LLM-as-an-Annotator: Training Lightweight Models with LLM-Annotated Examples for Aspect Sentiment Tuple Prediction

Nils Constantin Hellwig[1], **Jakob Fehle**[1], **Udo Kruschwitz**[2], **Christian Wolff**[1]
[1]Media Informatics Group, University of Regensburg, Regensburg, Germany
[2]Information Science Group, University of Regensburg, Regensburg, Germany
nils-constantin.hellwig@ur.de, jakob.fehle@ur.de, udo.kruschwitz@ur.de, christian.wolff@ur.de

Abstract

Training models for Aspect-Based Sentiment Analysis (ABSA) tasks requires manually annotated data, which is expensive and time-consuming to obtain. This paper introduces LA-ABSA, a novel approach that leverages Large Language Model (LLM)-generated annotations to fine-tune lightweight models for complex ABSA tasks. We evaluate our approach on five datasets for Target Aspect Sentiment Detection (TASD) and Aspect Sentiment Quad Prediction (ASQP). Our approach outperformed previously reported augmentation strategies and achieved competitive performance with LLM-prompting in low-resource scenarios, while providing substantial energy efficiency benefits. For example, using 50 annotated examples for in-context learning (ICL) to guide the annotation of unlabeled data, LA-ABSA achieved an F1 score of 49.85 for ASQP on the SemEval Rest16 dataset, closely matching the performance of ICL prompting with Gemma-3-27B (51.10), while requiring significantly lower computational resources.

Keywords: Aspect-Based Sentiment Analysis, Large Language Models, Data Annotation

1. Introduction

Sentiment Analysis (SA) involves recognizing the emotional mood expressed in a text, for example in domains like e-commerce or social media (Liu, 2022, p. 7). Traditional sentiment classification assigns a single polarity to a text, capturing its overall emotional tone (Liu, 2022, p. 59). In contrast, recent research considered more granular tasks, such as Target Aspect Sentiment Detection (TASD) and Aspect Sentiment Quad Prediction (ASQP), which fall under Aspect-based Sentiment Analysis (ABSA) (Zhang et al., 2021; Negi et al., 2024; Li et al., 2024; Liu et al., 2024). These tuple prediction tasks enable a detailed extraction of opinion structures, enhancing the analysis of nuanced sentiments in diverse linguistic contexts. TASD involves identifying (aspect term *a*, aspect category *c*, sentiment polarity *p*)-triplets, whereas ASQP allows for an even more precise insight by additionally predicting the opinion term *o*, resulting in quadruples (Zhang et al., 2021). For example, given the sentence *"The pizza was tasty."*, the corresponding ASQP annotation would be (*a*: *pizza*, *c*: *food general*, *o*: *tasty*, *p*: *positive*).

However, the increased complexity of these tasks comes at the cost of data annotation: creating labelled datasets for TASD and ASQP is even more labour-intensive than for traditional sentiment tasks, requiring annotators to identify multiple interrelated text spans, the related aspect category and sentiment polarities (Nasution and Onan, 2024; Negi et al., 2024; Wang et al., 2023a). As a result, annotated resources for these tasks are often limited in scale and domain coverage.

Large Language Models (LLMs) such as Gemma-3 (Team et al., 2025), LLaMA-3.1 (Grattafiori et al., 2024) and GPT-4 (Achiam et al., 2023) recently demonstrated strong performance in both zero-shot and few-shot settings, showing the ability to generalize well from task instructions and limited examples (Bai et al., 2024; Gou et al., 2023; Hellwig et al., 2025a).

Despite their considerable performance, LLMs are resource-intensive, requiring substantial computational power and memory (Xu et al., 2025). This poses challenges for practical deployment, especially at scale or in scenarios involving privacy-sensitive data, particularly when relying on external APIs offering access to an LLM (Chen et al., 2025; Li and Sun, 2025).

While zero-shot and few-shot methods are effective in scenarios without labelled data, they do not exploit the potentially available unlabelled textual resources. An alternative approach, which has been employed across various NLP tasks, involves using LLMs to annotate unlabelled data and subsequently leveraging these annotations to train smaller, more computationally efficient models (Huang et al., 2023; Malik et al., 2024; Yang et al., 2024; Zhang et al., 2023a).

In this study, we evaluate whether LLMs can be leveraged as data annotators for aspect sentiment tuple prediction, specifically, ASQP and TASD. We introduce LLM-as-an-Annotator (LA-ABSA), a novel approach that labels unlabelled examples using Gemma-3-27B and subsequently trains lightweight language models[1] on the LLM-annotated examples. We systematically evaluate

[1]Language models with fewer than 1 billion parameters were considered as lightweight.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7955–7972
11–16 May 2026.

this approach by addressing three key research questions:

- **RQ1:** Do lightweight language models trained on LLM-annotated examples perform as well as those trained on expert-annotated data?
- **RQ2:** Do lightweight language models trained on LLM-annotated examples perform better than zero-shot and few-shot prompted LLMs?
- **RQ3:** Does LA-ABSA outperform data augmentation methods?

Our main contributions are as follows:

- We introduce LA-ABSA, enabling lightweight language model training for aspect sentiment tuple prediction using LLM-generated annotations, while reducing dependency on expensive expert labelling.
- Evaluation across five datasets shows LA-ABSA achieves comparable performance to LLM prompting with lower energy consumption for large-scale deployments.
- We provide publicly available code (`https://github.com/NilsHellwig/LA-ABSA`) adaptable to other domains and languages.

2. Related Work

This section presents related work on two main aspects: (1) state-of-the-art (SOTA) methodologies for Target Aspect Sentiment Detection (TASD) and Aspect Sentiment Quad Prediction (ASQP), and (2) approaches aimed at mitigating annotation effort in Aspect-based Sentiment Analysis (ABSA).

Notably, all subsequently mentioned approaches for tackling tuple prediction tasks are primarily evaluated on the SemEval 2015 and 2016 datasets, which include TASD annotations (Pontiki et al., 2015, 2016). For the ASQP task, the extended versions of these datasets introduced by Zhang et al. (2021) were employed, which additionally include opinion term annotations.

2.1. Fine-tuning Language Models for Tuple Prediction

Fine-Tuning Approaches. For both ASQP and TASD, generative approaches achieved SOTA performance when small language models fine-tuned on annotated examples are employed. Most notably, Google's lightweight `T5-base` (223 million parameters) (Raffel et al., 2020) text-to-text model. The approaches primarily differ in how tuples are represented. When using the `T5-base` model, Multi-view Prompting (MvP), introduced by Gou et al. (2023), currently sets the SOTA performance for both ASQP and TASD on the Rest15 and Rest16 datasets, achieving, an F1 score of 72.76 on TASD-Rest16 and 60.39 on ASQP-Rest16.

Close contenders include Dataset-level Order (DLO) (Hu et al., 2022b) (TASD-Rest16: 71.79; ASQP-Rest16: 59.79) and Paraphrase (Zhang et al., 2021) (TASD-Rest16: 71.97; ASQP-Rest16: 57.93). MvP is notably more computationally intensive than other methods, as it introduces element order-based prompts to guide tuple generation from multiple perspectives, meaning that different positional configurations of sentiment elements within a tuple are considered. This results in a multiplication of training samples by the number of permutations. Five permutations are considered for ASQP and TASD (Gou et al., 2023). In contrast, DLO selects only the top-3 permutations, while Paraphrase relies on a single fixed ordering (Zhang et al., 2021).

Low-Rank Adaptation (LoRA) Fine-Tuning. Fine-tuning LLMs with billions of parameters with Low-Rank Adaptation (LoRA) (Hu et al., 2022a), a technique that updates only a small subset of parameters to reduce computational costs, has led to further improvements in F1 scores on both TASD and ASQP tasks. For example, Šmíd et al. (2024) report F1 scores of 78.82 and 76.10 on the TASD-Rest16 dataset using Microsoft's Orca-2-8B (Mitra et al., 2023) in its 13B and 7B variants, respectively.

