

LLaMA-Based Models for Aspect-Based Sentiment Analysis

Jakub Šmíd*, Pavel Přibán*, Pavel Král*,†

University of West Bohemia, Faculty of Applied Sciences,

*Department of Computer Science and Engineering,

†NTIS – New Technologies for the Information Society

Univerzita 2732/8, 301 00 Pilsen, Czech Republic

{jaksmid, pribanp, pkral}@kiv.zcu.cz

<https://nlp.kiv.zcu.cz>

Abstract

While large language models (LLMs) show promise for various tasks, their performance in compound aspect-based sentiment analysis (ABSA) tasks lags behind fine-tuned models. However, the potential of LLMs fine-tuned for ABSA remains unexplored. This paper examines the capabilities of open-source LLMs fine-tuned for ABSA, focusing on LLaMA-based models. We evaluate the performance across four tasks and eight English datasets, finding that the fine-tuned Orca 2 model surpasses state-of-the-art results in all tasks. However, all models struggle in zero-shot and few-shot scenarios compared to fully fine-tuned ones. Additionally, we conduct error analysis to identify challenges faced by fine-tuned models.

1 Introduction

Aspect-based sentiment analysis (ABSA) aims to extract detailed sentiment information from text (Zhang et al., 2022). ABSA includes four sentiment elements: aspect term (a), aspect category (c), opinion term (o), and sentiment polarity (p). Given the example review “The steak was delicious”, the elements are “steak”, “food quality”, “delicious” and “positive”, respectively.

Initially, ABSA research focused on extracting individual sentiment elements, e.g. aspect term extraction or aspect category detection (Pontiki et al., 2014). Recent research has transitioned towards compound tasks involving multiple sentiment elements, such as aspect sentiment triplet extraction (ASTE) (Peng et al., 2020), target aspect category detection (TASD) (Wan et al., 2020), aspect category opinion sentiment (ACOS) (Cai et al., 2021), and aspect sentiment quad prediction (ASQP) (Zhang et al., 2021a). Table 1 shows the output formats of these ABSA tasks.

Modern ABSA research often utilizes pre-trained language models, mainly focusing on sequence-to-sequence models. Compound ABSA

Task	Output	Example output
ASTE	$\{(a, o, p)\}$	$\{(\text{“steak”, “delicious”, POS})\}$
TASD	$\{(a, c, p)\}$	$\{(\text{“steak”, food quality, POS})\}$
ACOS	$\{(a, c, o, p)\}$	$\{(\text{“steak”, food quality, “delicious”, POS})\}$
ASQP	$\{(a, c, o, p)\}$	$\{(\text{“steak”, food quality, “delicious”, POS})\}$

Table 1: Output format for selected ABSA tasks for a review: “The steak was delicious”. ACOS focuses on implicit aspect and opinion terms in contrast to ASQP.

tasks are typically formulated as text generation problems (Zhang et al., 2021b,a; Gao et al., 2022; Hu et al., 2022; Gou et al., 2023), which allows to solve compound ABSA tasks simultaneously.

Lately, large language models (LLMs), such as ChatGPT (OpenAI, 2022), LLaMA 2 (Touvron et al., 2023b) and Orca 2 (Mitra et al., 2023), have made significant progress across various natural language processing tasks. However, more traditional approaches that fine-tune Transformer-based models with sufficient data have shown superior performance over ChatGPT in compound ABSA tasks (Zhang et al., 2023; Gou et al., 2023). Additionally, fine-tuning LLMs on a single GPU is challenging due to their large number of parameters. Techniques like QLoRA (Dettmers et al., 2023) address this challenge using a quantized 4-bit frozen backbone LLM with a small set of learnable LoRA weights (Hu et al., 2021). However, studies have yet to explore the capabilities of fine-tuned open-source LLMs for ABSA.

