Large Language Models for Czech Aspect-Based Sentiment Analysis

Jakub Šmíd$^{1,2[0000-0002-4492-5481]}$, Pavel Přibáň$^{1[0000-0002-8744-8726]}$, and
Pavel Král$^{1,2[0000-0002-3096-675X]}$

[1] University of West Bohemia in Pilsen
Faculty of Applied Sciences, Department of Computer Science and Engineering
[2] NTIS – New Technologies for the Information Society
Univerzitni 27328, 301 00 Plzeň, Czech Republic
{jaksmid,pribanp,pkral}@kiv.zcu.cz
https://nlp.kiv.zcu.cz

Abstract. Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task that aims to identify sentiment toward specific aspects of an entity. While large language models (LLMs) have shown strong performance in various natural language processing (NLP) tasks, their capabilities for Czech ABSA remain largely unexplored. In this work, we conduct a comprehensive evaluation of 19 LLMs of varying sizes and architectures on Czech ABSA, comparing their performance in zero-shot, few-shot, and fine-tuning scenarios. Our results show that small domain-specific models fine-tuned for ABSA outperform general-purpose LLMs in zero-shot and few-shot settings, while fine-tuned LLMs achieve state-of-the-art results. We analyze how factors such as multilingualism, model size, and recency influence performance and present an error analysis highlighting key challenges, particularly in aspect term prediction. Our findings provide insights into the suitability of LLMs for Czech ABSA and offer guidance for future research in this area.

Keywords: Aspect-based sentiment analysis · Sentiment analysis · Large language models · Prompting

1 Introduction

Aspect-based sentiment analysis (ABSA) is a natural language processing (NLP) task extends traditional sentiment analysis by targeting specific entities and their aspects, determining sentiment for each rather than providing an overall polarity. ABSA involves three sentiment elements [16]: the aspect term (a), denoting the opinion target; the aspect category (c), representing an attribute of an entity; and the sentiment polarity (p), reflecting the emotional tone. For instance, in the sentence *"Excellent soup"*, these elements correspond to *"soup"*, *"food quality"*, and *"positive"*. Aspect terms may also be implicit, as in *"Tasty!"*.

ABSA tasks vary in complexity depending on which elements they cover. Simple tasks, such as aspect term detection, focus on a single element. Recently more popular compound tasks integrate multiple sentiment elements, such as aspect category sentiment analysis (ACSA) [15], end-to-end ABSA (E2E-ABSA) [26], aspect category term extraction (ACTE) [13], and target-aspect-term-detection (TASD) [25]. Table 1 shows the input and output format of selected ABSA tasks.

Table 1: Outputs of selected ABSA tasks for input: *"Tasty tea but rude staff"*.

Task	**Output**	**Example output**
E2E-ABSA	$\{(a,\ p)\}$	{("tea", POS), ("staff", NEG)}
ACSA	$\{(c,\ p)\}$	{(drinks, POS), (service, NEG)}
ACTE	$\{(a,\ c)\}$	{("tea", drinks), ("staff", service)}
TASD	$\{(a,\ c,\ p)\}$	{("tea", drinks, POS), ("staff", service, NEG)}

While ABSA has been widely studied for English, other languages, including Czech, remain underrepresented. Early Czech ABSA studies [5, 21] relied on now-outdated sentiment classification methods. Recent research [14, 17, 20] has adopted modern Transformer-based [24] approaches.

Large language models (LLMs), such as GPT-4o [11], have transformed NLP via *prompting*, a technique that replaces fine-tuning by guiding model behaviour through task instructions. Few-shot prompting – providing input–output examples – can further enhance performance. However, for complex tasks like ABSA, fine-tuned smaller models still tend to outperform general-purpose LLMs [2, 29]. Although LLM fine-tuning is resource-intensive, methods like QLoRA [3] reduce memory usage, enabling efficient fine-tuning on consumer GPUs. Fine-tuned LLMs using QLoRA have outperformed smaller models for ABSA [18].

Despite these advancements, LLM-based ABSA for languages other than English is still underexplored [16]. Few studies have assessed LLM performance on multilingual ABSA tasks [19, 28], with no comprehensive evaluation for Czech ABSA. This study addresses this gap by evaluating multiple LLMs across zero-shot, few-shot, and fine-tuning scenarios on four Czech ABSA tasks.

