Evaluating Zero-Shot Multilingual Aspect-Based Sentiment Analysis with Large Language Models

Anonymous ACL submission

Abstract

Aspect-based sentiment analysis (ABSA), a sequence labeling task, has attracted increasing attention in multilingual contexts. While previous research has focused largely on fine-tuning or training models specifically for ABSA, we evaluate large language models (LLMs) under zero-shot conditions to explore their potential to tackle this challenge with minimal task-specific adaptation. We conduct a comprehensive empirical evaluation of a series of LLMs on multilingual ABSA tasks, investigating various prompting strategies, including vanilla zero-shot, chain-of-thought (CoT), self-improvement, self-debate, and self-consistency, across nine different models. Results indicate that while LLMs show promise in handling multilingual ABSA, they generally fall short of fine-tuned, task-specific models. Notably, simpler zero-shot prompts often outperform more complex strategies, especially in high-resource languages like English. These findings underscore the need for further refinement of LLM-based approaches to effectively address ABSA task across diverse languages.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a sequence labeling task in natural language processing (NLP) focused on identifying fine-grained information in sentences, such as aspect terms and their associated sentiment polarity (Liu, 2012). Although research on ABSA has achieved success on English texts with large corpora, real-world social media interactions often involve multiple languages (Mao et al., 2022; Zhang et al., 2021b), creating the need for multilingual sentiment analysis.

Figure 1 illustrates a simple multilingual ABSA example, where the same restaurant review is presented in five different languages. A well-trained model should be able to identify the corresponding aspect terms *service* and *food* in each language and predict the associated sentiment polarity (*service* is positive, *food* is negative).

Figure 1: An example of multilingual ABSA task, where LLMs identify and analyze aspect terms across languages, determining the associated sentiments.

With the rise of LLMs, there has been growing interest in leveraging in-context learning approaches for various downstream tasks (Tran and Matsui, 2024; van Schaik and Pugh, 2024; Roumeliotis et al., 2024, *inter alia*). Many studies have evaluated LLMs' capabilities in sequence labeling tasks such as named entity recognition (NER) (Xie et al., 2023) and part-of-speech (POS) tagging, particularly in multilingual settings (Ma et al., 2024; Nie et al., 2024). Similarly, work on LLMs for sentiment analysis has included some assessment of ABSA task (Zhang et al., 2024b). However, few studies have examined LLMs' performance on multilingual ABSA task specifically through various in-context learning strategies. As a result, the extent to which LLMs can effectively handle multilingual ABSA task under these conditions remains unclear. Understanding how well LLMs capture token-level aspect terms and their associated sentiments across languages will offer insights into their multilingual capabilities and aid in designing more effective models for specific tasks.

This work aims to comprehensively evaluate

LLMs on multilingual ABSA tasks by exploring different prompt strategies and languages. Specifically, we aim to answer the following questions: (1) How do LLMs perform on multilingual ABSA task? (2) How do single-turn and multi-turn dialogue prompting paradigms affect the LLM performance on this task? (3) How do LLMs compare with smaller models trained on domain-specific datasets?

To address these questions, we begin by reviewing sequence labeling methods in the context of multilingual ABSA and large language models. Our evaluation spans datasets across multiple domains and five languages. While prior ABSA studies often use a single prompt strategy and language, this narrow evaluation fails to demonstrate LLMs' adaptability. Therefore, it is necessary to perform a more comprehensive evaluation using various prompt strategies and LLMs.

Our experiments reveal several key findings: First, the LLM performance varies across languages, with higher performance in higher-resource languages. Second, when using single-turn and multi-turn dialogue methods requiring reasoning outputs, models show varying degrees of performance degradation, indicating the need for better solutions to optimize the reasoning process. Third, in extracting fine-grained structured information, LLMs still lag behind smaller language models trained on specific datasets. Finally, while LLM performance improves in few-shot settings with more demonstrations, it is constrained by context length, highlighting the need for more efficient solutions in the future.

2 Background

Multilingual ABSA. Multilingual ABSA aims to identify aspects and their associated sentiments in text across multiple languages. This task is complicated by linguistic diversity, cultural differences, and the scarcity of annotated datasets in non-English languages (Pontiki et al., 2016). Early approaches relied heavily on supervised learning models, necessitating manually annotated data for each language (Lin et al., 2014; Klinger and Cimiano, 2015; Lambert, 2015; Barnes et al., 2016). Recent advancements have shifted focus to cross-lingual embeddings and multilingual pre-trained language models, often fine-tuning these on specific languages for ABSA task (Phan et al., 2021; Xu et al., 2019; Zhao and Yu, 2021). While these approaches have enhanced generalization across languages, they still face difficulties when dealing with user-generated content, such as slang, abbreviations, and culturally dependent expressions (Zhang et al., 2021a). The emergence of LLMs offers new opportunities to tackle these challenges, driving our investigation into their performance on such a nuanced multilingual task.

LLMs for Sequence Labeling tasks. LLMs have recently shown promise in various sequence labeling tasks, such as NER and POS tagging (Xie et al., 2023; Zhang et al., 2024b; Wang et al., 2023b). Unlike traditional models that often require task-specific architectures, LLMs benefit from extensive pre-training on diverse data, enabling them to provide flexible solutions across multiple tasks. Techniques such as prompting and fine-tuning have been explored to improve LLMs' applicability to these tasks (Simmering and Huoviala, 2023; Wang et al., 2023c, 2024b). However, several challenges remain, particularly when adapting LLMs to multilingual contexts and balancing trade-offs between model size, computational cost, and performance (Simmering and Huoviala, 2023). In our work, we extend this research by evaluating LLMs' effectiveness on multilingual ABSA, a complex task that requires handling both aspect extraction and sentiment classification across languages.

3 Prompting Strategies

To evaluate the ability of LLMs to perform multilingual ABSA task, we employ a range of prompting strategies, from basic zero-shot prompts to more complex multi-turn reasoning prompts. Given that LLMs are highly sensitive to prompts and may generate varying responses even when the prompts are semantically similar (Perez et al., 2021; Lu et al., 2022; Zhang et al., 2024b), we ensure consistency across each prompting strategy. Figure 2 illustrates five prompting strategies assessed in this study. In all cases, we include a short introduction sentence in the beginning to make the model familiar with the task. Considering the token-level nature of the ABSA task, we include formatting instructions and explicitly instruct the models to output the results in a JSON format, containing the aspect term (entity) and its associated sentiment, in the prompts. The expected output structure is as follows:

```
{"entity": "<entity>", "sentiment": "<label>"}
```