2.2. Approaches for Minimizing Annotation Effort

To address the challenge of limited labelled resources, three approaches have received particular attention in recent NLP research: (1) LLM prompting (Brown et al., 2020; Zhang et al., 2024a; Wang et al., 2025; Goel et al., 2024), (2) traditional data augmentation methods such as EDA and back-translation (Imran et al., 2023; Sabty et al., 2021; Wei and Zou, 2019), and (3) generative, LLM-based approaches that either augment existing annotated examples or generate new ones (Ding et al., 2024; Chung et al., 2023; Omura et al., 2024; Schmidt et al., 2024). This section examines their application and adaptation within ABSA.

Zero-Shot and Few-Shot Learning. To reduce annotation costs, recent studies explored zero-shot and few-shot learning for tuple prediction tasks (Bai et al., 2024; Simmering and Huoviala, 2023; Šmíd et al., 2024; Zhang et al., 2024b; Zhou et al., 2024). Hellwig et al. (2025a) achieved SOTA in-context learning (ICL) performance on the ASQP and TASD tasks, reaching an F1 score of 66.03 on TASD-Rest16, though below SOTA fine-tuned performance slightly exceeding 70. However, their few-shot configurations outperformed established fine-tuning on TASD and ASQP tasks when only a few examples are given for training (Hellwig et al., 2025a; Varia et al., 2023).

Traditional Data Augmentation Methods. Data augmentation for ABSA has been primarily evaluated in fine-tuning settings with full annotated datasets (Li et al., 2023; Liesting et al., 2021; Wang et al., 2023b). Liesting et al. (2021) applied Easy Data Augmentation (EDA) to aspect sentiment classification (aspect term + sentiment polarity), achieving modest gains of 0.5-1.0 points on SemEval datasets. EDA uses lexical transformations: synonym replacement, random insertion, swapping, and deletion (Wei and Zou, 2019). Back-translation showed negligible improvements (Liesting et al., 2021).

Generative Augmentation Strategies. Leveraging the generative capabilities of language models, Wang et al. (2023a) propose a data augmentation approach for ASQP that augments new quads by exchanging opinion terms and sentiment polarity between tuples sharing the same aspect category, followed by training a quads-to-text (Q2T) model to generate diverse augmented texts based on the augmented quads. This approach yielded performance improvements of approximately 2 percentage points over the paraphrase-based method on both Rest15-ASQP and Rest16-ASQP datasets.[2]

Similarly, yang Lu et al. (2025) introduced an LLM-based approach named Quantity Augmentation and Information Enhancement (QAIE). QAIE uses PaLM (540 billion parameters) (Chowdhery et al., 2023) and assumes a small annotated set of examples with balanced distribution of *k* examples per category. This approach demonstrated substantial performance improvements, achieving an F1 score increase from 22.97 to 35.31 (*k=20*, 86 examples) on the ASQP-Rest15 dataset. QAIE enhances a given dataset by generating multiple paraphrased review texts with varied linguistic expressions and enriching them with additional implicit information before training a T5 model on the target task. The augmentation process generates more examples when both aspect and opinion terms are present in the source text, as this configuration enables the combination of synonym replacements and antonym substitutions for both opinion and aspect terms.

Finally, Xu et al. (2025) introduced DS^2-ABSA, a two-pronged data augmentation approach for E2E-ABSA (aspect term + sentiment polarity) that: (1) augments existing examples through sentence masking followed by paraphrasing of the masked instances, and (2) generates new training examples by first producing aspect and opinion terms, then synthesizing 20,000 complete examples based on these terms. Their approach outperformed other generative approaches adapted for E2E-ABSA.

[2] This approach was not adapted for our study, as we were unable to run the code provided on GitHub in its current form. Moreover, as reported in a GitHub issue, other users faced the same problem, and despite multiple attempts, we were unable to establish contact with the authors.

3. Methodology

3.1. LLM-as-an-Annotator (LA-ABSA)

We introduce LA-ABSA, a novel approach for aspect sentiment tuple prediction (e.g., TASD or ASQP) that is dedicated to scenarios with limited annotated data. As illustrated in Figure 1, our method leverages LLMs to automatically annotate unlabeled examples, thereby reducing the reliance on costly human annotation efforts.

3.1.1. Annotator Module

Algorithm 1: Annotator Module

Input: Unlabeled input set $\mathcal{D} = \{x_i\}_{i=1}^{N}$, language model $\mathcal{M}$, few-shot examples $\mathcal{E} = \{(x_j, y_j)\}_{j=1}^{k}$
Output: Annotated dataset $\mathcal{D}_{\text{annotated}} = \{(x_i, y_i)\}_{i=1}^{N}$

foreach $x_i \in \mathcal{D}$ **do**
  $p_i :=$ `construct_prompt`$(\mathcal{E}, x_i)$
  for $r = 1$ **to** *10* **do**
    $y_i \leftarrow \mathcal{M}(p_i)$
    if `validate_label`(y_i) **then**
      break
    end
  end
  if not `validate_label`(y_i) **then**
    $y_i \leftarrow []$
  end
  $\mathcal{D}_{\text{annotated}} \leftarrow \mathcal{D}_{\text{annotated}} \cup \{(x_i, y_i)\}$
end
return $\mathcal{D}_{\text{annotated}}$

To annotate the unlabeled training examples, we used the prompt template introduced by Gou et al. (2023). The prompt comprises an explanation on all sentiment elements and the output format, few-shot examples (except for zero-shot conditions), and the text to be labelled. However, we adopted the modification from Hellwig et al. (2025a), who additionally instructed the LLM to explicitly extract the exact aspect and opinion phrases from the input text. From the training split of the respective dataset and task, either 0, 10, or 50 few-shot examples were randomly sampled and inserted into the prompt, similar to Hellwig et al. (2025a).

Following prior work (Gou et al., 2023; Hellwig et al., 2025a; Hu et al., 2022b), we generated multi-

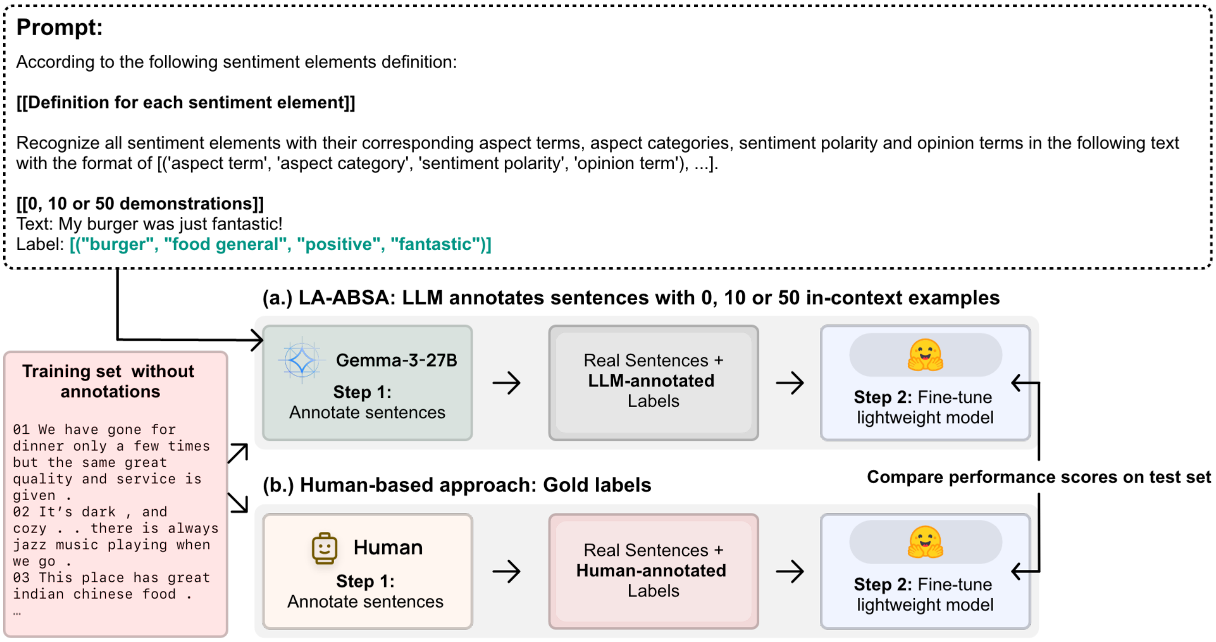

Figure 1: **Illustration of LLM-as-an-Annotator (LA-ABSA).** A LLM (Gemma-3-27B) is prompted to annotate training examples, which are subsequently used to fine-tune lightweight state-of-the-art models for Target Aspect Sentiment Detection (TASD) and Aspect Sentiment Quad Prediction (ASQP).

ple outputs with the LLM using five different random seeds. A predicted tuple was included in the final label set if it appeared in the majority (i.e., at least 3 out of 5) of the generated outputs. Formally:

$$T_{\text{final}} = \left\{ t \,\middle|\, \sum_{i=1}^{m} \mathbf{1}\left[t \in T_i\right] > \frac{m}{2} \right\} \tag{1}$$

where T_i is the set of predicted tuples from the i-th run, $m = 5$ is the number of generations, and $\mathbf{1}[\cdot]$ is the indicator function that returns 1 if the condition is true and 0 otherwise.

