Single Ground Truth Is Not Enough: Adding Flexibility to Aspect-Based Sentiment Analysis Evaluation

Soyoung Yang[♡] **Hojun Cho**[♡] **Jiyoung Lee**[♡] **Sohee Yoon**[♣]
Edward Choi[♡] **Jaegul Choo**[♡] **Won Ik Cho**[♣]
[♡] KAIST AI [♣] AI Center, Samsung Electronics
sy_yang@kaist.ac.kr

Abstract

Aspect-based sentiment analysis (ABSA) is a challenging task of extracting sentiments along with their corresponding aspects and opinion terms from the text. The inherent subjectivity of span annotation makes variability in the surface forms of extracted terms, complicating the evaluation process. Traditional evaluation methods often constrain ground truths (GT) to a single term, potentially misrepresenting the accuracy of semantically valid predictions that differ in surface form. To address this limitation, we propose a novel and fully automated pipeline that expands existing evaluation sets by adding alternative valid terms for aspect and opinion. Our approach facilitates an equitable assessment of language models by accommodating multiple-answer candidates, resulting in enhanced human agreement compared to single-answer test sets (achieving up to a 10%p improvement in Kendall's Tau score). Experimental results demonstrate that our expanded evaluation set helps uncover the capabilities of large language models (LLMs) in ABSA tasks, which is concealed by the single-answer GT sets. Consequently, our work contributes to the development of a flexible evaluation framework for ABSA by embracing diverse surface forms to span extraction tasks in a cost-effective and reproducible manner. Our code and dataset is open at https://github.com/dudrrm/zoom-in-n-out-absa.

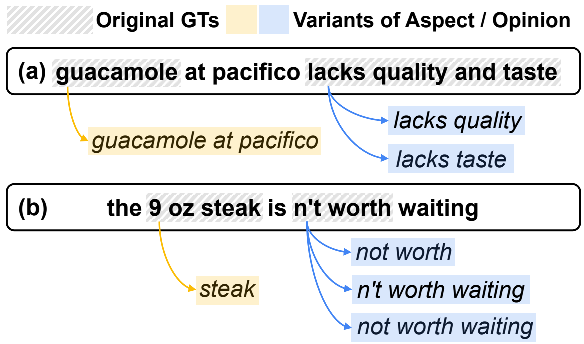

Figure 1: Original ground truth (GT) aspect and opinion terms are shown in gray background shading. Alternative valid expressions for these terms are highlighted: aspects in yellow and opinions in blue. Under conventional evaluation approaches using the original GT set, these highlighted candidates are assessed as *wrong* because they do not match exactly despite being semantically consistent with the original terms.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a sophisticated natural language processing task that aims to extract fine-grained sentiment information from text. ABSA is usually utilized to provide insights into opinions about specific attributes of products or services, enabling organizations or persons to precisely identify consumer preferences and criticisms. The granularity of ABSA facilitates targeted improvements and data-driven decision-making across diverse domains, including product development, customer service, political analysis, and e-commerce (Do et al., 2019; Zhang et al., 2022). ABSA includes a subtask known as aspect sentiment quadruple prediction (ASQP), which involves identifying four elements within the text: (aspect, category, sentiment, opinion). For illustration, consider the sentence in Figure 1 (a). From this, we can extract the quadruple of ("guacamole", Food Quality, Negative, "lacks quality and taste"). Here, the gray-shaded terms represent the ground truth (GT) annotations: "guacamole" as the aspect and "lacks quality and taste" as the opinion.

Due to the richness and variability of natural language, there are diverse forms in which these elements are expressed and consequently annotated. Different annotators may select different spans to represent the same semantic content, producing various surface forms in span annotations. In Figure 1, the terms highlighted in color illustrate alternative

[♣]Work is done while affiliated to SAIT.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 12071–12096
April 29 - May 4, 2025

candidates that can be considered valid annotations.

However, conventional ABSA benchmarks (Pontiki et al., 2014, 2015, 2016; Zhang et al., 2021; Cai et al., 2021) typically provide a single GT term for each aspect and opinion pair and employ exact match criteria for evaluation. This approach fails to account for the compatibility of multiple surface forms. As illustrated in Figure 1, semantically valid surface forms (highlighted in color), such as "guacamole at pacifico" for the aspect or "lacks quality" and "lacks taste" for the opinion, are evaluated as *incorrect*, while they preserve the essential meaning of the original GT quadruple. Consequently, this rigid evaluation schema may underestimate the performance of advanced language models in understanding and generating semantically valid predictions, raising concerns about the fairness and comprehensiveness of current model assessments in ABSA.

To address the limitations of conventional ABSA evaluation approaches, which often fail to account for the variability of natural language expressions, we propose ZOOM IN-N-OUT, a novel approach for evaluation in ABSA task [1]. Our method aims to accommodate diverse surface forms that preserve the semantic content of the original GT quadruple. ZOOM IN-N-OUT pipeline comprises three key steps: (i) identifying alternative terms within the original GT span, (ii) exploring potential expressions outside the span, and (iii) verifying that new candidates are semantically consistent with the original GT. This process leverages the capabilities of large language models (LLMs), specifically GPT-4, which has demonstrated quality in annotation tasks (Gilardi et al., 2023) and efficacy in generating synthetic datasets (Zheng et al., 2023; Li et al., 2023). Notably, our approach offers several advantages over existing methods, such as N-gram overlap metrics or direct LLM/human evaluation. N-gram-based methods like F-measure (Melamed et al., 2003) may fail to capture semantic nuances; for instance, they may incorrectly score "not worth" and "worth" as similar when GT is "n't worth". Direct LLM/human evaluation (Zheng et al., 2023; Wadhwa et al., 2023) can be costly and difficult to reproduce since the evaluator model should be run when a new prediction occurs. In contrast, our method provides a balanced solution by ensuring semantic consistency with the original GT and employing LLMs solely during the GT expansion phase, allowing the subsequent evaluations to be cost-effective and reproducible.