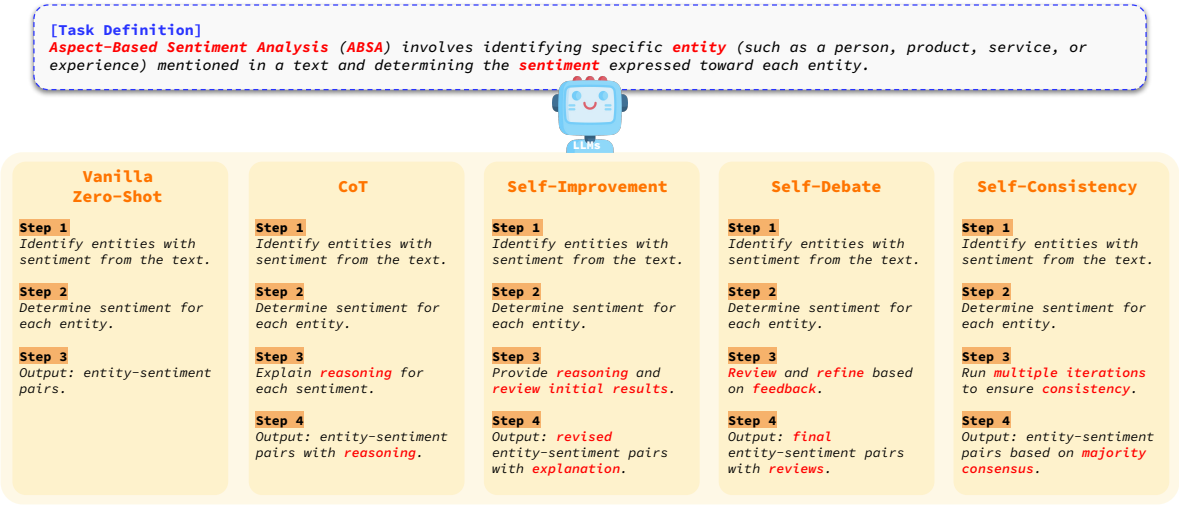

Figure 2: Prompting Strategies used for the experiments.

Vanilla Zero-Shot. The vanilla zero-shot prompt directly asks for the aspect terms and their sentiments, with a given input sentence. This is the major prompting method conducted in Zhang et al. (2024b).

CoT. Chain-of-Thought (CoT) prompting encourages the model to generate a sequence of intermediate reasoning steps before arriving at a final answer, which can significantly enhance its ability to handle complex, multi-step reasoning tasks (Wei et al., 2022; Wang et al., 2023a). In this study, CoT prompting extends the vanilla zero-shot approach by explicitly instructing the model to articulate its reasoning process, thereby providing greater transparency into the decision-making pathway.

Self-Improvement. Drawing from the self-improving framework, which uses unlabeled data and self-annotation and retrieving to enhance performance in zero-shot tasks such as NER (Xie et al., 2024), we introduce a simplified self-improvement prompting strategy. The self-improvement prompting strategy is a three-step, multi-turn approach designed to refine the model's initial output. The first step involves standard zero-shot prompting to generate an initial response. In the second step, the model is asked to explain its reasoning behind the initial response, encouraging introspection and evaluation of its own outputs. Finally, the model is prompted with both its initial response and its reasoning to generate a revised, more informed answer. This iterative process helps the model to self-correct and produce revised results.

Self-Debate. The self-debate prompting strategy introduces a simulated dialogue (Wang et al., 2024c; Zhang et al., 2024a) between multiple agents within the model. It begins with a zero-shot prompt to generate an initial response, as in the self-improvement approach. In the second step, the model is prompted to assume the role of a second agent tasked with critically reviewing and, if necessary, correcting the initial response. In the third step, the model acts as a third agent that evaluates both preceding outputs and engages in a simulated debate to arrive at a final decision. This approach can be iterated over multiple rounds to further explore the model's capability for self-assessment and refinement, although for efficiency, we limit this process to three steps.

Self-Consistency. The self-consistency prompting strategy aims to stabilize the model's outputs by running the same experiment multiple times (n iterations) with varying temperature settings. This approach leverages sampling diversity to generate multiple potential outputs, from which the most frequently occurring result is selected as the final answer. The empirical evaluation shows that self-consistency boosts the performance of CoT prompting on a few reasoning benchmarks (Wang et al., 2023c). This self-consistency is also tested to have improved performance on zero-shot NER tasks (Xie et al., 2023).

4 Experimental Setup

Models. To conduct a comprehensive evaluation, we select a diverse range of open-source

and closed-source instruction-tuned LLMs to perform the ABSA task: Llama-3.1 8B (AI@Meta, 2024), Mistral 7B (Jiang et al., 2023), Gemma-2 9B (Team, 2024b), Qwen-2.5 7B (Team, 2024c), Zephyr 7B (Tunstall et al., 2023), Phi-3.5-mini 3.8B (Microsoft, 2024), Gemini-1.5 (Team, 2024a), Claude-3.5 (Anthropic, 2024), and GPT-4o (OpenAI, 2024). These models are chosen for their diverse architectures and tuning strategies, offering a broad assessment of both state-of-the-art open-source and proprietary LLM capabilities.

Decoding Temperature. For all LLMs used in our experiments, unless stated otherwise, we set the decoding temperature to 0.0, employing a greedy decoding strategy to minimize random variations in the results. In the Self-Consistency setup, where the goal is to identify the most consistent output, we experiment with a range of temperatures {0.2, 0.4, 0.6, 0.8, 1.0}, in addition to the default 0.0 setting.

Dataset. We use the SemEval-2016 dataset (Pontiki et al., 2016) to perform a comprehensive evaluation of the selected models. This dataset consists of real user reviews across eight languages, with ABSA annotations available for English (EN), French (FR), Spanish (ES), Dutch (NL), Russian (RU), and Turkish (TK). However, due to the limited size of the Turkish test set (fewer than 150 sentences), we exclude it from our evaluation, consistent with prior multilingual ABSA studies (Zhang et al., 2021a; Lin et al., 2023). The data statistics are shown in Appendix A.4.

Baselines. Due to the limited research on multilingual or cross-lingual ABSA task in the past two years, we adopt commonly used supervised fine-tuning baselines, as used in Zhang et al. (2021a); Lin et al. (2023). Specifically, we fine-tune two multilingual models, mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), on the multilingual data set as baselines. These models provide strong points of comparison for our zero-shot and few-shot LLM prompting strategies. Detailed parameter settings for the fine-tuning process are provided in Appendix A.3.

Evaluation Metric. We evaluate five languages using the Micro-F1 metric on LLMs, which is well-suited for tasks requiring both aspect and sentiment prediction. In ABSA task, a prediction is considered correct only if both the aspect term and its associated sentiment match the ground truth. This ensures that the evaluation reflects the model's ability to accurately identify both components.

5 Results

5.1 Overall Observation

Table 1 summarizes the performance of five methods, including zero-shot, CoT, self-improvement, self-debate, and self-consistency methods on the multilingual ABSA task. For each method, we report the performance across five different languages, as well as the average performance, to provide a clearer comparison of the models' effectiveness under different approaches. For multi-turn dialogues, we select the output from the last turn as the final result.