Following Hellwig et al. (2025a), we re-executed the LLM in cases where it predicted an invalid aspect category or sentiment polarity, or generated phrases that were not present in the input sentence (see Algorithm 1). The label was set to an empty list (`[]`) if no valid tuples were predicted after 10 re-generations. To allow for sufficient variation across generations, the decoding temperature was set to 0.8, similar to Hellwig et al. (2025a).

Similar to Hellwig et al. (2025a), we employed Gemma-3-27B.[3] Gemma-3-27B is the most recent open-source LLM by *Google*, which comprises 27.4 billion parameters (Team et al., 2025).

This and all subsequently introduced approaches were conducted on a single NVIDIA RTX A5000 GPU with 24GB VRAM, ensuring consistent computational environments across model training and evaluation procedures.

[3] https://ollama.com/library/gemma3:27b

3.1.2. Trainer Module

Subsequently, the LLM-annotated examples were used as training data, combined with the given 0, 10, or 50 gold-standard examples. We evaluated two lightweight fine-tuning approaches based on `T5-base` (223 million parameters): (1) DLO (Hu et al., 2022b) and (2) Paraphrase (Zhang et al., 2021). For both DLO and Paraphrase, we adopted the same hyperparameters as proposed by Hu et al. (2022b) and Zhang et al. (2021) respectively: 20 epochs, a learning rate of 2e-4 (3e-4 for Paraphrase) and a batch size of 16.

3.2. Baselines

3.2.1. Fine-tuning on Human-Annotated Training Data

We included fine-tuning baselines Paraphrase and DLO on varying amounts of human-annotated data. Specifically, we conduct experiments using 10, 50 or all examples from the training set for each task and dataset combination. We employed the same hyperparameters as specified in Section 3.1.2 to ensure fair comparison.

3.2.2. LLMs for Zero-shot and Few-shot Prompting

Next, we incorporate the zero-shot and few-shot ICL results reported by Hellwig et al. (2025a), which demonstrated SOTA performance in low-resource scenarios for both TASD and ASQP. Following their experimental setup, we employed Gemma-3-27B for 0-, 10-, and 50-shot prompting configurations.

3.2.3. Low-resource Enhancement Methods.

Algorithm 2: EDA

Input: Dataset $\mathcal{D}$ with k examples (s, t), where s is a sentence and t the list of corresponding tuples; Number of augmentations per example α

Output: Augmented dataset $\mathcal{D}_{\text{aug}}$ with α examples per (s, t)

foreach $(s, t) \in \mathcal{D}$ **do**
  $tokens \leftarrow \texttt{tokenizer}(s)$
  for $i \leftarrow 1$ **to** α **do**
    $tokens_{\text{aug}} \leftarrow tokens.\text{copy}()$
    $t_{\text{aug}} \leftarrow t.\text{copy}()$
    $tokens_{\text{aug}} \leftarrow \texttt{insertion}(tokens_{aug})$
    $tokens_{\text{aug}} \leftarrow \texttt{deletion}(tokens_{aug})$
    $tokens_{\text{aug}} \leftarrow \texttt{swap}(tokens_{aug})$
    $tokens_{\text{aug}} \leftarrow \texttt{synonym}(tokens_{aug})$
    $s_{\text{aug}} \leftarrow \texttt{join}(tokens_{aug})$
    $\mathcal{D}_{\text{aug}} \leftarrow \mathcal{D}_{\text{aug}} \cup \{(s_{\text{aug}}, t_{\text{aug}})\}$
  end
end
return $\mathcal{D}_{\text{aug}}$

Easy Data Augmentation (EDA) was evaluated since it's widely used across various NLP tasks and ABSA tasks considering less sentiment elements (Hsu et al., 2021; Liesting et al., 2021; Rahamim et al., 2023). For a given set of annotated examples (10 or 50), we generated 10 augmented examples. For 10 augmentations per example and 10 given examples, we produced 100 augmented examples, resulting in a total of 110 examples when combined with the given 10 real examples.

The augmentation procedure is outlined in Algorithm 2. First, the insertion operation adds a synonym of a randomly selected word from the sentence at a random position, but not within aspect or opinion terms. The deletion operation removes a random token that is not part of a term, swap exchanges two tokens that are not located within any term. Finally, the replacement operation substitutes a random token in the sentence with its synonym. DLO was employed for training.

Quantity Augmentation and Information Enhancement (QAIE) by yang Lu et al. (2025) was included with the following modifications: (1) augmentation was performed based on 10 or 50 given examples, consistent with the other experimental conditions, and (2) instead of *Google*'s deprecated PaLM API[4], we used Gemma-3-27B, which further improves comparability across the evaluated approaches. In cases where an example did not contain any explicit aspect term, term-based augmentation was impossible, resulting in fewer augmentations compared to the ASQP task, where an opinion term is always given for all examples. Notably, yang Lu et al. (2025) exclusively augmented ASQP examples.

DS^2-ABSA was adapted with several modifications. To maintain consistency with LA-ABSA and the aforementioned baseline methods, we employed Gemma-3-27B instead of GPT-3.5-Turbo for data synthesis. Furthermore, we did not implement the iterative noisy self-training mechanism from the original DS^2-ABSA framework, where a teacher model iteratively refines synthetic labels and trains successive student models until validation performance stabilizes. Our scenario assumes strict data constraints: apart from 0, 10, or 50 training examples, no additional annotated data is assumed.

3.3. Evaluation and Datasets

Dataset	Train		Test	
	TASD	ASQP	TASD	ASQP
Rest15	1,120	834	582	537
Rest16	1,708	1,264	587	544
FlightABSA	1,351	1,351	387	387
Coursera	1,400	1,400	400	400
Hotels	1,400	1,400	400	400

Table 1: **Dataset statistics overview.** Distribution of annotated examples for TASD and ASQP tasks across five domain-specific datasets, showing training and test set sizes.

We evaluated the tuple prediction tasks ASQP and TASD on five diverse datasets, which, to the best of our knowledge, constitute the only publicly available resources annotated for both tasks. For the TASD task, we utilized the SemEval 2015 (Pontiki et al., 2015) and 2016 (Pontiki et al., 2016) datasets. For ASQP, we employed the extended versions of the SemEval datasets by Zhang et al. (2021), which additionally include opinion term annotations.

Furthermore, we incorporated an adapted version of the OATS dataset (Chebolu et al., 2024) by Hellwig et al. (2025a), which comprises hotel and e-learning (Coursera) domain data. These adaptations included converting ASQP quadruples to TASD triplets by removing opinion terms and eliminating duplicate triplets from the resulting dataset. Finally, we included FlightABSA, an ABSA-dataset comprising airline reviews that was introduced by Hellwig et al. (2025a). The distribution of aspect categories and polarities across the considered datasets and tasks is illustrated in Appendix A.