This paper examines the unexplored potential of LLaMA-based models fine-tuned for English ABSA alongside their performance in zero-shot and few-shot scenarios. Our key contributions include: 1) Introducing the capabilities of fine-tuned LLaMA-based models for ABSA. 2) Conducting a comparative analysis of two LLaMA-based models against state-of-the-art results across four ABSA tasks and eight datasets. 3) Evaluating models’ performance in zero-shot, few-shot, and fine-tuning scenarios, demonstrating the superior performance

of the fine-tuned Orca 2 model, surpassing state-of-the-art results across all datasets and tasks. 4) Presenting error analysis of the top-performing model.¹

2 Related Work

Early ABSA studies focused on predicting one or two sentiment elements (Liu et al., 2015; Zhou et al., 2015; He et al., 2019; Cai et al., 2020) before progressing to more complex tasks involving triplets and quadruplets, such as ASTE (Peng et al., 2020), TASD (Wan et al., 2020), ASQP (Zhang et al., 2021a) and ACOS (Cai et al., 2021).

Recent ABSA research focuses primarily on text generation initiated by GAS (Zhang et al., 2021b). PARAPHRASE (Zhang et al., 2021a) converts labels to natural language. LEGO-ABSA (Gao et al., 2022) explores multi-tasking, DLO (Hu et al., 2022) optimizes element ordering, MVP (Gou et al., 2023) combines differently ordered outputs, and Scaria et al. (2023) adopt instruction tuning.

Gou et al. (2023) and Zhang et al. (2023) show that ChatGPT struggles with compound ABSA tasks in zero-shot and few-shot settings. Simmering and Huoviala (2023) report promising results with close-source LLMs for a single simple ABSA task.

3 Experimental Setup

We employ the 7B and 13B versions of LLaMA 2 (Touvron et al., 2023b) and Orca 2 (Mitra et al., 2023) models from the HuggingFace Transformers library² (Wolf et al., 2020). LLaMA 2 offers models of various sizes tailored for dialogue tasks, building upon the LLaMA framework (Touvron et al., 2023a). Orca 2 extends this collection with enhanced reasoning capabilities.

3.1 Experimental Details

For fine-tuning, we follow recommendations from Dettmers et al. (2023) and use QLoRA with the following settings: 4-bit NormalFloat (NF4) with double quantization and bf16 computation datatype, batch size of 16, constant learning rate of $2e-4$, AdamW optimizer (Loshchilov and Hutter, 2019), LoRA adapters (Hu et al., 2021) on all linear Transformer block layers, and LoRA $r = 64$ and $\alpha = 16$.

¹Code and datasets are available at <https://github.com/biba10/LLaMA-ABSA>.

²<https://github.com/huggingface/transformers>

We fine-tune the models for up to 5 epochs and choose the best-performing model based on validation loss. Following Mitra et al. (2023), we compute loss only on tokens generated by the model, excluding the prompt with instructions.

For zero-shot and few-shot experiments, we use 4-bit quantization of the models. Preliminary experiments indicated that 4-bit quantized models performed similarly to 8-bit quantized models and non-quantized models.

All experiments, including zero-shot and few-shot scenarios, employ greedy search decoding and are conducted on an NVIDIA A40 with 48 GB GPU memory.

3.2 Evaluation Metrics

We use micro F1-score as the primary evaluation metric, chosen based on related work, and report average results from 5 runs with different seeds. We consider a predicted sentiment tuple correct only if all its elements exactly match the gold tuple.

3.3 Tasks & Datasets

We evaluate the LLMs on four tasks: two involving quadruplets (ASQP and ACOS) and two involving triplets (TASD and ASTE). We select two datasets for each task and use the same data splits as previous works for a fair comparison. Table 1 displays the output targets for each task.

We use Rest15 and Rest16 datasets for ASQP in the restaurant domain, initially introduced in SemEval tasks (Pontiki et al., 2015, 2016), later aligned and supplemented by Zhang et al. (2021a). For ACOS, we employ ACOS-Rest and ACOS-Lap datasets from Cai et al. (2021), focusing on implicit aspects and opinions and providing comprehensive evaluation. We use the dataset from Xu et al. (2020) and Wan et al. (2020) for ASTE and TASD, respectively. Table 2 shows the detailed data statistics. ASTE datasets are the only ones that do not include implicit sentiment elements.