Prompt Sensitivity Across Models, Methods, Languages in Performance. Different prompt designs can lead to significant performance variations. To fairly investigate the impact of this sensitivity on the multilingual ABSA task, we establish uniform performance variations for prompt evaluation for each method. For fine-grained structured outputs, the impact of prompts on performance under the same prompt is significant. Generally, we observe that simple zero-shot prompts lead to better performance than other complex prompts in most cases. In terms of overall performance, GPT-4o achieves the best average performance in all five methods, reaching an F1 score of 45 95%, which can be attributed to the stronger capabilities of the model and its use of a more diverse multilingual corpus during pre-training, leading to better handling of downstream tasks. When comparing the performance between the languages, we clearly observe that the results on English are the best, followed by Spanish and French, and the results on Russian are generally the worst. This is not surprising, as current LLMs are mostly English-centric (Nie et al., 2024) and we apply English prompts during the experiments. Details about the prompts can be found in the Appendix A.1.

Challenges in CoT Reasoning for Multilingual ABSA. For single-turn dialogues, we can observe that the zero-shot method significantly outperforms the CoT method across all benchmarks (with an average 1.55% Micro-F1 difference). In multilingual ABSA task, the quality and structure of the dataset are critical to the performance of sentiment classification models, especially when we apply step-by-step reasoning methods. However, previous studies (Wang et al., 2024a) have shown

Methods	Lang.	Claude-3.5	Gemma-2	Gemini-1.5	GPT-4o	Llama-3.1	Mistral	Phi-3.5	Qwen-2.5	Zephyr	Avg.
Zero-Shot	EN	53.37	53.95	60.39	55.81	45.60	48.26	52.19	54.64	39.15	51.49
	ES	44.36	48.29	48.80	49.91	32.49	38.32	40.08	48.10	33.48	42.65
	FR	42.73	48.88	50.94	48.43	23.15	37.21	39.81	43.60	32.21	40.77
	NL	42.03	46.75	50.24	49.94	33.53	33.98	37.41	42.59	26.79	40.36
	RU	33.90	40.25	39.34	45.15	30.18	26.58	28.71	37.15	22.67	33.77
	Avg.	43.28	47.63	49.94	49.85	32.99	36.87	39.64	45.22	30.86	41.81
CoT	EN	45.14	56.08	55.85	59.84	45.95	44.14	47.81	52.62	44.50	50.22
	ES	40.58	48.99	45.38	54.01	32.63	33.85	37.58	44.21	32.81	41.12
	FR	34.50	49.80	43.27	52.28	23.77	32.10	37.40	38.36	33.14	38.29
	NL	37.84	47.73	46.78	53.06	34.83	27.43	31.18	39.36	29.28	38.61
	RU	34.31	40.56	38.14	44.88	31.17	25.86	24.79	31.87	26.24	33.09
	Avg.	38.48	48.63	45.88	52.81	33.67	32.68	35.75	41.28	33.19	40.26
Self-Improve.	EN	51.69	47.69	53.07	50.18	38.83	38.10	48.68	49.14	40.61	46.44
	ES	42.69	42.56	43.66	42.83	31.10	31.50	38.29	40.84	34.38	38.65
	FR	42.05	42.42	45.70	41.09	21.72	29.80	37.87	38.23	32.75	36.85
	NL	38.52	37.53	42.91	37.17	30.71	26.20	34.68	34.72	27.45	34.43
	RU	34.26	33.51	35.05	34.69	26.86	21.23	27.03	29.11	23.73	29.50
	Avg.	41.84	40.74	44.08	41.19	29.84	29.37	37.31	38.41	31.78	37.17
Self-Debate	EN	51.44	40.93	46.86	45.86	25.30	27.00	45.88	42.59	35.78	40.18
	ES	40.03	36.89	37.70	39.45	18.67	22.69	37.08	37.05	31.33	33.43
	FR	39.02	36.90	40.94	38.11	15.43	23.80	34.54	32.02	27.36	32.01
	NL	38.41	34.79	37.67	36.24	20.79	19.40	32.14	29.81	23.36	30.29
	RU	31.58	29.95	32.39	32.60	19.82	16.62	24.33	25.43	19.86	25.84
	Avg.	40.10	35.89	39.11	38.45	20.00	21.90	34.79	33.38	27.54	32.35
Self-Consist.	EN	48.21	49.95	56.07	56.60	38.18	42.54	43.18	52.09	29.29	46.23
	ES	36.84	44.49	41.25	46.28	24.93	29.04	29.08	44.58	22.06	35.39
	FR	38.84	45.98	44.13	47.25	21.87	27.55	30.07	41.82	20.09	35.29
	NL	34.72	43.31	42.68	45.61	25.10	24.46	25.88	39.66	16.25	33.07
	RU	29.97	36.56	34.73	41.62	24.18	19.74	20.85	33.97	14.67	28.48
	Avg.	37.72	44.06	43.77	47.47	26.85	28.67	29.81	42.42	20.47	35.69

Table 1: Performance comparison of various methods on different languages.

that the SemEval-2016 dataset (Pontiki et al., 2016) contains a significant number of sentences with incomplete structures, which lack explicit sentiment-bearing entities or grammatical completeness, making it difficult to find explicit evidence from the sentence. When encountering incomplete sentences, the CoT method may attempt to "force" the generation of reasoning steps, leading to over-extrapolation or reasoning errors.

Query Text: Yum !.
Result: {"entity": "Yum","sentiment": "positive", "explanation":"The text 'Yum !' expresses a positive sentiment towards an implicit entity, likely referring to food or a dining experience. The exclamation mark further emphasizes the positive sentiment."},
Ground True:{[]}

Figure 3: Case study of an incomplete sentence in CoT scenario. Results are extracted from the Claude-3.5 output.

For example, we present a representative example in Figure 3 where the sentence is composed of phrases and punctuation marks with a positive sentiment polarity. The CoT method fails to identify explicit evidence within the sentence, leading to over-reasoning and ultimately resulting in an incorrect inference. Moreover, CoT methods not only

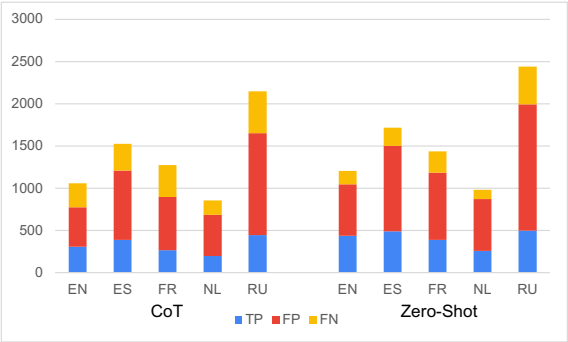

Figure 4: Comparison of True Positives, False Positives, and False Negatives of Claude-3.5 outputs in Zero-Shot and CoT Methods.

require the model to identify sentiment-polarized entities but also to generate a detailed reasoning process. However, the model often produces incorrect reasoning chains or fails to find reasoning evidence when the sentences lack clear sentiment clues, leading to entity omission. In Figure 4, we present the statistics of true positives, false positives, and false negatives for Claude-3.5 under both zero-shot and CoT methods, further illustrating our conclusions. Compared to the zero-shot method, the CoT method exhibits a significant de-

crease in the number of correct predictions and over-predictions, as well as a noticeable increase in the number of entity omissions, which aligns with our hypothesis.