[4] *Google* Generative AI: `https://pypi.org/project/google-generativeai/`

Approach	# Train		Rest15		Rest16		FlightABSA		Coursera		Hotels		AVG	
	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP
Scenario: 0 annotated examples given														
Gemma-3-27B (0-shot)	0	0	**30.36**	**24.73**	45.51	**28.96**	51.81	42.37	29.50	13.36	38.97	23.02	39.23	**26.49**
LA-ABSA														
w/ Paraphrase	full	full	27.70	18.54	**47.73**	26.07	**52.10**	42.58	**30.92**	**13.41**	**40.74**	**26.27**	**39.84**	25.37
w/ DLO	full	full	25.82	18.88	47.41	26.97	51.75	**42.82**	30.60	13.38	40.70	24.97	39.26	25.40
Scenario: 10 annotated examples given														
Gemma-3-27B ICL	10	10	**54.47**	**39.95**	**66.75**	**46.23**	60.36	45.24	**41.69**	22.31	**56.51**	31.41	**55.96**	**37.03**
Paraphrase	10	10	8.75	1.32	6.66	3.56	8.82	3.44	15.94	4.75	14.91	2.63	11.02	3.14
DLO	10	10	15.84	4.37	13.59	5.18	16.07	4.87	22.93	4.47	18.07	3.53	17.30	4.48
LA-ABSA														
w/ Paraphrase	full	full	49.09	35.04	62.74	44.59	59.87	45.89	38.46	22.47	55.69	**32.83**	53.17	36.16
w/ DLO	full	full	49.23	37.19	62.37	46.20	**61.40**	**46.47**	39.22	**23.37**	55.27	31.44	53.50	36.93
EDA	110	110	30.18	9.34	19.16	11.80	22.66	11.95	28.83	14.19	25.65	7.88	25.30	11.03
QAIE	26.4	45.2	20.11	9.96	12.37	14.78	17.76	15.38	25.18	17.54	21.69	9.86	19.42	13.51
DS^2-ABSA	21,1k	21,1k	29.04	18.19	32.09	23.52	27.01	16.49	17.81	7.61	31.87	11.78	27.56	15.52
Scenario: 50 annotated examples given														
Gemma-3-27B ICL	50	50	**62.12**	**41.74**	**68.53**	**51.10**	**64.60**	48.37	**44.80**	**25.86**	**62.97**	43.83	**60.60**	**42.18**
Paraphrase	50	50	36.92	25.55	35.87	23.50	33.57	17.98	34.26	19.38	40.10	23.09	36.14	21.90
DLO	50	50	39.54	26.63	43.95	29.57	42.92	28.74	36.04	19.08	44.72	27.20	41.44	26.24
LA-ABSA														
w/ Paraphrase	full	full	56.21	37.61	62.20	46.76	62.47	47.16	44.36	25.71	60.58	43.19	57.16	40.09
w/ DLO	full	full	58.40	40.38	62.03	49.85	62.57	**48.98**	44.39	25.69	61.43	**44.24**	57.76	41.83
EDA	550	550	44.14	30.93	45.72	34.59	45.64	36.73	37.89	24.57	45.72	32.51	43.82	31.86
QAIE	141.8	248.8	45.01	33.87	45.09	35.21	48.44	33.98	36.23	22.45	50.51	35.95	45.06	32.29
DS^2-ABSA	21,7k	21,6k	37.94	28.93	43.53	35.79	33.39	23.87	29.56	16.41	40.26	23.61	36.94	25.72
Full set of human-annotated examples: SOTA approaches														
Paraphrase	full	full	**<u>63.06</u>**	46.93	**<u>71.97</u>**	57.93	**<u>69.74</u>**	57.76	51.86	32.34	67.70	53.87	64.87	49.77
DLO	full	full	62.95	**<u>48.18</u>**	71.79	**<u>59.79</u>**	68.95	**<u>58.33</u>**	**<u>52.58</u>**	**<u>32.54</u>**	**<u>68.56</u>**	**<u>55.45</u>**	**<u>64.97</u>**	**<u>50.86</u>**

Table 2: **F1 scores of LA-ABSA**. Results are evaluated against EDA-based data augmentation methods, QAIE (FT), DS^2-ABSA (FT), and prompting baselines (0, 10, and 50 annotated examples) as reported by Hellwig et al. (2025a), as well as fully supervised models including DLO (Hu et al., 2022b) (FT) and Paraphrase (Zhang et al., 2021) (FT). The highest F1 scores within each annotation regime (0, 10, 50 or all examples) are shown in bold; the best overall scores across all settings are underlined.

As commonly done in ABSA research, the reported evaluation metric is the micro-averaged F1 score (Zhang et al., 2023b).

4. Results

4.1. Overall Results

Table 2 presents the performance scores of LA-ABSA. Detailed precision, recall, and macro-averaged F1 scores are reported in Appendix D.

LA-ABSA achieves competitive or superior performance relative to LLM prompting. Table 2 shows that LA-ABSA achieved F1 scores only about 4 percentage points below those achieved by LLM-prompting (e.g., Rest15-TASD, 50-shot: 58.40 vs. 62.12). However, several notable exceptions emerge: in zero-shot scenarios, LA-ABSA with DLO fine-tuning outperformed zero-shot LLM-prompting across most datasets and tasks. Additionally, occasional superior performance is observed in 10-shot and 50-shot settings for the ASQP task.

LA-ABSA demonstrates higher performance scores than data augmentation approaches. Augmentation approaches that expand labelled examples, including EDA-based methods, QAIE and DS^2-ABSA demonstrated substantially lower performance than LA-ABSA across 0-shot, 10-shot, and 50-shot learning scenarios. Performance gaps reach approximately 10 percentage points in several cases, with differences exceeding 20 percentage points in 10-shot learning configurations.

To investigate the impact of different augmentation ratios (2, 5, 10, and 15 augmentations per annotated example) and fine-tuning approaches (DLO, Paraphrase), we conducted an ablation study for EDA detailed in Appendix B. Our results demonstrate that 10 augmentations in combination with DLO achieve the highest performance.

Expert-annotated training data yielded superior performance over LLM-generated annotations. Fine-tuning SOTA approaches for tuple prediction tasks using expert annotations is still

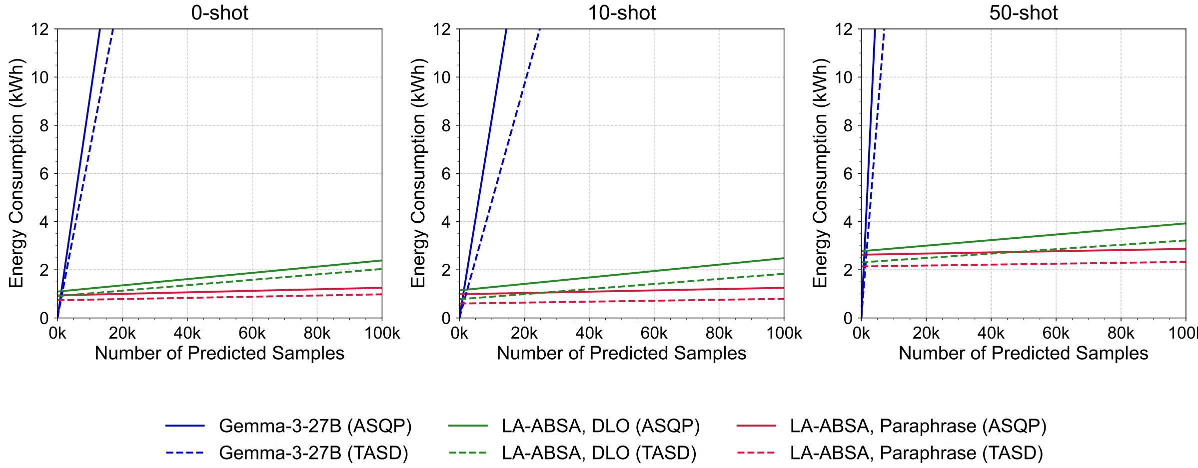

Figure 2: **Energy consumption analysis**. Comparative evaluation of energy usage (kilowatt-hours, kWh) for ASQP and TASD tasks across different settings (0, 10 or 50 annotated examples given). Results are shown for LLM-prompting (Gemma-3-27B) and LA-ABSA methods, using either DLO (Hu et al., 2022b) or Paraphrase (Zhang et al., 2021) for fine-tuning. Each line represents the average energy usage to predict up to 100,000 examples per method and task across the five datasets. LA-ABSA approaches generally require much lower energy due to their smaller underlying base model T5-base.

superior to using LLM-generated annotations. Furthermore, fine-tuning on expert annotations still surpasses all other evaluated approaches in our experimental setup.