Our main contributions include: 1) We provide a comprehensive evaluation of 19 large language models of varying architectures and sizes for Czech aspect-based sentiment analysis, being the first to do so. 2) We compare the zero-shot, few-shot, and fine-tuned performance of LLMs, showing that fine-tuned LLMs achieve new state-of-the-art results, while smaller ABSA-specific models from previous work outperform general-purpose LLMs in zero-shot and few-shot settings. 3) We provide an analysis of the impact of model properties, such as multilingualism, size, and recency, on ABSA performance. 4) We conduct a detailed error analysis identifying key challenges in Czech ABSA, particularly in aspect term prediction.

2 Related Work

Early Czech ABSA research [5, 21] rely on traditional methods like conditional random fields and maximum entropy classifiers. Recent approaches adopt

Transformer-based models. Some enhance ABSA with semantic role labelling in a multitask setup [14], while others explore prompt-based learning and the use of Czech-specific models and in-domain pre-training [17, 20].

LLMs have been evaluated for ABSA, but fine-tuned smaller models often outperform LLMs in zero- and few-shot settings [2, 29]. Fine-tuning LLMs has been shown to improve performance across languages [18, 19, 28], highlighting the value of task-specific fine-tuning.

3 Experimental Setup

We conduct experiments on ACSA, E2E-ABSA, ACTE, and TASD. We utilize the `CsRest-M` dataset [20] consisting of real-world restaurant reviews in Czech designed for compound ABSA tasks, with annotations linking aspect terms, aspect categories, and sentiment polarities. The dataset is already split into training, validation, and test sets. Table 2 shows the statistics of the dataset.

Table 2: Statistics of the dataset.

Count	**Train**	**Dev**	**Test**
Sentences	2,151	240	798
Triplets	4,386	483	1,609

3.1 Models

We utilize two closed-source LLMs and several open-source LLMs of varying sizes. Table 3 provides an overview of the models used in this paper, including their sizes and language support. *English-centric* indicates that while the models were primarily pre-trained and instruction-tuned in English, they may also include data from other languages[3].

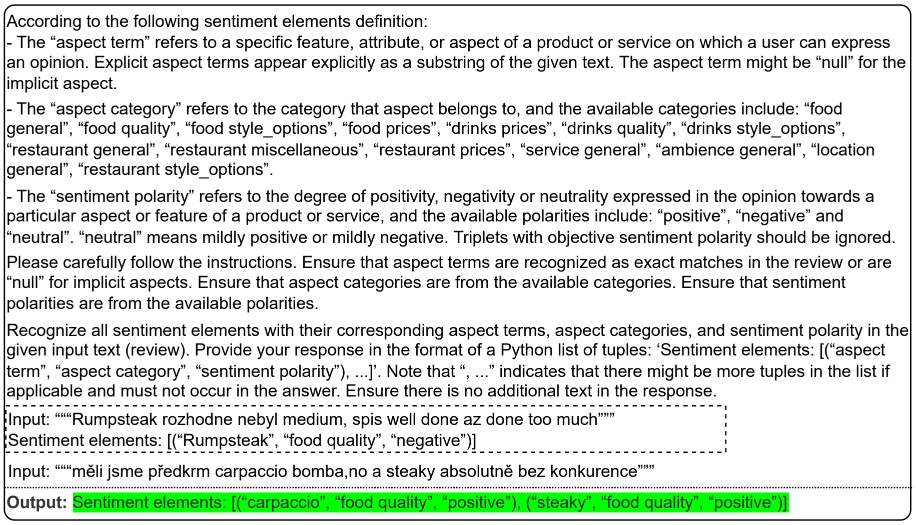
3.2 Prompting Strategy & Fine-Tuning

We design our prompts based on prior work [18, 19], ensuring they are simple, clear, and standardized for ABSA. These prompts define sentiment elements and output format. Sentiment elements specify the permitted label space, such as aspect categories and sentiment polarities or that aspect terms must be found in the text or be *"null"* for implicit ones, while the output format ensures consistency in model responses. We use the standard zero-shot prompt, as those have been shown to often outperform more complex strategies like chain-of-thought for E2E-ABSA in different languages [28].