Multi-Turn vs. Single-Turn in Performance. As seen in Table 1, compared to the zero-shot single-turn dialogue method, the multi-turn dialogue method (self-improvement & self-debate) exhibits a significant performance drop across all languages. The best average performance achieves an F1 score of 37.17%, which is 4.64% lower than that of the zero-shot method. The performance variation indicates that while multi-turn dialogues can enrich the interaction by allowing for more detailed exchanges, they may inadvertently lead to misunderstandings or misinterpretations of the sentiment expressed. This is particularly evident in languages where sentiment cues are heavily context-dependent. We will delve deeper into these findings and discuss the potential implications of adopting multi-turn dialogue strategies for multilingual aspect-based sentiment analysis in §5.3.

Self-Consistency in Different Temperatures. Our analysis of self-consistency in multilingual ABSA task reveals a decrease in performance when averaging the outputs from five different temperature settings (0 to 1) compared to using a singular temperature of 0. The average performance decreased by 5.26%, 7.26%, 5.48%, 7.29%, and 5.29% respectively in five languages compared to the initial zero-shot setting (greedy decoding, temperature=0), indicating that while the varying temperature can introduce diversity in responses, it may also lead to inconsistencies that negatively impact overall performance. These results emphasize the importance of carefully selecting temperature settings to maintain output stability in multilingual contexts.

5.2 Comparison with Fine-Tuning Baseline

The comparison between prompt-engineering methods with LLMs and fine-tuning method reveals distinct performance trends in ABSA task. As shown in Table 1, we can observe that the prompt-engineering method with LLMs struggles to extract fine-grained structured aspect terms and corresponding sentiment compared to fine-tuning methods. We show the fully-supervised baseline results from the mBERT and XLM-R models in Table 2, as well as the results from the best-performing prompting method and the model (GPT-4o-CoT). We notice, even the best-performing LLM prompting method (GPT-4o-CoT) only reaches around 77% of the average performance achieved by fine-tuned models (compared to XLM-R). This performance gap becomes especially obvious in languages with complex morphological structures and domain-specific datasets such as Russian and French, where fine-tuned models, leveraging task-specific training, exhibit more consistent and reliable results. The uniqueness of prompting lies in its flexibility and adaptability, but applying it effectively in specific scenarios can be challenging. This is because prompt-engineering approaches often require carefully crafted instructions that can vary significantly depending on the task, making consistent performance harder to achieve across diverse settings.

	EN	ES	FR	NL	RU	Avg.
mBERT	66.68	67.97	62.78	56.31	59.49	62.65
XLM-R	72.92	72.74	67.26	64.72	66.66	68.86
GPT-4o-CoT	59.84	54.01	52.28	53.06	44.88	52.81

Table 2: The fully-supervised baseline results, compared with GPT-4o-CoT results.

5.3 The Effect of Multi-Turn in Performance

As discussed in §5.1, we observe that the performance of both multi-turn approaches, i.e. self-improvement and self-debate is worse than the zero-shot strategies. To better understand this outcome, we analyze the performance across different rounds. Figure 5 presents the average performance differences across all languages in the three rounds for both multi-turn setups. Detailed performance for each language is provided in Appendix A.2.

Interestingly, for all LLMs, we find that performance consistently declines with each additional round. Notably, in the self-debate setup for GPT-4o, the performance between the first and second rounds drops by approximately 10%. One possible explanation for this decline is that during the multi-turn process, when LLMs are prompted to self-correct, they tend to assume that errors exist in the prior responses, even when instructed to only revise incorrect predictions (Bai et al., 2024). This assumption may lead to unnecessary changes and performance degradation.

5.4 Few-Shot Ablation

In addition to the main prompting strategies described in §3, we conduct few-shot experiments

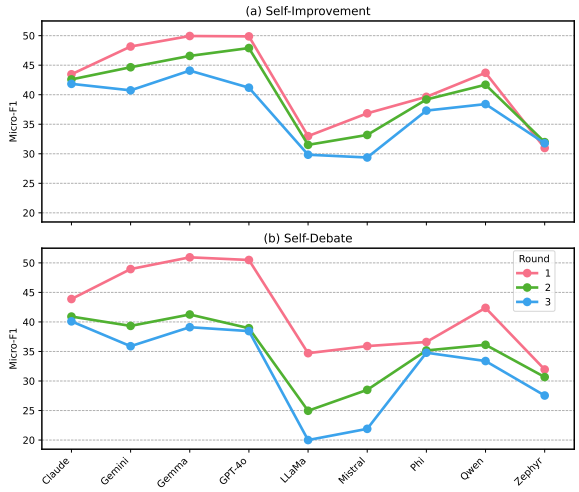

Figure 5: Multi-turn performance across different rounds in different models. In each model, The result is averaged across all the five languages.

using the same setup as the zero-shot experiments. This allows us to examine how including examples in the prompt affects final performance.

In the few-shot setup, the model is provided with n examples from the training data as contextual demonstrations. These examples are selected based on their similarity to the input sentence, as determined using Sentence Transformers (Reimers and Gurevych, 2019). The top n most similar sentences are included as examples, with experiments conducted using the following values for n: $n \in [1, 2, 4, 8, 16]$.

Figure 6 shows the results of the few-shot experiments across different shot sizes for all LLMs. We observe that in most cases, as the number of examples increases, model performance improves accordingly. This supports the assumption that including semantically related examples in the prompt enhances the LLMs' inference capabilities in ABSA task. The exception is Zephyr, whose performance remains unstable and consistently lower than other models, as also noted in the main experiment results in all different setups (see Table 1). Detailed results for the few-shot experiments are presented in Table 7 in the Appendix.