Notably, we conducted two-sided paired t-tests with Bonferroni-Holm correction (Holm, 1979) to examine significant differences ($p_{\text{adj}} \leq 0.05$) between LA-ABSA (based on Paraphrase or DLO), fine-tuned approaches trained on the full human-labelled training set (DLO and Paraphrase), and the strongest baseline in low-resource scenarios, ICL with Gemma-3-27B. Considering these five comparison groups, we performed 10 pairwise tests per task (TASD and ASQP) and shot setting (0, 10, and 50 shots), yielding a total of 56 comparisons. Each group comprised five values (one per dataset). No significant differences were found, except that Paraphrase and DLO fine-tuned on the full set achieved significantly higher performance scores than 0-shot prompting for the TASD task.

4.2. Energy Consumption Analysis

A key advantage of LA-ABSA lies in its superior energy efficiency compared to LLM-prompting approaches. To quantify this benefit, we conducted an energy consumption analysis comparing the two best-performing low-resource methods LA-ABSA and LLM-prompting across up to 100,000 generated examples.

Our analysis uses the average energy consumption per prediction as the baseline metric, computed across all five datasets. It is important to note that LA-ABSA's energy consumption profile includes initial overhead costs from two prerequisite steps: (1) LLM-based annotation of training examples, and (2) fine-tuning of the lightweight approach (DLO or Paraphrase). Consequently, the energy curve for LA-ABSA does not originate at zero, reflecting these upfront investments.

As illustrated in Figure 2, LA-ABSA demonstrates significantly lower energy consumption than LLM-prompting when processing fewer than 2,000 examples, regardless of the task type or the number of given examples (0, 10, or 50). This finding indicates that despite the initial energy investment required for model preparation, LA-ABSA becomes substantially more energy-efficient at scale, positioning it as a more sustainable solution for large-scale ABSA deployments.

Table 3 provides detailed energy consumption metrics for LA-ABSA, expressed in milliwatt-hours (mWh) per predicted example. The results highlight the particular efficiency of the paraphrase approach, which evaluates only a single ordering of sentiment elements per test example, compared to the DLO approach that requires assessment of multiple sentiment element permutations.

We note that all energy measurements were conducted using an NVIDIA A5000 GPU (24GB VRAM), and the reported metrics are inherently hardware-dependent. Generalization to other GPU architectures would require appropriate scaling factors. Additionally, we provide a complementary analysis of prediction times, which exhibits trends consistent with our energy consumption findings (see Appendix C).

Method	0-shot		10-shot		50-shot	
	TASD	ASQP	TASD	ASQP	TASD	ASQP
LA-ABSA (Paraphrase)	2.47 ± 0.53	3.17 ± 0.31	1.96 ± 0.40	2.66 ± 0.28	1.83 ± 0.38	2.41 ± 0.46
LA-ABSA (DLO)	11.16 ± 0.72	13.06 ± 0.89	10.68 ± 2.11	13.19 ± 1.62	8.95 ± 0.85	11.65 ± 2.23
Gemma-3-27B ICL	703.10 ± 63.97	912.25 ± 72.67	481.39 ± 46.11	843.68 ± 356.29	1688.76 ± 255.30	2963.65 ± 1360.88

Table 3: **Average energy consumption analysis**. Energy usage (milliwatt-hours, mWh) per sample of LA-ABSA and LLM-prompting (Gemma-3-27B) across few-shot settings. Results show mean ± standard deviation across the five datasets for TASD and ASQP tasks. We performed 18 pairwise comparisons between DLO, Paraphrase, and Prompting (3 comparisons $\times$ 2 tasks $\times$ 3 shot settings) using paired t-tests with Bonferroni-Holm correction ($p_{\text{adj}} \leq 0.05$), and all differences were found to be significant.

5. Discussion and Limitations

In this section, we discuss the key findings and potential limitations of our study.

First, our proposed approach LA-ABSA demonstrated superior performance compared to previous approaches that rely on only a few annotated examples and utilize smaller models (e.g., T5) rather than LLMs, which is an observation also made in other NLP tasks (Li et al., 2025; Huang et al., 2023). Data augmentation approaches such as Easy Data Augmentation (EDA) (Liesting et al., 2021), QAIE (yang Lu et al., 2025) and DS^2-ABSA (Xu et al., 2025) were adapted for TASD and ASQP tasks for the first time in this study, as previous work only applied these augmentation strategies to ABSA tasks with fewer sentiment elements. Nevertheless, LA-ABSA demonstrated much higher F1 scores across scenarios and datasets.

In contrast, LLM-based ICL marginally outperforms LA-ABSA in most cases, with few exceptions. Figure 2 demonstrates that LA-ABSA achieves substantially higher energy efficiency when processing more than 2,000 examples. Given the magnitude of this difference (e.g., for the TASD task, LA-ABSA + Paraphrase requires 2.47 mWh/example vs. Gemma-3-27B at 703.1 mWh/example), similar efficiency gains would likely be observed even when comparing ICL with smaller LLMs (e.g., 12B parameters) or single-execution approaches without self-consistency learning (majority voting mechanism).

In this context, a limitation of this study is the evaluation of only one LLM, primarily due to the extensive computational requirements of the investigated conditions. Examining reasoning models that generate intermediate reasoning steps, which have demonstrated superiority over non-reasoning approaches across various NLP tasks (Wei et al., 2022; Fei et al., 2023; Sun et al., 2023) might be a strategy to further improve the annotation quality. However, such models substantially increase inference time due to the higher number of output tokens generated.

Finally, LA-ABSA did not reach performance levels comparable to models fine-tuned on human-annotated data. On the ASQP task, LA-ABSA demonstrated an average performance decrease of 9 percentage points relative to the approach fine-tuned on human annotations.

6. Conclusion & Future Work

In this work, we introduced LA-ABSA, a novel approach that leverages LLM-generated annotations to fine-tune lightweight models for ABSA tasks. Our comprehensive evaluation across five datasets and two tuple prediction tasks (TASD and ASQP) demonstrated competitive performance with LLM-based in-context learning approaches while offering substantial energy efficiency advantages.

Future work could explore several promising directions. First, crowd-sourcing approaches could be explored as an alternative or complement to LLM-based annotation, potentially offering a middle ground between fully automated and expert annotation. Second, data synthesis approaches present an interesting avenue for investigation. While this study assumed the availability of non-annotated texts, low-resource scenarios may arise where such texts are unavailable. In such cases, similar to approaches practiced in other NLP tasks or ABSA tasks, methods could be developed to synthetically generate opinion pieces (e.g., reviews) using LLMs (Xu et al., 2025; Ye et al., 2022; Yu et al., 2023), as well as approaches that generate annotated training examples in a single step (Xu et al., 2025; Hellwig et al., 2025b).

7. Ethics Statement

This research was conducted without industrial funding or commercial sponsorship. All datasets contain annotated examples that have not been anonymized, except for FlightABSA. To ensure research transparency and reproducibility, code is available on GitHub. Claude Sonnet 4.5[5] was used to assist in the formulation of this publication.

[5] Claude Sonnet: https://www.anthropic.com/claude/sonnet

Finally, we want to highlight that LA-ABSA contributes to the reduction of the carbon footprint associated with large-scale model inference. By utilizing a substantially smaller base model and requiring fewer computational resources, LA-ABSA minimizes energy consumption. This resource-efficient design supports more sustainable research practices and aligns with broader efforts to mitigate the environmental impact of AI systems.

8. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yinhao Bai, Zhixin Han, Yuhua Zhao, Hang Gao, Zhuowei Zhang, Xunzhi Wang, and Mengting Hu. 2024. Is compound aspect-based sentiment analysis addressed by LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7836–7861, Miami, Florida, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: A challenge dataset for opinion aspect target sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12336–12347, Torino, Italia. ELRA and ICCL.

Xiang Chen, Chaoyang Gao, Chunyang Chen, Guangbei Zhang, and Yong Liu. 2025. An empirical study on challenges for llm application developers. *ACM Trans. Softw. Eng. Methodol.*, 34(7).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Bálint Csanády, Lajos Muzsai, Péter Vedres, Zoltán Nádasdy, and András Lukács. 2024. Llambert: Large-scale low-cost data annotation in nlp. *arXiv preprint arXiv:2403.15938*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Mansi Goel, Ayush Agarwal, Shubham Agrawal, Janak Kapuriya, Akhil Vamshi Konam, Rishabh Gupta, Shrey Rastogi, Niharika, and Ganesh Bagler. 2024. Deep learning based named entity recognition models for recipes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4542–4554, Torino, Italia. ELRA and ICCL.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nils Constantin Hellwig, Jakob Fehle, Udo Kruschwitz, and Christian Wolff. 2025a. Do we still need human annotators? prompting large language models for aspect sentiment quad prediction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 153–172, Vienna, Austria. Association for Computational Linguistics.