[3] For example, approximately 90% of LLaMA 2's pre-training data is English [23], with the remainder in other languages.

Table 3: Alphabetically sorted LLMs used in our experiments, their sizes (in billions of parameters), and language support. † indicates models with official support for Czech. * indicates models without official documentation on language support, assumed to be primarily English-centric.

Model	**Sizes (B)**	**Language Support**	**Open-source**
Aya 23 [1]	8, 35	Multilingual†	Yes
Gemma 3 [22]	1, 4, 12, 27	1B: English-centric, others: Multilingual†	Yes
GPT-3.5 Turbo [12]	–	Multilingual†	No
GPT-4o mini [11]	–	Multilingual†	No
LLaMA 2 [23]	7, 13	English-centric	Yes
LLaMA 3 [4]	8	English-centric	Yes
LLaMA 3.1 [4]	8, 70	Multilingual	Yes
LLaMA 3.2 [4]	1, 3	Multilingual	Yes
LLaMA 3.3 [4]	70	Multilingual	Yes
Mistral (v0.3) [7]	7	English-centric*	Yes
Orca 2 [10]	7, 13	English-centric*	Yes



According to the following sentiment elements definition:
- 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: "food general", "food quality", "food style_options", "food prices", "drinks prices", "drinks quality", "drinks style_options", "restaurant general", "restaurant miscellaneous", "restaurant prices", "service general", "ambience general", "location general", "restaurant style_options".

- 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. Triplets with objective sentiment polarity should be ignored.

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.

Recognize all sentiment elements with their corresponding aspect terms, aspect categories, and sentiment polarity 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"), ...]'. 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.

Input: """Rumpsteak rozhodne nebyl medium, spis well done az done too much"""
Sentiment elements: [("Rumpsteak", "food quality", "negative")]

Input: """měli jsme předkrm carpaccio bomba,no a steaky absolutně bez konkurence"""

Output: Sentiment elements: [("carpaccio", "food quality", "positive"), ("steaky", "food quality", "positive")]

Fig. 1: Prompt for the TASD task, showing an example input (English translation: *"we had carpaccio as a starter – amazing – and the steaks were absolutely unmatched"*), the expected output in the green box, and one demonstration in the dashed box. Demonstrations are included only in few-shot scenarios.

Figure 1 shows a TASD prompt, which we adapt for other tasks by omitting irrelevant elements (e.g. sentiment polarity for ACTE). For few-shot experiments, we use the first ten training examples due to their balanced label distribution.

We also test Czech-translated prompts, as prior work shows language alignment helps, especially with English-centric LLMs [8]. Instead of translating the dataset into English – which risks misalignment and errors – we translate the prompt to Czech to preserve evaluation quality.

For fine-tuning, we use QLoRA [3] on models up to 13B parameters, which adds LoRA [6] weights to a 4-bit quantized backbone, reducing memory use while maintaining performance. Since prompt language has no effect during fine-tuning, we use English-only prompts and the task-specific training set, fine-tuning the model to generate outputs in the desired format.

3.3 Experimental Details

We use the official API[4] for GPT models but exclude GPT-3.5 Turbo with Czech prompts due to budget limits. For open-source LLMs, we use instruction-tuned models from HuggingFace Transformers [27]. We use 4-bit quantized models, which offer performance similar to 8-bit or full-precision versions [3].

Fine-tuning follows QLoRA [3], with 4-bit NF4 quantization, bf16 precision, AdamW [9], a learning rate of 2e-4, batch size 16, and LoRA adapters on all linear layers. While $r = 64$, $\alpha = 16$ works for most models, Gemma 3 (4B/12B) required tuning. A grid search found $r = 64$, $\alpha = 128$ performed best. Models are trained for up to 5 epochs, selecting the best by validation loss. Following prior work [10, 18, 19], we compute loss only on generated tokens. All experiments use greedy decoding and run on an NVIDIA L40 GPU with 48 GB of VRAM.