5.5 Error Cases

In this section, we analyze the types of errors observed during the experiments. To classify the errors, we follow the categorization proposed by Xie et al. (2023), which also addresses token-level tasks like NER:

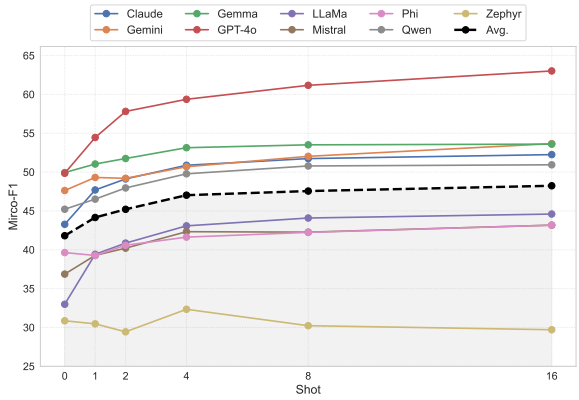

Figure 6: Few-shot results across different shots. In each shot, the result is averaged across all five languages.

- **Out-of-Domain (OOD) types**: Predicted entity types that are not in the predefined label set.
- **Wrong types**: Predicted types that are incorrect but still belong to the label set.
- **Contain gold**: Predicted mentions that include the correct (gold) mentions.
- **Contained by gold**: Predicted mentions that are fully contained within the gold mentions.
- **Overlap with gold**: Predicted mentions that overlap with gold mentions but do not fall into the two previous categories.
- **Completely-O**: Predicted mentions that are entirely outside the gold mentions, yet overlap with other tokens.
- **OOD mentions**: Predicted mentions that do not appear in input text at all.
- **Omitted mentions**: Mentions that should have been predicted but were omitted by the model.

We use the zero-shot performance of the GPT-4o model as a representative example. Figure 8 illustrates the distribution of error types for GPT-4o under the zero-shot setting. The most frequent error types are "Completely-O" and "Omitted mentions", together accounting for over 70% of the total errors. This observation is partially consistent with the findings of Xie et al. (2023) on NER tasks, where "Completely-O" and "Wrong types" are the most common error types. The "Completely-O" errors may result from incomplete annotations or the model's tendency to guess entities based on prior knowledge. The "Omitted mentions" errors likely stem from inadequate contextual understanding.

As shown in Figure 7, we select three typical cases to illustrate reasoning errors, which include OOD mentions, contain gold, and wrong types. In Case 1, the aspect "service" is over-predicted during the first review, which subsequently leads to an incorrect final prediction. Moreover, despite providing a specific task description before each round of debate, the model focuses on the sentiment polarity of the entire sentence rather than the specific aspect terms. Consequently, the entire sentence is generalized as "service is highly satisfactory", re-

Query Text: The entire staff was extremely accomodating and tended to my every need.
Initial Response: {"entity": "staff","sentiment": "positive"}
Review:{"entity": "staff","sentiment": "positive","review": "The sentiment towards 'staff' is appropriately identified as positive, as the description indicates that they were 'extremely accommodating' and attentive to the needs of the person."},
{"entity": "service","sentiment": "positive","review": "The overall experience described, indicated by the phrase 'tended to my every need,' reflects a positive sentiment towards the service provided by the staff."}
Second Round Response: {"entity": "staff","sentiment": "positive","review": "The sentiment towards 'staff' is correctly identified as positive due to the description of them being 'extremely accommodating' and attentive to the needs of the person. This reflects a high level of service and care."},
{"entity": "service","sentiment": "positive","review": "The sentiment towards 'service' is also positive. The phrase 'tended to my every need' indicates that the service provided was attentive and satisfactory, further supporting a positive sentiment overall."}
Ground True:{"entity": "staff","sentiment": "positive"}

Query Text: La comida estuvo muy sabrosa.
Zero-Shot:{"entity": "La comida","sentiment": "positive"}
Self-Consistency: {"entity": "comida","sentiment": "positive"}
Few-Shot:{"entity": "comida","sentiment": "positive"}
CoT: {"entity": "La comida","sentiment": "positive","explanation": "The phrase 'La comida estuvo muy sabrosa' translates to 'The food was very tasty' in English. The word 'sabrosa' indicates that the food is pleasant or enjoyable, which conveys a positive sentiment toward the entity 'la comida'."}
Ground True:{"entity": "comida","sentiment": "positive"}

Type Error: Contain gold

Reasoning Error: The CoT and Zero-Shot methods include the unnecessary definite article "la" within the entity "comida," leading to entity mismatches. The model interprets "la comida" as a single entity rather than focusing on the actual target entity "comida." This demonstrates that the model may overly rely on syntactic structures, neglecting the need for treating entities independently and simplifying the processing in sentiment analysis tasks.

Query Text: Quien sea amante de la carne tiene una carta bastante amplia para elegir, aunque ayer no tenían chuleton .
Initial Response: {"entity": "carne","sentiment": "positive"}, {"entity": "chuleton","sentiment": "negative"}
Commentary:{"entity": "carne","sentiment": "positive","explanation": "The phrase 'amante de la carne' suggests a strong appreciation or enjoyment of meat. The sentiment is inferred to be positive as it indicates that individuals who love meat would find the menu appealing, describing it as 'bastante amplia para elegir' (quite extensive to choose from)."},
{"entity": "chuleton","sentiment": "negative", "explanation": "The mention of 'chuleton' alongside 'aunque ayer no tenían' (although yesterday they did not have) implies a lack of availability of this specific item. This indicates disappointment or a negative sentiment since the absence of a desired item is often viewed unfavorably by customers."}
Second Round Response: {"entity": "carne","sentiment": "positive","explanation": "The phrase 'amante de la carne' indicates a positive sentiment, as it suggests that those who love meat have a favorable view towards the options provided. The description 'bastante amplia para elegir' indicates that there are many appealing choices on the menu, reinforcing the positive sentiment."},
{"entity": "chuleton","sentiment": "neutral","explanation": "While the mention of 'chuleton' is tied to the phrase 'aunque ayer no tenían' (although yesterday they did not have), the sentiment isn't strictly negative or positive. It indicates a simple fact about availability without an explicit emotional response. The tone does not express disappointment or satisfaction, thus it is classified as neutral."}
Ground True:{"entity": "carta","sentiment": "positive"}, {"entity": " chuleton","sentiment":"negative"}

Figure 7: Case study of error increase with the proposed methods. Texts in blue are involved entities in the error cases. The items highlighted are false predictions.