3.4 Prompting Strategy & Fine-Tuning

LLMs show varied responses despite similar prompts (Perez et al., 2021; Lu et al., 2022). Our goal is to design simple, clear, and straightforward prompts to standardize evaluations across datasets and ensure consistent assessment of LLMs.

Our prompts define sentiment elements and output format. Sentiment element definitions include the permitted label space, e.g. allowed sentiment

		ASQP		ACOS		TASD		ASTE	
		Rest15	Rest16	Lap	Rest	Rest15	Rest16	Rest15	Rest16
Train	Sentences	834	1,264	2,934	1,530	1,120	1,708	605	857
	Tuples	1,354	1,989	4,172	2,484	1,654	2,507	1,013	1,394
	Categories	13	12	114	12	13	12	0	0
	POS/NEG/NEU	1,005/315/34	1,369/558/62	2,583/1,362/227	1,656/733/95	1,198/403/53	1,657/749/101	783/205/25	1,015/329/50
Dev	Sentences	209	316	326	171	10	29	148	210
	Tuples	347	507	440	261	13	44	249	339
	Categories	12	13	71	13	6	9	0	0
	POS/NEG/NEU	252/81/14	341/143/23	279/137/24	180/69/12	6/7/0	23/20/1	185/53/11	252/76/11
Test	Sentences	537	544	816	583	582	587	322	326
	Tuples	795	799	1,161	916	845	859	485	514
	Categories	12	12	81	12	12	12	0	0
	POS/NEG/NEU	453/305/37	583/176/40	716/380/65	667/205/44	454/346/45	611/204/44	317/143/25	407/78/29

Table 2: Statistics for each dataset. POS, NEG and NEU denote the number of positive, negative and neutral examples, respectively.

Prompt for quadruplet tasks	
<p>According to the following sentiment elements definition:</p> <ul style="list-style-type: none"> - The “aspect term” refers to a specific feature, attribute, or aspect of a product or service on which a user can express an opinion. Explicit aspect terms appear explicitly as a substring of the given text. The aspect term might be “null” for the implicit aspect. - The “aspect category” refers to the category that aspect belongs to, and the available categories include: “ambience general”, “drinks prices”, “drinks quality”, “drinks style_options”, “food general”, “food prices”, “food quality”, “food style_options”, “location general”, “restaurant general”, “restaurant miscellaneous”, “restaurant prices”, “service general”. - The “sentiment polarity” refers to the degree of positivity, negativity or neutrality expressed in the opinion towards a particular aspect or feature of a product or service, and the available polarities include: “positive”, “negative” and “neutral”. “neutral” means mildly positive or mildly negative. Quadruplets with objective sentiment polarity should be ignored. - The “opinion term” refers to the sentiment or attitude expressed by a user towards a particular aspect or feature of a product or service. Explicit opinion terms appear explicitly as a substring of the given text. The opinion term might be “null” for the implicit opinion. <p>Please carefully follow the instructions. Ensure that aspect terms are recognized as exact matches in the review or are “null” for implicit aspects. Ensure that aspect categories are from the available categories. Ensure that sentiment polarities are from the available polarities. Ensure that opinion terms are recognized as exact matches in the review or are “null” for implicit opinions.</p> <p>Recognize all sentiment elements with their corresponding aspect terms, aspect categories, sentiment polarity, and opinion terms in the given input text (review). Provide your response in the format of a Python list of tuples: ‘Sentiment elements: [(“aspect term”, “aspect category”, “sentiment polarity”, “opinion term”), ...]’. Note that “...” indicates that there might be more tuples in the list if applicable and must not occur in the answer. Ensure there is no additional text in the response.</p>	
<p>Input: ““““We have gone for dinner only a few times but the same great quality and service is given.”“““</p> <p>Sentiment elements: [(“service”, “service general”, “positive”, “great”), (“dinner”, “food quality”, “positive”, “great quality”)]</p>	
<p>Input: ““““It is n’t the cheapest sushi but has been worth it every time.”“““</p> <p>Output: Sentiment elements: [(“sushi”, “food prices”, “neutral”, “is n’t the cheapest”), (“sushi”, “food quality”, “positive”, “worth”)]</p>	