Nils Constantin Hellwig, Jakob Fehle, and Christian Wolff. 2025b. Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Systems with Applications*, 261:125514.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4417–4422, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022a. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022b. 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.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Mia Mohammad Imran, Yashasvi Jain, Preetha Chatterjee, and Kostadin Damevski. 2023. Data augmentation for improving emotion recognition in software engineering communication. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ASE '22, New York, NY, USA. Association for Computing Machinery.

Guangmin Li, Hui Wang, Yi Ding, Kangan Zhou, and Xiaowei Yan. 2023. Data augmentation for aspect-based sentiment analysis. *International Journal of Machine Learning and Cybernetics*, 14(1):125–133.

Juanhui Li, Sreyashi Nag, Hui Liu, Xianfeng Tang, Sheikh Muhammad Sarwar, Limeng Cui, Hansu Gu, Suhang Wang, Qi He, and Jiliang Tang. 2025. Learning with less: Knowledge distillation from large language models via unlabeled data. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2627–2641, Albuquerque, New Mexico. Association for Computational Linguistics.

Lin Li, Shaopeng Tang, and Renwei Wu. 2024. Majority rules guided aspect-category based sentiment analysis via label prior knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10952–10957, Torino, Italia. ELRA and ICCL.

Yijie Li and Yuan Sun. 2025. EasyJudge: an easy-to-use tool for comprehensive response evaluation of LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 91–103, Abu Dhabi, UAE. Association for Computational Linguistics.

Tomas Liesting, Flavius Frasincar, and Maria Mihaela Truşcă. 2021. Data augmentation in a hybrid approach for aspect-based sentiment analysis. In *Proceedings of the 36th annual ACM* *symposium on applied computing*, pages 828–835.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Yiding Liu, Jingjing Wang, Jiamin Luo, Tao Zeng, and Guodong Zhou. 2024. ChatASU: Evoking LLM's reflexion to truly understand aspect sentiment in dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3075–3085, Torino, Italia. ELRA and ICCL.

Usman Malik, Simon Bernard, Alexandre Pauchet, Clément Chatelain, Romain Picot-Clémente, and Jérôme Cortinovis. 2024. Pseudo-labeling with large language models for multi-label emotion classification of french tweets. *IEEE Access*, 12:15902–15916.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Gaurav Negi, Rajdeep Sarkar, Omnia Zayed, and Paul Buitelaar. 2024. A hybrid approach to aspect based sentiment analysis using transfer learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 647–658, Torino, Italia. ELRA and ICCL.

Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. 2024. An empirical study of synthetic data generation for implicit discourse relation recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1073–1085, Torino, Italia. ELRA and ICCL.

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

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Adir Rahamim, Guy Uziel, Esther Goldbraich, and Ateret Anaby Tavor. 2023. Text augmentation using dataset reconstruction for low-resource classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7389–7402, Toronto, Canada. Association for Computational Linguistics.

Caroline Sabty, Islam Omar, Fady Wasfalla, Mohamed Islam, and Slim Abdennadher. 2021. Data augmentation techniques on arabic data for named entity recognition. *Procedia Computer Science*, 189:292–299. AI in Computational Linguistics.

Maximilian Schmidt, Andrea Bartezzaghi, and Ngoc Thang Vu. 2024. Prompting-based synthetic data generation for few-shot question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13168–13178, Torino, Italia. ELRA and ICCL.

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

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2024. LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Bangkok, Thailand. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational*

Linguistics: EMNLP 2023, pages 8990–9005, Singapore. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction tuning for few-shot aspect-based sentiment analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. 2023a. Generative data augmentation for aspect sentiment quad prediction. pages 128–140.

An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. 2023b. Generative data augmentation for aspect sentiment quad prediction. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 128–140, Toronto, Canada. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Hongling Xu, Yice Zhang, Qianlong Wang, and Ruifeng Xu. 2025. DS^2-ABSA: Dual-stream data synthesis with label refinement for few-shot aspect-based sentiment analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15460–15478, Vienna, Austria. Association for Computational Linguistics.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Heng yang Lu, Tian ci Liu, Rui Cong, Jun Yang, Qiang Gan, Wei Fang, and Xiao jun Wu. 2025. Qaie: Llm-based quantity augmentation and information enhancement for few-shot aspect-based sentiment analysis. *Information Processing Management*, 62(1):103917.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: a tale of diversity and bias.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023a. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. 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 Pan, and Lidong Bing. 2024a. 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, 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, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023b. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges . *IEEE Transactions on Knowledge & Data Engineering*, 35(11):11019–11038.

Changzhi Zhou, Dandan Song, Yuhang Tian, Zhijing Wu, Hao Wang, Xinyu Zhang, Jun Yang, Ziyi Yang, and Shuhao Zhang. 2024. A comprehensive evaluation of large language models on aspect-based sentiment analysis. *arXiv preprint arXiv:2412.02279*.

A. Datasets: Aspect Category and Sentiment Distributions

Figure 3: **Distribution of aspect categories and sentiments across datasets for TASD and ASQP tasks.** Each subplot shows the top 10 aspect categories (sorted by total frequency), with stacked bars representing positive (green), neutral (yellow), and negative (red) sentiments. The 'Others' category aggregates the remaining aspects. Results are aggregated across five datasets: Rest15 (Zhang et al., 2021; Pontiki et al., 2015), Rest16 (Zhang et al., 2021; Pontiki et al., 2016), FlightABSA (Hellwig et al., 2025a), Coursera (Chebolu et al., 2024), and Hotels (Chebolu et al., 2024). This visualization highlights the imbalances in aspect-level sentiment annotations, showing varying distributions of polarities and aspect categories across datasets.

B. Ablation Study: EDA

Approach	# Aug	# Train		Rest15		Rest16		FlightABSA		Coursera		Hotels		AVG	
		TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP
Scenario: 10 annotated examples given															
EDA w/ Paraphrase	2	30	30	13.15	1.74	7.67	11.41	10.79	4.42	17.97	6.61	22.05	3.47	14.33	5.53
EDA w/ Paraphrase	5	60	60	23.20	10.42	16.35	12.30	19.26	10.99	23.62	9.47	25.85	6.15	21.66	9.86
EDA w/ Paraphrase	10	110	110	26.07	**10.52**	16.82	**12.68**	21.83	**13.46**	25.67	10.35	24.95	5.84	23.07	10.57
EDA w/ Paraphrase	15	160	160	25.53	10.08	18.34	12.32	21.57	12.39	25.60	12.33	24.94	7.03	23.20	10.83
EDA w/ DLO	2	30	30	24.58	6.74	16.97	9.27	21.18	12.13	28.14	9.75	22.73	5.98	22.72	8.77
EDA w/ DLO	5	60	60	26.65	8.68	18.06	10.70	21.66	12.40	28.25	11.73	23.95	7.51	23.71	10.20
EDA w/ DLO	10	110	110	**30.18**	9.34	19.16	11.80	**22.66**	11.95	**28.83**	**14.19**	25.65	**7.88**	**25.30**	**11.03**
EDA w/ DLO	15	160	160	29.10	8.07	**19.59**	11.83	22.05	12.13	27.40	13.36	**27.04**	6.94	25.04	10.46
Scenario: 50 annotated examples given															
EDA w/ Paraphrase	2	150	150	42.70	27.55	41.59	28.78	40.73	28.26	36.47	19.90	43.44	27.96	40.98	26.49
EDA w/ Paraphrase	5	300	300	42.29	28.96	42.57	29.40	42.80	30.32	36.85	21.08	44.67	28.04	41.84	27.56
EDA w/ Paraphrase	10	550	550	41.83	28.70	43.01	30.20	41.62	31.04	35.84	21.37	42.70	27.97	41.00	27.85
EDA w/ Paraphrase	15	800	800	40.78	27.81	40.81	28.80	42.02	31.92	36.49	21.22	43.45	27.38	40.71	27.43
EDA w/ DLO	2	150	150	43.86	30.41	45.77	33.52	44.11	34.84	37.81	23.53	**45.89**	31.39	43.49	30.74
EDA w/ DLO	5	300	300	43.95	30.90	**46.28**	34.42	44.90	36.37	**38.32**	23.53	45.20	**32.55**	43.73	31.55
EDA w/ DLO	10	550	550	**44.14**	**30.93**	45.72	**34.59**	**45.64**	**36.73**	37.89	**24.57**	45.72	32.51	**43.82**	**31.86**
EDA w/ DLO	15	800	800	41.71	28.82	42.10	31.68	45.57	34.09	36.55	23.33	44.61	31.13	42.11	29.81

Table 4: **Results of ablation study for EDA-based data augmentation.** F1 scores are reported across different augmentation ratios (2, 5, 10, 15) and fine-tuning approaches (DLO, Paraphrase) for scenarios with 10 and 50 annotated examples. Results demonstrate that 10 augmentations combined with DLO achieves optimal performance across most datasets. Bold values indicate the highest F1 scores within a scenario assuming either 10 or 50 given annotated examples.