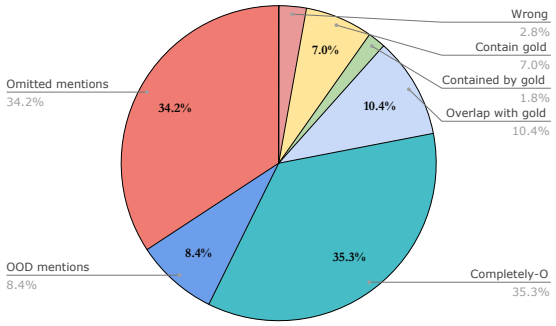

Figure 8: Percentage of different error types with GPT-4o under the zero-shot prompting strategy. The percentages are averaged across the five languages.

sulting in an OOD prediction. We hypothesize that the model suffers from catastrophic forgetting during complex contextual reasoning. In Case 2, the CoT method first translates the original sentence back into English before making predictions, focusing heavily on analyzing the sentence's structural features. This leads to an incorrect identification of "La comida" as an aspect term. We can explain this as the CoT method places too much emphasis on syntactic structure analysis rather than accurately identifying aspect boundaries. In Case 3, after two rounds of self-improvement corrections, the model incorrectly revises a previously correct prediction, changing the sentiment polarity of "chuleton" to neutral. We suspect that the model encounters ambiguity in the implicit sentiment reasoning process for the aspect "chuleton" and lacks confidence during self-correction, ultimately leading to an incorrect prediction type.

Overall, the errors in the selected cases mentioned can be attributed to the following reasons: neglecting the reasoning of entity boundaries during the inference process, error accumulation from multiple rounds of dialogue, and ambiguity in implicit sentiment reasoning.

6 Conclusion

In this study, we conduct an empirical evaluation of zero-shot multilingual ABSA using different LLMs, exploring their performance across five different prompting methods. Inspired by LLMs' strong reasoning capabilities in logical and arithmetic tasks, we apply these strategies and assess their effectiveness across five languages: English, French, Spanish, Dutch, and Russian, as well as domain-specific datasets. Our findings indicate that LLMs do not surpass fully-supervised multilingual models in the ABSA task. Performance varies significantly across languages, with higher-resource languages like English yielding better results. We also find that closed-source models generally outperform open-source ones, and simpler zero-shot prompts prove more effective than more complex prompting strategies. These insights suggest that while LLMs show potential for multilingual tasks, their effectiveness remains highly dependent on the task, language, and model configuration, highlighting the need for further exploration to optimize LLMs for cross-lingual tasks like ABSA.

Limitations

This study on multilingual ABSA has several limitations that highlight areas for future investigation. First, due to constraints in existing multilingual ABSA datasets, we use only the SemEval-2016 dataset, which covers five languages for the ABSA task. While this dataset provides a useful foundation, its limited size and language coverage restrict us from evaluating the model's generalizability across a broader range of languages. Second, the prompt design in our study is limited to fixed English prompts without incorporating variations across other languages. This approach may overlook potential sensitivities to language-specific nuances in prompt formulation, which could impact model performance in non-English contexts. Finally, our exploration is confined to zero-shot prompting without investigating fine-tuning techniques. While prompt engineering is an accessible approach, incorporating fine-tuning could potentially yield more substantial performance gains. Future research should consider employing diverse multilingual datasets and exploring fine-tuning methods to comprehensively assess and improve LLM performance in multilingual ABSA task.

References

AI@Meta. 2024. Llama 3.1 model card.

Anthropic. 2024. Claude 3.5 sonnet model card addendum. Technical report, Anthropic. Technical Report.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Jeremy Barnes, Patrik Lambert, and Toni Badia. 2016. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Roman Klinger and Philipp Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 153–163, Beijing, China. Association for Computational Linguistics.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 781–787, Beijing, China. Association for Computational Linguistics.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2935–2946.

Zheng Lin, Xiaolong Jin, Xueke Xu, Weiping Wang, Xueqi Cheng, and Yuanzhuo Wang. 2014. A cross-lingual joint aspect/sentiment model for sentiment analysis. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1089–1098.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Bolei Ma, Ercong Nie, Shuzhou Yuan, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schuetze. 2024. ToPro: Token-level prompt decomposition for cross-lingual sequence labeling tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2685–2702, St. Julian's, Malta. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2022. A joint training dual-mrc framework for aspect based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 13543–13551.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Ercong Nie, Shuzhou Yuan, Bolei Ma, Helmut Schmid, Michael Färber, Frauke Kreuter, and Hinrich Schütze. 2024. Decomposed prompting: Unveiling multilingual linguistic structure knowledge in english-centric large language models. *Preprint*, arXiv:2402.18397.

OpenAI. 2024. Gpt-4o model card.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.

Khoa Thi-Kim Phan, Duong Ngoc Hao, Dang Van Thin, and Ngan Luu-Thuy Nguyen. 2021. Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. Llms in e-commerce: a comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Paul F. Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *Preprint*, arXiv:2310.18025.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2403.05530.

Qwen Team. 2024c. Qwen2.5: A party of foundation models.

Vu Tran and Tomoko Matsui. 2024. Improving llm prompting with ensemble of instructions: A case study on sentiment analysis. In *JSAI International Symposium on Artificial Intelligence*, pages 299–305. Springer.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Tempest A van Schaik and Brittany Pugh. 2024. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2832–2836.

Bing Wang, Ximing Li, Zhiyao Yang, Yuanyuan Guan, Jiayin Li, and Shengsheng Wang. 2024a. Unsupervised sentence representation learning with frequency-induced adversarial tuning and incomplete sentence filtering. *Neural Networks*, 175:106315.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Qianlong Wang, Hongling Xu, Keyang Ding, Bin Liang, and Ruifeng Xu. 2024b. In-context example retrieval from multi-perspectives for few-shot aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8975–8985, Torino, Italia. ELRA and ICCL.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024c. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593, Mexico City, Mexico. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024a. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024b. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

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

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Anping Zhao and Yu Yu. 2021. Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, 227:107220.

A Appendix

A.1 Prompt

We show all of our prompts for six methods. The prompts are in Figure 9, 10 and 11.

A.2 Detailed Results

The details of the overall performance under the self-improvement, self-debate, and few-shot settings are shown in Table 5, 6 and 7.

A.3 Fine-Tuning Parameter Setting

The details of the parameter setting under supervised methods are shown in Table 3.

A.4 Dataset Details

The details of the dataset statistics in each language are shown in Table 4. # S and # A denote the number of sentences and aspect terms in different sets, respectively.

Parameter	**Value**
Epoch	20
Batch size	[8, 16]
Learning rate (BERT)	5e-5
Learning rate (XLM-R)	2e-5
Hidden size	768
Dropout rate	0.1
Max steps	[2000,2500]
Adam epsilon	1e-8
Warm factor	0.1

Table 3: Fine-tuning hyper-parameter settings.

		EN	FR	ES	NL	RU
Train	# S	2000	1664	2070	1722	3655
	# A	1743	1641	1856	1231	3077
Test	# S	676	668	881	575	1209
	# A	612	650	713	373	949

Table 4: Statistic of the original dataset.

Zero-Shot & Self-Consistency

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity. Each entity is associated with a sentiment that can be [positive, negative, or neutral].

Your task is to:

1. Identify the entity with a sentiment mentioned in the given text.
2. For each identified entity, determine the sentiment in the label set (positive, negative, or neutral).
3. The output should be a list of dictionaries, where each dictionary contains the entity with a sentiment and its corresponding sentiment. If there are no sentiment-bearing entities in the text, the output should be an empty list.