Figure 1: Prompt for quadruplet tasks (ASQP and ACOS) with example input, expected output in a green box, and one demonstration enclosed in a dashed box. The demonstrations are used solely in few-shot scenarios.

polarities and aspect categories. The output format describes the expected structure of model responses, allowing us to decode the responses into our desired format. We supplement the prompts with the first ten training examples for a given task for few-shot learning. We use the same prompts for fine-tuning as for zero-shot experiments. Figure 1 illustrates a prompt for quadruplet tasks. Appendix A presents the prompts for the triplet tasks.

During the fine-tuning experiments, we train the model to generate the output in the desired format, as shown in Figure 1.

4 Results

Table 3 shows the results of LLaMA-based models.

The results demonstrate the remarkable potential of Orca 2, especially in its 13B version, which sur-

passes previous benchmarks across all four tasks and eight datasets. Notably, the TASD task shows the most significant improvement, with 6% and 8% enhancements for the Rest15 and Rest16 datasets, respectively. While improvements for other tasks are relatively smaller, they remain noteworthy. There are marginal enhancements, within 1%, for the ASQP and ASTE tasks and the ACOS-Lap dataset. However, the ACOS-Rest dataset sees a significant improvement exceeding 4%, indicating notable progress. The remarkable advancements in the TASD task suggest that predicting opinion terms not included in the TASD task presents the most significant challenge for these models. The larger Orca 2 achieves a substantial improvement of 2.87% on average.

The 7B version of Orca 2 performs similarly to

Method	ASQP		ACOS		TASD		ASTE		AVG
	R15	R16	Lap	Rest	R15	R16	R15	R16	
GAS (Zhang et al., 2021b)	45.98	56.04	-	-	60.63	68.31	60.23	69.05	-
PARAPHRASE (Zhang et al., 2021a)	46.93	57.93	43.51	61.16	63.06	71.97	62.56	71.70	59.85
LEGO-ABSA (Gao et al., 2022)	46.10	57.60	-	-	62.30	71.80	64.40	69.90	-
MvP (Gou et al., 2023)	51.04	<u>60.39</u>	<u>43.92</u>	61.54	64.53	72.76	65.89	<u>73.48</u>	61.69
MvP (multi-task) (Gou et al., 2023)	<u>52.21</u>	58.94	43.84	60.36	64.74	70.18	<u>69.44</u>	73.10	61.60
ChatGPT (zero-shot) (Gou et al., 2023)	22.87	-	-	27.11	-	34.08	-	-	-
ChatGPT (few-shot) (Gou et al., 2023)	34.27	-	-	37.71	-	46.50	-	-	-
Orca 2 7B (zero-shot)	1.19	1.66	0.87	2.52	7.77	9.80	23.04	24.58	8.93
Orca 2 7B (few-shot)	11.34	14.21	4.50	16.00	27.32	34.13	37.70	42.18	23.42
Orca 2 7B	51.50	58.63	43.48	<u>63.01</u>	<u>69.74</u>	<u>76.10</u>	65.62	73.18	<u>62.66</u>
Orca 2 13B (zero-shot)	7.83	10.23	3.20	10.98	15.62	22.84	27.74	31.64	17.46
Orca 2 13B (few-shot)	21.13	23.47	9.10	23.80	32.00	39.08	39.50	44.16	30.16
Orca 2 13B	52.29	60.82	44.09	65.80	70.49	78.82	69.91	74.23	64.56
LLaMA 2 7B (zero-shot)	0.80	1.85	0.05	2.39	2.28	7.45	3.47	5.00	3.21
LLaMA 2 7B (few-shot)	11.20	17.48	2.68	26.43	28.10	33.85	38.88	45.04	25.46
LLaMA 2 7B	42.48	55.46	36.49	57.81	64.80	71.39	57.41	67.69	56.69
LLaMA 2 13B (zero-shot)	7.54	6.86	0.72	7.79	13.65	18.04	17.43	18.66	11.34
LLaMA 2 13B (few-shot)	12.08	19.37	2.36	23.08	35.22	38.80	31.49	38.06	25.06
LLaMA 2 13B	47.16	52.98	38.44	60.92	67.70	74.08	61.95	69.95	59.15