3.4 Evaluation Metrics & Compared Methods

We use micro F1-score, a standard metric in ABSA research, and consider a predicted sentiment tuple correct only if all its elements match the gold tuple exactly. For fine-tuning experiments, we report the average results over five runs.

We compare the performance of LLMs against the best results reported in [20], who fine-tuned multilingual and Czech-only Transformer-based models. For the ACSA task, there are no prior results on the employed dataset.

4 Results

Table 4 presents the zero-shot and few-shot results with Czech and English prompts on four ABSA tasks with different LLMs compared to fine-tuned models. There are several observations:

1) **Effect of Prompt Language**: The impact of using Czech versus English prompts is inconsistent. While Czech prompts sometimes yield slightly better results, English prompts generally perform better. In some cases, the differences are significant; for instance, in the zero-shot ACSA task, LLaMA 3 8B performs about 50% better with an English prompt than a Czech one. However, such large margins are uncommon.

2) **Impact of Model Size, Recency, and Multilingualism**: As expected, larger, newer, and multilingual models tend to achieve better results. Older models such as Orca 2 and LLaMA 2 significantly underperform compared to more

[4] https://platform.openai.com/docs/overview

Table 4: Zero- and few-shot results on different tasks with English (En) and Czech (Cs) prompts with different LLMs compared to the best results with fine-tuned models achieved in [20]. For each column, the best result is in **bold**, the second best is underlined. We group the LLMs by architecture and sort by size.

	ACSA				ACTE				E2E-ABSA				TASD			
	Zero-shot		Few-shot		Zero-shot		Few-shot		Zero-shot		Few-shot		Zero-shot		Few-shot	
	En	Cs	En	Cs	En	Cs	En	Cs	En	Cs	En	Cs	En	Cs	En	Cs
[20]	–				67.30				74.80				59.30			
GPT-3.5 Turbo	57.29	–	61.64	–	26.32	–	45.79	–	44.58	–	54.75	–	25.39	–	42.60	–
GPT-4o mini	61.43	61.65	69.90	70.94	34.22	21.30	51.75	49.32	**54.38**	46.45	60.72	59.51	35.53	24.18	46.07	46.21
Aya 23 8B	41.86	43.26	61.81	62.53	17.70	9.38	39.60	38.33	26.16	16.50	47.66	44.50	13.74	6.99	35.62	35.67
Aya 23 35B	61.67	61.75	67.00	67.27	28.43	26.94	52.88	53.15	43.94	28.90	59.79	55.80	25.98	25.71	46.34	48.37
Gemma 3 1B	5.52	2.64	38.20	32.01	4.99	0.55	19.12	13.91	4.39	0.08	23.87	15.95	5.39	0.72	14.96	12.55
Gemma 3 4B	57.82	59.66	63.13	65.12	39.34	18.26	49.21	48.86	42.43	32.18	54.35	52.46	32.68	13.97	47.72	44.56
Gemma 3 12B	69.25	69.93	69.97	69.27	49.24	41.24	56.65	56.79	53.98	45.85	59.81	59.81	44.61	37.10	51.66	52.47
Gemma 3 27B	**69.79**	**70.91**	**72.76**	**72.74**	**51.47**	**47.18**	**58.60**	**58.26**	51.89	**47.44**	**64.23**	**63.65**	**46.68**	**41.89**	**54.53**	**54.64**
LLaMA 2 7B	14.15	3.82	40.96	40.47	5.97	0.46	29.08	32.61	12.24	0.74	35.04	37.94	3.58	0.94	25.94	27.06
LLaMA 2 13B	32.73	27.78	49.03	52.17	9.57	4.94	37.66	35.86	18.95	13.21	44.03	44.03	10.21	6.43	35.73	35.72
LLaMA 3 8B	53.32	3.01	58.97	47.28	16.74	2.80	39.45	31.64	34.54	11.79	42.32	39.18	7.91	8.31	34.64	28.86
LLaMA 3.1 8B	29.72	26.95	48.28	1.92	8.90	12.24	27.36	7.62	12.30	23.19	41.65	6.22	11.51	1.97	22.31	0.12
LLaMA 3.1 70B	55.15	54.53	68.58	67.47	27.04	24.99	50.62	51.13	44.37	37.84	59.38	57.35	26.08	23.33	47.79	45.47
LLaMA 3.2 1B	0.12	2.76	0.12	1.09	0.00	0.11	0.85	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00
LLaMA 3.2 3B	0.00	6.96	0.00	2.55	0.89	0.47	2.40	2.02	0.12	3.91	9.37	1.06	0.00	0.71	3.59	0.00
LLaMA 3.3 70B	55.59	54.41	70.08	68.75	28.35	25.18	52.92	53.54	48.89	42.46	59.20	54.15	27.85	24.20	49.72	47.92
Mistral 7B	43.56	47.32	57.17	56.12	11.63	7.55	41.13	37.83	21.47	17.21	44.52	39.24	11.58	8.76	37.22	32.63
Orca 2 7B	35.73	0.95	54.28	53.29	7.61	1.23	28.43	27.54	16.06	6.19	32.97	31.16	4.58	0.43	26.75	17.68
Orca 2 13B	49.51	45.72	63.39	62.88	13.72	11.49	35.09	34.81	22.67	20.81	41.04	39.99	11.19	11.19	32.63	32.66