Example Output format:

["entity": "<entity>", "sentiment": "<label>"]

Please return the final output based on the following text in json format.

CoT

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity. Each entity is associated with a sentiment that can be [positive, negative, or neutral].

Your task is to:

1. Identify the entity with a sentiment mentioned in the given text.
2. For each identified entity, determine the sentiment in the label set (positive, negative, or neutral).
3. Provide a reasoning process for how you identified the entities and assigned their sentiments.
4. The output should be a list of dictionaries, where each dictionary contains the entity with a sentiment and its reasoning process. If there are no sentiment-bearing entities in the text, the output should be an empty list.

Example Output format:

["entity": "<entity>", "sentiment": "<label>", "Explanation": "<reasoning process>"]

Please return the final output base on the following text in json format.

Few-Shot

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity.

Each entity is associated with a sentiment that can be [positive, negative, or neutral].

Your task is to:

1. Identify the entity with a sentiment mentioned in the given text.
2. For each identified entity, determine the sentiment in the label set (positive, negative, or neutral).
3. The output should be a list of dictionaries, where each dictionary contains the entity with a sentiment and its corresponding sentiment. If there are no sentiment-bearing entities in the text, the output should be an empty list.

Example Output format:

["entity": "<entity>", "sentiment": "<label>"]

Here are {top_k} similar sentences from the training set:

{similar_sentences and corresponding entity with its sentiment}

Please return the final output based on the following text in json format.

Figure 9: Single-turn prompts used in this paper.

Round 1

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity. Each entity is associated with a sentiment that can be [positive, negative, or neutral].

Your task is to:

1. Identify the entity with a sentiment mentioned in the given text.
2. For each identified entity, determine the sentiment in the label set (positive, negative, or neutral).
3. The output should be a list of dictionaries, where each dictionary contains the entity with a sentiment and its corresponding sentiment. If there are no sentiment-bearing entities in the text, the output should be an empty list.

Example Output format:

["entity": "<entity>", "sentiment": "<label>"]

Please return the final output based on the following text in json format.

Round 2

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity. Each entity is associated with a sentiment that can be [positive, negative, or neutral].

Here is the sentence: {sentence}. You have classified the sentiment of the entities in this sentence.

Here is your initial result: {initial_result}.

Please explain why you classified them in this way.

Example Output format:

["entity": "<entity>", "sentiment": "<label>", "Explanation": "<reasoning process>"]

Please return the final output based on the above sentence in json format.

Round 3

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity.

Each entity is associated with a sentiment that can be [positive, negative, or neutral].

Here is the sentence: {sentence}. You have given a sentiment classification for the entities in this sentence.

Here is your initial result: {initial_result}.

Here is your explanation: {explanation_result}.

Please recheck your classification and explanation. If you find any errors or better classifications, please update your response accordingly.

Example Output format:

["entity": "<entity>", "sentiment": "<label>", "Explanation": "<reasoning process>"]

Please return the final output based on the above sentence in json format.

Figure 10: Self-improvement prompts used in this paper.

Round 1

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity. Each entity is associated with a sentiment that can be [positive, negative, or neutral].

Your task is to:

1. Identify the entity with a sentiment mentioned in the given text.
2. For each identified entity, determine the sentiment in the label set (positive, negative, or neutral).
3. The output should be a list of dictionaries, where each dictionary contains the entity with a sentiment and its corresponding sentiment. If there are no sentiment-bearing entities in the text, the output should be an empty list.

Example Output format:

Round 1: ["entity": "<entity>", "sentiment": "<label>"]

Please return the final output based on the following text in json format.

Round 2

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity. Each entity is associated with a sentiment that can be [positive, negative, or neutral].

The source sentence is: {sentence}.

The first response result: {previous_response}

Please review and comment on the following response. Provide corrections if necessary or add more details to improve the result.

Example Output format:

Round 2: ["entity": "<entity>", "sentiment": "<label>","Review":"<review>"]

Please return the final output based on the above sentence in json format.

Round 3

Aspect-Based Sentiment Analysis (ABSA) involves identifying specific entity (such as a person, product, service, or experience) mentioned in a text and determining the sentiment expressed toward each entity. Each entity is associated with a sentiment that can be [positive, negative, or neutral].

The source sentence is: {sentence}.

The first response result: {first_response}.

The first commentary result: {first_commentary}.

Based on the initial response and commentary, please further debate and refine the analysis. If there are any conflicting opinions or uncertainties, resolve them (both entity and review) and provide a more detailed and accurate response.

Example Output format:

Round 3: ["entity": "<entity>", "sentiment": "<label>","Review":"<review>"]

Please return the final output based on the above sentence in json format.

Figure 11: Self-debate prompts used in this paper.

Round	Lang.	Claude-3.5	Gemma-2	Gemini-1.5	GPT-4o	Llama-3.1	Mistral	Phi-3.5	Qwen-2.5	Zephyr	Avg.
1	EN	54.87	57.35	61.14	53.72	47.05	47.81	48.39	53.20	40.45	51.55
	ES	43.02	48.60	49.84	51.71	35.18	36.42	39.24	45.75	34.36	42.68
	FR	43.85	49.44	50.74	49.18	25.41	36.48	36.13	42.13	32.64	40.67
	NL	42.55	48.16	52.05	51.80	34.68	31.82	33.22	38.21	28.04	40.06
	RU	35.13	41.19	40.92	46.07	31.20	27.03	26.00	32.61	24.28	33.82
	Avg.	43.88	48.95	50.94	50.49	34.71	35.91	36.60	42.38	31.95	41.76
2	EN	52.64	45.19	49.10	46.40	34.05	36.47	47.61	46.96	40.66	44.34
	ES	40.25	39.91	40.44	40.03	24.96	30.48	37.00	39.91	32.68	36.18
	FR	39.96	40.43	41.85	38.65	17.48	28.68	34.65	34.14	30.97	34.09
	NL	39.62	38.55	40.09	36.76	25.78	25.33	31.73	31.63	26.77	32.92
	RU	32.07	32.55	34.81	32.87	22.50	21.54	24.85	27.99	22.25	27.94
	Avg.	40.91	39.33	41.26	38.94	24.96	28.50	35.17	36.13	30.67	35.09
3	EN	51.44	40.93	46.86	45.86	25.30	27.00	45.88	42.59	35.78	40.18
	ES	40.03	36.89	37.70	39.45	18.67	22.69	37.08	37.05	31.33	33.43
	FR	39.02	36.90	40.94	38.11	15.43	23.80	34.54	32.02	27.36	32.01
	NL	38.41	34.79	37.67	36.24	20.79	19.40	32.14	29.81	23.36	30.29
	RU	31.58	29.95	32.39	32.60	19.82	16.62	24.33	25.43	19.86	25.84
	Avg.	40.10	35.89	39.11	38.45	20.00	21.90	34.79	33.38	27.54	32.35

Table 5: Detailed performance comparison of multiple runs in self-improvement.