Table 3: F1 scores on eight datasets of ASQP, ACOS, TASD, and ASTE tasks, along with the average score. The best results are in **bold**, and the second-best results are underlined.

the state-of-the-art (SOTA) for most tasks. However, it falls behind by over 2% in the Rest15 dataset and ASTE task. Nonetheless, it notably exceeds previous SOTA results for the TASD task by 3–5%, highlighting the challenge of predicting opinion terms absent in the TASD task. Nevertheless, the smaller Orca 2 performs almost 1% better on average than the previous best results.

Orca 2 significantly outperforms LLaMA 2, with the smaller Orca 2 model even surpassing the larger LLaMA 2 model, underscoring the superior reasoning capabilities of Orca 2. Additionally, it suggests that opting for more advanced but smaller models may be more beneficial than using larger models with less sophistication. The TASD task is the only task LLaMA 2 outperforms previous SOTA results. Compared to previous SOTA results, on average, the larger version is more than 2% worse, and the smaller version is 5% worse.

In zero-shot and few-shot scenarios, both evaluated LLaMA-based models exhibit notably inferior performance compared to their fine-tuned counterparts, particularly in quadruplet tasks. ChatGPT, with significantly more parameters, notably outperforms these models across zero-shot and few-shot scenarios. However, ChatGPT notably underperforms compared to fine-tuned models.

4.1 Error Analysis

To gain insights into the challenges of sentiment prediction, we conduct an error analysis focusing on identifying the most difficult sentiment elements to predict. We manually investigate predictions of 100 random test samples from the best-performing run of Orca 2 with 13B parameters for each dataset. Figure 2 depicts the results of the error analysis.

In most cases, the most challenging element to predict is the opinion term, often comprising multiple words. The model frequently struggles to predict the text span precisely, for instance, predicting “mild” instead of “too mild”. Following closely in difficulty is typically the aspect term, which encounters similar mistakes as opinion terms, but aspect terms are more often just one word, making such errors less frequent. Sentiment polarity proves to be the easiest to predict. However, an exception arises in the ACOS-Lap dataset, where the aspect category emerges as the most challenging due to the extensive category variety of the dataset (81 categories in the test set, compared to only 12 in the restaurant datasets).

The model also occasionally confuses semantically similar aspect categories, such as “restaurant general” with “restaurant miscellaneous” or “keyboard usability” with “keyboard general”.