recent multilingual models of similar or even smaller sizes. Additionally, despite being multilingual, LLaMA 3.2 models perform extremely poorly, often scoring 0%. Upon closer examination, we found that these models generated Python code instead of task-relevant outputs, suggesting they failed to understand the task. Interestingly, even few-shot prompting does not help these models. Similarly, Gemma 3 1B struggles in zero-shot scenarios but improves substantially when provided with few-shot examples.

3) **Effect of Few-Shot Examples**: Providing few-shot examples generally improves results, particularly for smaller and more English-centric models. These findings suggest that these models struggle to understand the task from a zero-shot prompt alone, but demonstrations help guide them toward the correct interpretation.

4) **Performance of Proprietary Models**: Among proprietary models, GPT-4o mini consistently outperforms GPT-3.5 Turbo, likely due to its newer architecture and improved capabilities.

5) **Strong Performance of Aya and Gemma Models**: Among open-source models, the Aya and Gemma models perform particularly well, likely due to their official support for Czech and recent release. Notably, Gemma 3 27B performs best in most cases, with Gemma 3 12B frequently ranking second. Their strong results are particularly impressive given that they often outperform proprietary GPT models with significantly more parameters. Aya 23 35B is usually about 5%

worse than the Gemma 3 27B model and only slightly worse than Gemma 3 12B in few-shot scenarios. However, the difference in zero-shot settings is larger; for example, Aya 23 35B is about 20% worse for TASD than Gemma 3 27B. The smaller 8B version of Aya 23 is often more than 10% worse than the 35B version, while the 4B version of Gemma 3 is about 10% worse than the 12B version and is often comparable or only slightly worse than the much larger Aya 23 35B.

6) **Task Difficulty Ranking**: The models generally perform best on ACSA, followed by E2E-ABSA and ACTE, with TASD being the most challenging task. This ranking likely reflects differences in label complexity. ACSA is the easiest because it does not require predicting aspect terms, whereas ACTE and E2E-ABSA involve more complex label spaces. TASD is the hardest since it requires predicting three sentiment elements rather than just two.

7) **Comparison to Fine-Tuned Models**: The best-performing LLMs achieve zero-shot results approximately 20% lower than fine-tuned models. With few-shot prompting, this gap shrinks to around 5–10%. While fine-tuned models still offer superior performance, LLMs provide a viable alternative when annotated data is scarce. Their ability to generate results quickly without the need for fine-tuning makes them attractive for rapid deployment, though fine-tuned models remain the preferred choice when performance is the primary concern.

Table 5: Results with different fine-tuned LLMs compared to the best results with fine-tuned models achieved in [20], alongside the average score. For each task, the best result is in **bold**, the second best is underlined.