C. LA-ABSA: Efficiency in Terms of Time

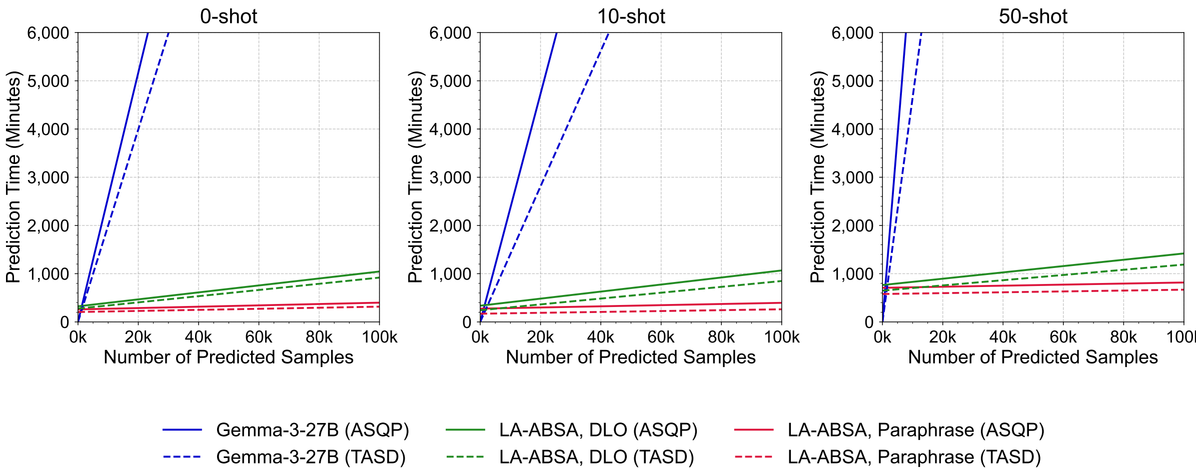

Figure 4: **Prediction time analysis**. Prediction time in minutes for ASQP and TASD tasks across different settings (0, 10, or 50 annotated examples given). Results are shown for LLM-prompting (Gemma-3-27B) and LA-ABSA using DLO (Hu et al., 2022b) and Paraphrase (Zhang et al., 2021) for fine-tuning. Each line represents the average time required to predict up to 100,000 examples per method and task across the five datasets. LA-ABSA approaches generally require substantially less time due to their smaller underlying base model T5-base.

D. Performance: Precision, Recall and F1 Macro

D.1. Precision

Approach	**# Train**		**Rest15**		**Rest16**		**FlightABSA**		**Coursera**		**Hotels**		**AVG**	
	TASD	**ASQP**	**TASD**	**ASQP**	**TASD**	**ASQP**	**TASD**	**ASQP**	**TASD**	**ASQP**	**TASD**	**ASQP**	**TASD**	**ASQP**
Scenario: 0 annotated examples given														
Gemma-3-27B (0-shot)	0	0	**29.41**	**23.35**	44.49	**27.75**	47.55	**39.70**	26.95	**11.95**	37.12	22.88	37.11	**25.13**
LA-ABSA														
w/ Paraphrase	full	full	26.91	17.09	**45.93**	24.00	**48.95**	39.41	**28.32**	11.68	**38.53**	**25.41**	**37.73**	23.52
w/ DLO	full	full	24.44	17.21	45.68	24.62	47.41	39.06	27.52	11.48	37.98	23.83	36.61	23.24
Scenario: 10 annotated examples given														
Gemma-3-27B ICL	10	10	**56.40**	**39.41**	**68.38**	**44.64**	59.85	45.39	**43.11**	**23.41**	**57.93**	**35.29**	**57.13**	**37.63**
Paraphrase	10	10	10.72	1.64	7.61	4.02	10.44	4.34	17.69	5.35	18.74	3.66	13.04	3.80
DLO	10	10	19.23	4.64	13.27	5.49	19.02	6.15	25.45	5.03	18.84	3.68	19.16	5.00
LA-ABSA														
w/ Paraphrase	full	full	51.41	33.75	63.75	43.35	61.12	**45.85**	38.28	21.98	57.70	34.80	54.45	35.95
w/ DLO	full	full	50.33	35.54	63.51	43.85	**61.29**	45.63	38.82	22.79	56.07	33.10	54.00	36.18
EDA	110	110	32.49	9.16	19.32	11.92	26.82	14.10	31.08	15.52	26.15	8.61	27.17	11.86
QAIE	26.4	45.2	26.44	12.71	15.71	17.42	22.37	20.19	28.49	19.70	29.39	12.40	24.48	16.49
DS^2-ABSA	21,1k	21,1k	29.55	18.10	33.30	22.72	27.34	17.33	17.40	7.43	32.17	12.03	27.95	15.52
Scenario: 50 annotated examples given														
Gemma-3-27B ICL	50	50	**68.03**	**44.57**	**71.52**	**54.55**	**66.33**	**51.95**	**51.32**	**31.96**	**70.47**	**53.39**	**65.54**	**47.28**
Paraphrase	50	50	39.57	24.58	37.18	23.75	36.10	18.58	37.64	20.72	45.21	23.67	39.14	22.26
DLO	50	50	40.48	24.92	44.59	29.09	42.01	28.30	39.26	20.44	49.18	28.54	43.10	26.26
LA-ABSA														
w/ Paraphrase	full	full	60.49	38.23	63.32	47.39	64.28	48.06	46.39	27.13	65.43	45.94	59.98	41.35
w/ DLO	full	full	62.07	40.23	63.47	49.64	62.88	49.02	46.50	27.01	65.58	46.61	60.10	42.50
EDA	550	550	45.93	31.12	46.84	34.87	46.81	38.17	39.96	26.28	49.55	33.48	45.82	32.78
QAIE	141.8	248.8	48.89	35.07	48.29	37.15	51.71	37.13	37.70	23.57	58.26	38.78	48.97	34.34
DS^2-ABSA	21,7k	21,6k	37.80	27.92	43.13	34.96	32.25	23.58	29.42	16.03	42.16	24.01	36.95	25.30
Full set of human-annotated examples: SOTA approaches														
Paraphrase	full	full	-	46.16	-	56.63	**70.22**	**57.37**	52.73	**32.06**	**68.41**	52.61	**63.79**	48.97
DLO	full	full	-	**47.08**	-	**57.92**	68.60	56.67	**52.79**	32.03	**68.41**	**54.39**	63.27	**49.62**

Table 5: **Precision scores of LA-ABSA**. Results are evaluated against EDA-based data augmentation methods, QAIE, DS^2-ABSA, and prompting baselines (0, 10, and 50 annotated examples) as reported by Hellwig et al. (2025a), as well as fully supervised models including DLO (Hu et al., 2022b) and Paraphrase (Zhang et al., 2021). The highest precision scores within each annotation regime (0, 10, 50 or all examples) are shown in bold; the best overall scores across all settings are underlined.

