - laptop
    1. Query: "the high resolution screen is easy to read and has a backlight." A: "The high resolution display is easy to read and has a backlight."
    2. Query: "i received the apex ad-2600 before i expected." A: "I received the laptop before I expected."
    3. Query: "the mms technology is very well integrated with this phone, which you will enjoy." A: "The MMS software is very well integrated with this laptop, which you will enjoy."
  - rest
    1. Query: "the high resolution screen is easy to read and has a backlight." A: "The menu is easy to read and has a clear layout."
    2. Query: "i received the apex ad-2600 before i expected." A: "I received my reservation confirmation before I expected."
    3. Query: "the mms technology is very well integrated with this phone, which you will enjoy." A: "The reservation system is very well integrated with this restaurant, which you will enjoy."
  - service
    1. Query: "the high resolution screen is easy to read and has a backlight." A: "The user interface is easy to read and has a clear design."
    2. Query: "i received the apex ad-2600 before i expected." A: "I received the service request confirmation before I expected."
    3. Query: "the mms technology is very well integrated with this phone, which you will enjoy." A: "The notification system is very well integrated with this service, which you will enjoy."
- **laptop**
  - laptop (same as device→laptop)
  - rest (same as device→rest)
  - service(same as device→service)
  - device
    1. Query: "I love the operating system and the preloaded software." A: "I love the firmware and the preloaded applications."
    2. Query: "And these are some reasons you should get a macbook pro." A: "And these are some reasons you should get this device."
    3. Query: "I had something else go wrong... and buy the warranty." A: "I had something else go wrong... and buy the service plan."
- **rest**
  - laptop
    1. Query: "After really enjoying ourselves at the bar... had dinner." A: "After really enjoying ourselves at the keyboard... continued working."
    2. Query: "Nice Family owned traditional restaurant." A: "Nice family-owned traditional laptop."
    3. Query: "The Dim Sum was so-so, but not spectacular." A: "The battery life was so-so, but not spectacular."
  - device
    1. Query: "After really enjoying ourselves at the bar... had dinner." A: "After really enjoying ourselves with the speakers... continued using it."
    2. Query: "Nice Family owned traditional restaurant." A: "Nice family-owned traditional device."
    3. Query: "The Dim Sum was so-so, but not spectacular." A: "The sound quality was so-so, but not spectacular."
  - service
    1. Query: "After really enjoying ourselves at the bar... had dinner." A: "After really enjoying ourselves at the reception... received the service."
    2. Query: "Nice Family owned traditional restaurant." A: "Nice family-owned traditional service."
    3. Query: "The Dim Sum was so-so, but not spectacular." A: "The customer support was so-so, but not spectacular."
- **service**
  - laptop
    1. Query: "I love the idea of this site... MapQuest." A: "I love the idea of this software... the latest updates."
    2. Query: "I want to write about my inside view of E*Trade." A: "I want to write about my inside view of this laptop model."
    3. Query: "Egroups would be 5 stars to me." A: "This laptop would be 5 stars to me."
  - device
    1. Query: "I love the idea of this site... MapQuest." A: "I love the idea of this device... its performance."
    2. Query: "I want to write about my inside view of E*Trade." A: "I want to write about my inside view of this gadget."

3. Query: “Egroups would be 5 stars to me.” A: “This device would be 5 stars to me.”

- rest
  1. Query: “I love the idea of this site…MapQuest.” A: “I love the idea of this restaurant…the service.”
  2. Query: “I want to write about my inside view of E*Trade.” A: “I want to write about my inside view of this dining experience.”
  3. Query: “Egroups would be 5 stars to me.” A: “This restaurant would be 5 stars to me.”