	ACSA	**ACTE**	**E2E**	**TASD**	**AVG**
[20]	–	67.30	74.80	59.30	–
Aya 23 8B	76.62	73.02	74.04	68.08	72.94
Gemma 3 1B	68.09	63.52	64.74	53.68	62.50
Gemma 3 4B	73.00	70.57	73.02	65.27	70.46
Gemma 3 12B	76.78	74.30	**75.10**	**69.36**	73.89
LLaMA 2 7B	73.31	66.13	66.53	60.20	66.54
LLaMA 2 13B	73.17	67.01	69.39	60.75	67.58
LLaMA 3 8B	70.77	63.07	62.97	56.84	63.41
LLaMA 3.1 8B	**77.51**	**75.46**	**75.10**	69.06	**74.28**
LLaMA 3.2 1B	65.26	64.16	63.35	55.71	62.12
LLaMA 3.2 3B	73.75	69.14	68.54	61.07	68.13
Mistral 7B	61.13	55.14	54.41	48.52	54.80
Orca 2 7B	74.26	69.99	70.75	63.36	69.59
Orca 2 13B	75.37	72.61	71.83	65.62	71.36

Table 5 presents the results with fine-tuned models, showing significant improvements over previous state-of-the-art approaches. The largest gain is observed in the TASD task, where our best-performing model surpasses prior results by approximately 10%. The top-performing models are LLaMA 3.1 8B, Gemma 3 12B, and Aya 23 8B, demonstrating the effectiveness of fine-tuning for enhancing LLM-based sentiment analysis. Notably, fine-tuning yields greater improvements for English-centric models than multilingual ones, suggesting that language-specific adaptations play a crucial role. Mistral 7B achieves the lowest scores, possibly due to suboptimal training hyperparameters rather than inherent model limitations. The results with 1B and 3B models improve substantially over the zero-shot and few-shot performance, even by 70% in some cases. These results confirm that fine-tuned LLMs are strong alternatives to traditional models for ABSA tasks not only in English, but also in Czech.

4.1 Effect of Few-Shot Example Count

We analyze how the number of few-shot examples impacts performance for selected models. Figure 2 presents the results, averaged across tasks, as their behaviour is generally consistent. Even a single few-shot example provides a noticeable improvement over zero-shot performance. Generally, increasing the number of examples leads to better results, though gains tend to plateau around 5 to 10 examples. Notably, LLaMA 3.1 8B exhibits a performance drop beyond 10 examples, primarily due to declines in ACSA and ACTE tasks. Given these trends, our choice of 10 few-shot examples appears to be a reasonable balance between performance gains and diminishing returns.

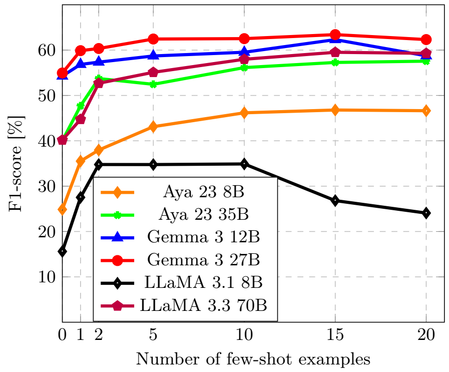
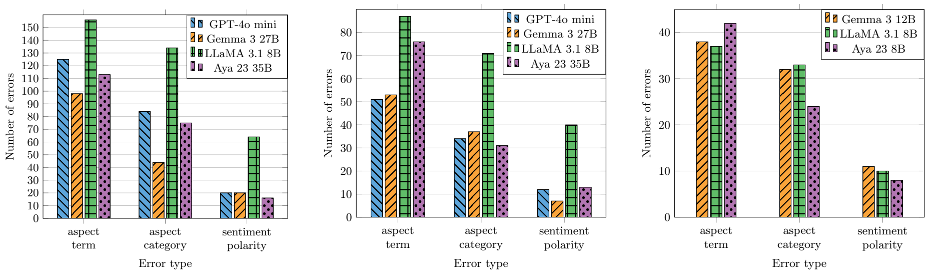


Fig. 2: Impact of the number of few-shot examples on model performance. Results are averaged across all four tasks.

4.2 Error Analysis

We conduct an error analysis to evaluate model performance and identify key challenges. For this purpose, we randomly select 100 test examples and assess multiple models using the same set. Our analysis focuses on the TASD task in zero-shot, few-shot, and fine-tuning scenarios with an English prompt, manually comparing model predictions to ground truth labels. Figure 3 presents the results.