D.2. Recall

Approach	# Train		Rest15		Rest16		FlightABSA		Coursera		Hotels		AVG	
	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP
Scenario: 0 annotated examples given														
Gemma-3-27B (0-shot)	0	0	**31.36**	**26.29**	46.57	**30.29**	56.90	45.42	32.58	15.14	41.02	23.16	41.69	28.06
LA-ABSA														
w/ Paraphrase	full	full	28.54	20.25	**49.68**	28.54	55.69	46.31	34.06	15.74	43.22	**27.18**	42.24	27.60
w/ DLO	full	full	27.36	20.91	49.27	29.81	**56.98**	**47.39**	**34.47**	**16.02**	**43.85**	26.21	**42.38**	**28.07**
Scenario: 10 annotated examples given														
Gemma-3-27B ICL	10	10	**52.66**	**40.50**	**65.19**	47.93	60.87	45.08	**40.37**	21.31	**55.17**	28.29	**54.85**	36.63
Paraphrase	10	10	7.38	1.11	5.93	3.23	7.64	2.85	14.51	4.26	12.39	2.06	9.57	2.70
DLO	10	10	13.47	4.13	13.95	4.91	13.91	4.03	20.86	4.02	17.39	3.41	15.92	4.10
LA-ABSA														
w/ Paraphrase	full	full	46.98	36.43	61.76	45.91	58.68	45.93	38.65	22.99	53.82	**31.07**	51.98	36.47
w/ DLO	full	full	48.19	38.99	61.28	**48.81**	**61.51**	**47.36**	39.63	**23.98**	54.50	29.93	53.02	**37.82**
EDA	110	110	28.19	9.53	19.02	11.69	19.62	10.37	26.89	13.07	25.18	7.27	23.78	10.39
QAIE	26.4	45.2	16.24	8.19	10.24	12.84	14.73	12.42	22.56	15.81	17.19	8.19	16.19	11.49
DS^2-ABSA	21,1k	21,1k	28.54	18.29	30.96	24.38	26.69	15.73	18.24	7.81	31.59	11.54	27.21	15.55
Scenario: 50 annotated examples given														
Gemma-3-27B ICL	50	50	57.16	39.25	65.77	48.06	**62.95**	45.25	39.75	21.71	56.92	37.17	**56.51**	38.29
Paraphrase	50	50	34.60	26.62	34.66	23.25	31.38	17.42	31.43	18.21	36.05	22.59	33.63	21.62
DLO	50	50	38.65	28.60	43.33	30.06	43.89	29.22	33.32	17.89	41.02	25.99	40.04	26.35
LA-ABSA														
w/ Paraphrase	full	full	52.50	37.01	61.12	46.16	60.76	46.31	**42.50**	24.42	56.40	40.75	54.65	38.93
w/ DLO	full	full	55.15	**40.53**	60.65	**50.06**	62.27	**48.95**	42.46	**24.50**	**57.77**	**42.11**	55.66	**41.23**
EDA	550	550	42.49	30.74	44.66	34.32	44.54	35.39	36.02	23.07	42.45	31.60	42.03	31.02
QAIE	141.8	248.8	41.70	32.78	42.30	33.46	45.57	31.33	34.87	21.44	44.60	33.52	41.81	30.51
DS^2-ABSA	21,7k	21,6k	38.08	30.01	43.94	36.67	34.63	24.17	29.71	16.81	38.54	23.23	36.98	26.18
Full set of human-annotated examples: SOTA approaches														
Paraphrase	full	full	-	47.72	-	59.30	69.26	58.17	51.02	32.63	67.01	55.19	62.43	50.60
DLO	full	full	-	**49.33**	-	**61.80**	**69.30**	**60.10**	**52.38**	**33.07**	**68.71**	**56.56**	**63.46**	**52.17**

Table 6: **Recall scores of LA-ABSA**. Results are evaluated against EDA-based data augmentation methods, QAIE, DS^2-ABSA, and prompting baselines (0, 10, and 50 annotated examples) as reported by Hellwig et al. (2025a), as well as fully supervised models including DLO (Hu et al., 2022b) and Paraphrase (Zhang et al., 2021). The highest recall scores within each annotation regime (0, 10, 50 or all examples) are shown in bold; the best overall scores across all settings are underlined.

D.3. F1 Macro

Approach	# Train		Rest15		Rest16		FlightABSA		Coursera		Hotels		AVG	
	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP	TASD	ASQP
Scenario: 0 annotated examples given														
Gemma-3-27B (0-shot)	0	0	**33.68**	**23.64**	**45.94**	23.64	**55.24**	**39.50**	**33.78**	**13.23**	**29.70**	**18.49**	**39.67**	**23.70**
LA-ABSA														
w/ Paraphrase	full	full	22.88	19.93	35.84	26.20	45.95	33.61	11.85	4.09	16.20	12.44	26.54	19.25
w/ DLO	full	full	24.12	21.02	36.12	**29.16**	45.37	36.32	12.10	3.70	17.18	13.10	26.98	20.66
Scenario: 10 annotated examples given														
Gemma-3-27B ICL	10	10	**54.53**	**38.67**	**64.33**	**43.21**	**56.32**	36.58	**41.75**	**20.44**	**31.15**	18.05	**49.62**	**31.39**
Paraphrase	10	10	3.47	0.46	2.59	0.87	2.38	1.03	2.17	0.66	2.65	0.47	2.65	0.70
DLO	10	10	6.10	1.42	6.01	1.25	4.74	1.35	3.33	0.60	4.10	0.83	4.86	1.09
LA-ABSA														
w/ Paraphrase	full	full	34.17	23.98	48.59	34.59	51.20	36.29	17.01	8.54	26.90	**19.19**	35.57	24.52
w/ DLO	full	full	35.03	26.77	50.39	36.25	53.19	**39.01**	17.14	9.38	29.02	18.18	36.95	25.92
EDA	110	110	11.33	5.57	7.94	2.22	8.05	3.64	4.22	2.02	5.49	2.27	7.40	3.15
QAIE	26.4	45.2	14.86	5.37	7.84	14.29	12.02	7.00	4.30	1.90	3.52	2.86	8.51	6.28
DS^2-ABSA	21,1k	21,1k	18.83	12.17	26.68	15.60	22.66	13.04	9.65	3.17	16.97	7.89	18.96	10.37
Scenario: 50 annotated examples given														
Gemma-3-27B ICL	50	50	**56.01**	**36.01**	**64.81**	**44.56**	**54.06**	36.67	**38.24**	**17.80**	**35.47**	**25.24**	**49.72**	**32.06**
Paraphrase	50	50	19.75	14.05	16.35	8.35	12.43	7.23	9.59	2.99	13.87	7.54	14.40	8.03
DLO	50	50	27.78	18.10	22.46	13.05	28.90	19.71	8.68	3.11	16.50	9.74	20.86	12.74
LA-ABSA														
w/ Paraphrase	full	full	35.74	26.57	48.34	31.01	52.88	36.65	17.96	7.37	28.20	19.79	36.62	24.28
w/ DLO	full	full	41.64	28.42	49.37	37.37	52.17	**40.88**	18.07	8.10	30.19	22.59	38.29	27.47
EDA	550	550	27.35	17.61	22.38	16.12	33.71	24.74	12.22	5.72	15.85	13.07	22.30	15.45
QAIE	141.8	248.8	44.81	30.84	45.50	32.23	37.41	22.29	20.79	10.20	24.19	18.51	34.54	22.82
DS^2-ABSA	21,7k	21,6k	24.95	18.14	31.43	21.81	28.52	17.92	11.86	4.69	18.35	13.06	23.02	15.12
Full set of human-annotated examples: SOTA approaches														
Paraphrase	full	full	-	-	-	-	**58.48**	48.47	30.67	**15.66**	37.49	30.31	42.21	31.48
DLO	full	full	-	-	-	-	57.24	**49.41**	**32.25**	14.91	**38.14**	**33.04**	**42.54**	**32.45**

Table 7: **Macro-averaged F1 scores of LA-ABSA.** Results are evaluated against EDA-based data augmentation methods, QAIE, DS^2-ABSA, and prompting baselines (0, 10, and 50 annotated examples) as reported by Hellwig et al. (2025a), as well as fully supervised models including DLO (Hu et al., 2022b) and Paraphrase (Zhang et al., 2021). The highest F1 macro scores within each annotation regime (0, 10, 50 or all examples) are shown in bold; the best overall scores across all settings are underlined.