Round	Lang.	Claude-3.5	Gemma-2	Gemini-1.5	GPT-4o	Llama-3.1	Mistral	Phi-3.5	Qwen-2.5	Zephyr	Avg.
1	EN	53.37	56.52	60.39	57.10	45.60	48.20	52.19	53.91	39.66	51.88
	ES	44.57	48.71	48.80	52.25	32.49	38.34	40.08	46.46	33.79	42.83
	FR	43.70	48.74	50.94	49.05	23.15	37.16	39.81	42.94	30.91	40.71
	NL	40.76	46.45	50.24	47.73	33.53	33.98	37.41	40.70	26.76	39.73
	RU	34.91	40.38	39.34	43.24	30.18	26.59	28.71	34.54	23.79	33.52
	Avg.	43.88	48.95	50.94	50.49	34.71	35.91	36.60	42.38	31.95	41.76
2	EN	51.48	52.82	55.83	55.73	41.76	42.86	50.06	52.83	41.15	49.39
	ES	43.47	45.59	45.94	50.05	32.07	36.15	40.02	44.72	34.31	41.37
	FR	43.10	46.23	49.22	47.40	22.88	33.68	39.83	41.42	32.97	39.64
	NL	39.94	41.15	44.73	45.55	32.04	30.03	37.71	37.35	27.50	37.33
	RU	34.94	37.44	37.12	40.77	28.80	23.25	28.31	32.10	23.94	31.85
	Avg.	42.59	44.65	46.57	47.90	31.51	33.19	39.19	41.69	31.97	39.92
3	EN	51.69	47.69	53.07	50.18	38.83	38.10	48.68	49.14	40.61	46.44
	ES	42.69	42.56	43.66	42.83	31.10	31.50	38.29	40.84	34.38	38.65
	FR	42.05	42.42	45.70	41.09	21.72	29.80	37.87	38.23	32.75	36.85
	NL	38.52	37.53	42.91	37.17	30.71	26.20	34.68	34.72	27.45	34.43
	RU	34.26	33.51	35.05	34.69	26.86	21.23	27.03	29.11	23.73	29.50
	Avg.	41.84	40.74	44.08	41.19	29.84	29.37	37.31	38.41	31.78	37.17

Table 6: Detailed performance comparison of multiple runs in self-debate.

Shot	Lang.	Claude-3.5	Gemma-2	Gemini-1.5	GPT-4o	Llama-3.1	Mistral	Phi-3.5	Qwen-2.5	Zephyr	Avg.
0	EN	53.37	53.95	60.39	55.81	45.60	48.26	52.19	54.64	39.15	51.49
	ES	44.36	48.29	48.80	49.91	32.49	38.32	40.08	48.10	33.48	42.65
	FR	42.73	48.88	50.94	48.43	23.15	37.21	39.81	43.60	32.21	40.77
	NL	42.03	46.75	50.24	49.94	33.53	33.98	37.41	42.59	26.79	40.36
	RU	33.90	40.25	39.34	45.15	30.18	26.58	28.71	37.15	22.67	33.77
	Avg.	43.28	47.63	49.94	49.85	32.99	36.87	39.64	45.22	30.86	41.81
1	EN	54.77	56.86	59.08	59.06	50.17	47.93	50.12	55.74	39.58	52.59
	ES	48.33	50.37	52.16	57.08	40.07	41.70	40.98	48.23	32.27	45.69
	FR	52.12	50.21	52.19	56.19	37.95	38.65	39.14	48.51	31.39	45.15
	NL	43.30	46.49	48.52	51.26	34.97	36.42	35.00	41.64	24.82	40.27
	RU	39.97	42.62	43.23	48.65	33.92	31.67	31.11	38.50	24.31	37.11
	Avg.	47.70	49.31	51.04	54.45	39.42	39.28	39.27	46.52	30.47	44.16
2	EN	57.07	56.93	59.71	63.88	51.74	50.20	52.95	57.90	36.58	54.11
	ES	50.24	49.62	53.37	59.14	42.28	42.59	41.58	49.74	29.97	46.50
	FR	51.84	50.08	53.53	56.89	39.34	40.22	40.74	50.18	30.56	45.93
	NL	44.88	44.91	48.28	56.46	35.45	36.12	36.65	42.50	25.58	41.20
	RU	41.55	44.48	43.87	52.66	35.51	31.90	30.96	39.48	24.53	38.33
	Avg.	49.12	49.20	51.75	57.81	40.86	40.21	40.58	47.96	29.44	45.21
4	EN	59.39	57.21	61.16	66.87	53.78	52.30	53.68	59.66	40.76	56.09
	ES	52.13	52.69	55.93	62.57	44.37	44.03	43.22	51.91	34.22	49.01
	FR	53.36	51.85	55.26	56.49	42.09	43.03	42.02	53.28	32.45	47.76
	NL	46.49	46.15	47.97	56.40	38.65	37.88	36.10	42.70	28.61	42.33
	RU	43.02	45.52	45.38	54.50	36.51	34.44	33.13	41.34	25.65	39.94
	Avg.	50.88	50.69	53.14	59.37	43.08	42.33	41.63	49.78	32.34	47.03
8	EN	59.34	59.10	61.15	68.84	54.30	51.96	54.24	61.34	39.61	56.65
	ES	53.32	54.49	55.68	63.04	45.64	44.55	44.35	53.42	30.19	49.41
	FR	55.02	53.33	56.22	59.02	43.82	43.44	42.32	53.50	32.00	48.74
	NL	47.60	46.89	48.56	59.09	39.91	36.35	36.86	43.18	25.85	42.70
	RU	43.43	46.36	45.94	55.86	36.79	35.03	33.43	42.52	23.48	40.32
	Avg.	51.74	52.03	53.51	61.17	44.09	42.27	42.24	50.79	30.23	47.56
16	EN	59.61	60.62	61.19	69.67	53.65	52.25	55.27	58.36	36.23	56.32
	ES	54.06	55.00	55.11	64.04	46.07	46.04	44.99	54.05	31.28	50.07
	FR	55.03	55.35	56.31	61.85	44.93	45.38	42.87	54.30	32.24	49.81
	NL	46.79	48.22	48.63	61.13	39.62	37.07	38.83	44.35	24.70	43.26
	RU	45.76	49.11	46.75	58.41	38.73	35.13	33.88	43.66	24.09	41.73
	Avg.	52.25	53.66	53.60	63.02	44.60	43.17	43.17	50.94	29.71	48.24

Table 7: Detailed performance comparison of few-shot results.