(a) Zero-shot (b) Few-shot (c) Fine-tuning

Fig. 3: Error type distribution for different models on 100 TASD task examples.

Aspect term prediction poses the greatest challenge, as aspect terms can be any word or phrase in the text. Common errors include missing aspect terms, incorrect spans, and partial matches (e.g. omitting or adding words). Implicit aspect terms are particularly problematic – models frequently fail to recognize them or incorrectly predict explicit terms from the text instead. Notably, Aya 23 35B in zero-shot scenarios frequently predicts implicit aspect terms, though often incorrectly, whereas other models rarely identify implicit aspects at all. Additionally, in some cases, models predict aspect terms in their base (nominative) form, even when they appear in a different grammatical case in the text. For example, a model may predict *"obsluha"* (*"service"*) instead of the instrumental form *"obsluhou"*. While technically a mismatch, such predictions are not necessarily incorrect. We recommend developing improved evaluation metrics tailored to LLMs, as the strict matching criteria commonly used in ABSA can be overly harsh in these situations and may unfairly penalize otherwise valid predictions.

Aspect category prediction is relatively easier due to the limited label space. However, models struggle with semantically similar categories, such as *"restaurant miscellaneous"* and *"restaurant general"*, and with rare categories like *"location general"*. LLaMA 3.1 8B exhibits notably higher error rates in aspect category prediction in zero-shot and few-shot settings compared to other models.

Sentiment polarity prediction is the easiest task, with most errors occurring in the *"neutral"* class. Models often misclassify mildly positive or mildly negative sentiment, which implies *"neutral"* polarity, as *"positive"* or *"negative"*, respectively. These errors are significantly less frequent than those related to aspect terms or categories, likely because traditional sentiment analysis is well-represented in pre-training and instruction tuning data for LLMs.

Our analysis reveals that few-shot prompting reduces errors across all sentiment elements, with the greatest impact on aspect term prediction. Sentiment polarity, already the least error-prone element, benefits the least from it.

Fine-tuned models produce the fewest errors, particularly in aspect term prediction. Interestingly, all evaluated models in all scenarios incorrectly predicted the sentiment polarity for the phrase *"Fajn bar"* (*"Cool bar"*) as negative, while the correct sentiment polarity is *"positive"*. The term *"Fajn"* is from Common Czech, suggesting that the models struggle with these types of vernacular expressions.

5 Conclusion

This paper comprehensively evaluates large language models for Czech aspect-based sentiment analysis. We compare 19 LLMs of varying sizes and architectures, assessing their performance across zero-shot, few-shot, and fine-tuning scenarios. Our results highlight the strong influence of model properties – such as multilingualism, size, and recency – on ABSA performance. We find that small models fine-tuned specifically for ABSA outperform LLMs in zero-shot and few-shot settings, while fine-tuned LLMs achieve state-of-the-art results. Additionally, our error analysis identifies key challenges in Czech ABSA, offering insights into the strengths and limitations of LLMs for this task.

Acknowledgements

This work has been supported by the Grant No. SGS-2025-022 – New Data Processing Methods in Current Areas of Computer Science. 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.

References

1. Aryabumi, V., Dang, J., et al.: Aya 23: Open weight releases to further multilingual progress (2024), https://arxiv.org/abs/2405.15032
2. Bai, Y., Han, Z., Zhao, Y., Gao, H., Zhang, Z., Wang, X., Hu, M.: Is compound aspect-based sentiment analysis addressed by LLMs? In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 7836–7861. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.460, https://aclanthology.org/2024.findings-emnlp.460/
3. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: efficient finetuning of quantized llms. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
4. Dubey, A., Jauhri, A., Pandey, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
5. Hercig, T., Brychcín, T., Svoboda, L., Konkol, M., Steinberger, J.: Unsupervised methods to improve aspect-based sentiment analysis in czech. Computación y Sistemas **20**(3), 365–375 (2016)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
7. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825