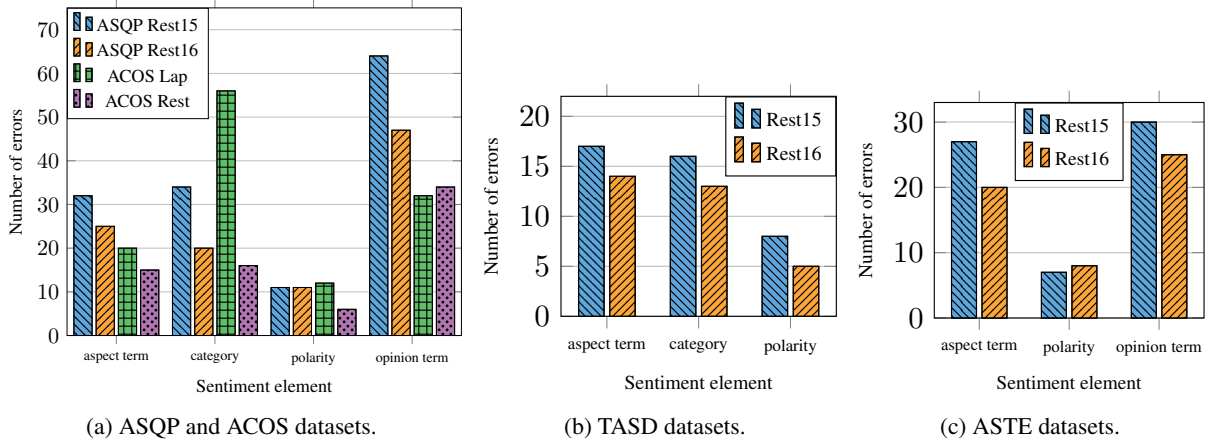


Figure 2: Number of error types for each dataset.

The most common error considering sentiment polarity is in predicting the “neutral” class, possibly due to imbalanced label distribution, since the “neutral” class is the least frequent in all datasets.

In contrast to observations made by Zhang et al. (2021a), we did not encounter errors related to text generation, such as generating words for aspect or opinion terms that are absent in the original text.

Additionally, we identified mistakes in the dataset labels. For example, in the ACOS-Rest dataset, the aspect “service” in the sentence “worst service i ever had” is labelled as “positive”, despite being clearly “negative”, a prediction the model also makes correctly. Similarly, we noticed inconsistencies in the datasets, such as in the sentence “One of the best hot dogs I have ever eaten”, where the expression “hot dogs” is not labelled as an aspect term for the “food quality” category; instead, it is labelled as an implicit aspect term (“NULL”), contrary to other examples. These labelling errors could potentially negatively impact the final scores of evaluated models.

5 Conclusion

This paper presents a comprehensive evaluation of LLaMA-based models for compound ABSA tasks. We show that these models underperform in zero-shot and few-shot scenarios compared to smaller models fine-tuned specifically for ABSA. However, we demonstrate that fine-tuning the LLaMA-based models for ABSA significantly improves their performance, and the best model outperforms previous state-of-the-art results on all eight datasets and four tasks. Error analysis reveals that predicting opinion terms is generally the most challenging for the evaluated models.

Acknowledgements

This work has been partly supported by the OP JAC project DigiTech no. CZ.02.01.01/00/23_021/0008402 and by the Grant No. SGS-2022-016 Advanced methods of data processing and analysis. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Limitations

Results highlight LLaMA-based models’ ineffectiveness in compound ABSA tasks in zero-shot and few-shot scenarios. Additionally, their performance in non-English languages remains unclear. Future work could also consider other open-source models based on a different architecture.

Ethics Statement

We experiment with well-known datasets used in prior scientific studies, ensuring fair and honest analysis while conducting our work ethically and without harming anybody.

References

- Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. [Aspect-category based sentiment analysis with hierarchical graph convolutional network](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 833–843, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. [Aspect-category-opinion-sentiment quadruple extraction](#)

- with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. *Qlora: Efficient finetuning of quantized llms*. *Preprint*, arXiv:2305.14314.
- Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. *LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis*. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. *MvP: Multi-view prompting improves aspect sentiment tuple prediction*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.
- Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. *An interactive multi-task learning network for end-to-end aspect-based sentiment analysis*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*. *Preprint*, arXiv:2106.09685.
- Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. *Improving aspect sentiment quad prediction via template-order data augmentation*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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.
- Ilya Loshchilov and Frank Hutter. 2019. *Decoupled weight decay regularization*. *Preprint*, arXiv:1711.05101.
- 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.
- Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. *Orca 2: Teaching small language models how to reason*. *Preprint*, arXiv:2311.11045.
- OpenAI. 2022. *Openai: Introducing chatgpt*.
- 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.
- 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.
- 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 Eryigit. 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, Haris 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.
- Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. *Instructabsa: Instruction learning for aspect based sentiment analysis*. *arXiv preprint arXiv:2302.08624*.
- Paul F Simmering and Paavo Huoviala. 2023. *Large language models for aspect-based sentiment analysis*. *arXiv preprint arXiv:2310.18025*.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. [Target-aspect-sentiment joint detection for aspect-based sentiment analysis](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9122–9129.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. [Position-aware tagging for aspect sentiment triplet extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.
- Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. [Aspect sentiment quad prediction as paraphrase generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. [Towards generative aspect-based sentiment analysis](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.
- Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. [Representation learning for aspect category detection in online reviews](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