8. Liu, C., Zhang, W., Zhao, Y., Luu, A.T., Bing, L.: Is translation all you need? a study on solving multilingual tasks with large language models (2024), https://arxiv.org/abs/2403.10258
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
10. Mitra, A., Corro, L.D., Mahajan, S., Codas, A., Simoes, C., Agarwal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., Palangi, H., Zheng, G., Rosset, C., Khanpour, H., Awadallah, A.: Orca 2: Teaching small language models how to reason (2023), https://arxiv.org/abs/2311.11045
11. OpenAI, :, Hurst, A., Lerer, A., et al.: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
12. OpenAI: GPT-3.5 Turbo (2024), https://platform.openai.com/docs/models/gpt-3.5-turbo, accessed March 2025
13. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: Aspect based sentiment analysis. In: Nakov, P., Zesch, T., Cer, D., Jurgens, D. (eds.) Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 486–495. Association for Computational Linguistics, Denver, Colorado (Jun 2015). https://doi.org/10.18653/v1/S15-2082, https://aclanthology.org/S15-2082/
14. Přibáň, P., Prazak, O.: Improving aspect-based sentiment with end-to-end semantic role labeling model. In: Mitkov, R., Angelova, G. (eds.) Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. pp. 888–897. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria (Sep 2023), https://aclanthology.org/2023.ranlp-1.96/
15. Schmitt, M., Steinheber, S., Schreiber, K., Roth, B.: Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1109–1114. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1139, https://aclanthology.org/D18-1139/
16. Šmíd, J., Král, P.: Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges. Information Fusion **120**, 103073 (2025). https://doi.org/https://doi.org/10.1016/j.inffus.2025.103073, https://www.sciencedirect.com/science/article/pii/S1566253525001460
17. Šmíd, J., Přibáň, P.: Prompt-based approach for Czech sentiment analysis. In: Mitkov, R., Angelova, G. (eds.) Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. pp. 1110–1120. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria (Sep 2023), https://aclanthology.org/2023.ranlp-1.118/
18. Šmíd, J., Přibáň, P., Kral, P.: LLaMA-based models for aspect-based sentiment analysis. In: De Clercq, O., Barriere, V., Barnes, J., Klinger, R., Sedoc, J., Tafreshi, S. (eds.) Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. pp. 63–70. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.wassa-1.6, https://aclanthology.org/2024.wassa-1.6/
19. Šmíd, J., Přibáň, P., Král, P.: Advancing cross-lingual aspect-based sentiment analysis with llms and constrained decoding for sequence-to-sequence models.

In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART. pp. 757–766. INSTICC, SciTePress (2025). https://doi.org/10.5220/0013349400003890

20. Šmíd, J., Přibáň, P., Prazak, O., Kral, P.: Czech dataset for complex aspect-based sentiment analysis tasks. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 4299–4310. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.384/
21. Steinberger, J., Brychcín, T., Konkol, M.: Aspect-level sentiment analysis in Czech. In: Balahur, A., van der Goot, E., Steinberger, R., Montoyo, A. (eds.) Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 24–30. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/W14-2605, https://aclanthology.org/W14-2605/
22. Team, G., Kamath, A., et al.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786
23. Touvron, H., Martin, L., et al.: Llama 2: Open foundation and fine-tuned chat models (2023), https://arxiv.org/abs/2307.09288
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
25. Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., Pan, J.Z.: Target-aspect-sentiment joint detection for aspect-based sentiment analysis. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 9122–9129 (Apr 2020). https://doi.org/10.1609/aaai.v34i05.6447, https://ojs.aaai.org/index.php/AAAI/article/view/6447
26. Wang, F., Lan, M., Wang, W.: Towards a one-stop solution to both aspect extraction and sentiment analysis tasks with neural multi-task learning. In: 2018 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2018)
27. Wolf, T., Debut, L., Sanh, V., Chaumond, J., et al.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6/
28. Wu, C., Ma, B., Zhang, Z., Deng, N., He, Y., Xue, Y.: Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models (2024), https://arxiv.org/abs/2412.12564
29. Zhang, W., Deng, Y., Liu, B., Pan, S., Bing, L.: Sentiment analysis in the era of large language models: A reality check. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3881–3906. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.findings-naacl.246, https://aclanthology.org/2024.findings-naacl.246/