A Prompts

Figure 3 shows the prompt for the TASD task, while Figure 4 presents the prompts for the ASTE task. The prompts are also available in our code.

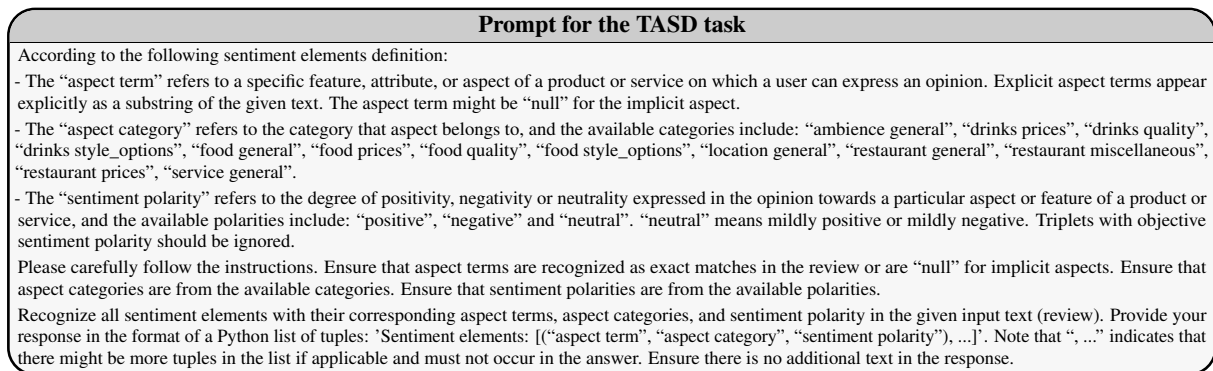


Figure 3: Prompt for the TASD task.

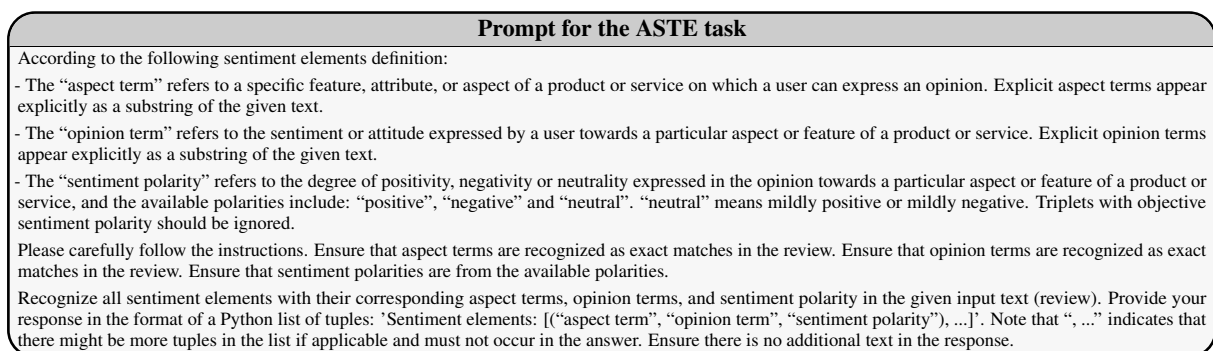


Figure 4: Prompt for the ASTE task.