Experimental results demonstrate the quality and effectiveness of our method. Our expanded GT set achieves over 90% validity in human assessment and shows improved human alignment, with up to a 10%p increase in Kendall Tau score compared to the original GT set. Furthermore, compared to T5-based models, the F1 scores of LLMs jump up in our evaluation set when the train set (in-context examples) comes from the original dataset. For instance, four LLMs increase their F1 score by 9.8%p, but T5 models increase by only 2.3%p on average. This significant score gap suggests that the conventional evaluation framework may have underestimated the capabilities of LLMs. It highlights how our method uncovers their hidden ability through our flexible evaluation framework.

In summary, the contributions of our paper can be summarized as follows:

- We introduce ZOOM IN-N-OUT, a fully automated pipeline that addresses the diversity of expressions in the ABSA ground truth set by expanding aspect and opinion terms in existing datasets.
- Our experimental results demonstrate that our expanded truth set aligns closely with human judgments more than the conventional test set, highlighting that our dataset has effectively reflected the linguistic variability in ABSA evaluation.
- We conduct an extensive comparative study of ABSA performance between LLMs and ABSA-specific T5 models, revealing the underestimated potential of LLMs in ABSA tasks and proposing a more equitable evaluation framework for the LLM era.

2 Related work

2.1 LLM in Information Extraction

The advent of proprietary LLMs, such as GPT (Brown et al., 2020) and Gemini (GeminiTeam, 2024), along with open-sourced models like Llama (Touvron et al., 2023; LlamaTeam, 2024) and its variants (Taori et al., 2023; Chiang et al., 2023), has led to their widespread adoption in natural language processing (NLP) (Liu et al., 2022; Min et al., 2022; Yoo et al., 2022). In infor-

[1]While implicit terms (*e.g.*, "null") are considered in ABSA benchmarks (Zhang et al., 2021; Cai et al., 2021), our method aims to expand explicit GT expressions in aspect and opinion terms. We leave this exploration for future work.

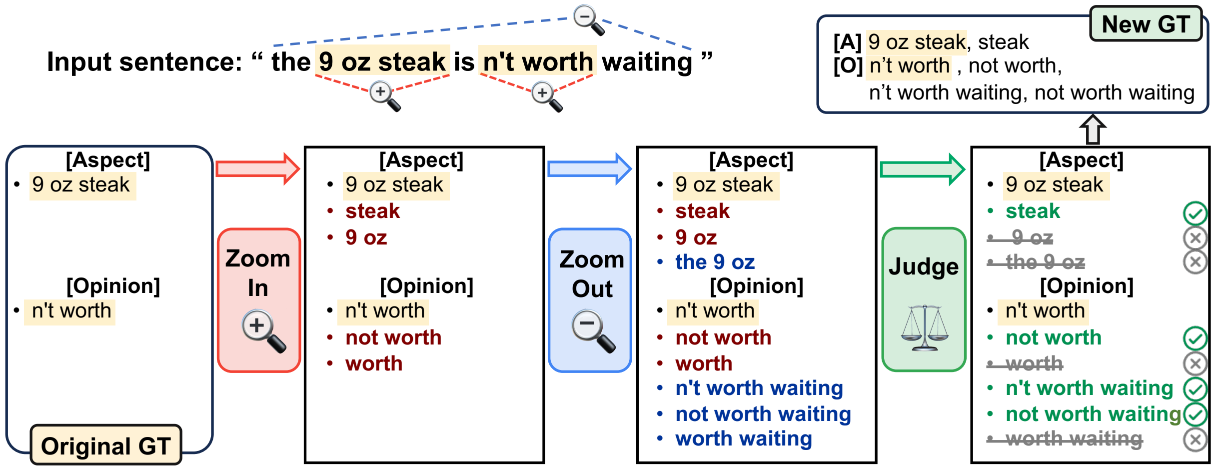

Figure 2: Overview of ZOOM IN-N-OUT, illustrating the extraction of diverse candidates of the aspect and opinion terms in ABSA task. The three stages of our pipeline are depicted by three colored blocks. Starting with the input sentence and the original ground truth term highlighted in yellow, the model generates alternative representations of the target term at each stage. The original ground truth quadruple is ("9 oz steak", Food Quality, Negative, "n't worth"), representing aspect, category, sentiment, and opinion, respectively. Since our pipeline preserves the provided categorical values, the category and sentiment are not included in the figure.

Task	Output
Aspect term extraction (ATE)	a
Aspect-opinion pair extraction (AOPE)	a, o
Aspect-sentiment pair extraction (ASPE)	a, s
Aspect-sentiment triplet extraction (ASTE)	a, o, s
Aspect-category-sentiment detection (ACSD)	$a, c\ s$
Aspect-sentiment quad prediction (ASQP)	a, c, s, o

Table 1: Target output elements for ABSA tasks. In ASQP, the output is represented as a quadruple: (aspect term a, aspect category c, sentiment polarity s, opinion term o). The names of the tasks follow Zhang et al. (2022).

mation extraction tasks, LLMs have been applied to various tasks, including named entity recognition (Xie et al., 2023; Wu et al., 2024), relation extraction (Wan et al., 2023; Wadhwa et al., 2023; Ma et al., 2024), and event extraction (Pang et al., 2023; Wang et al., 2023b). Data augmentation with LLMs has also been actively studied (Xu et al., 2023; Wang et al., 2023a). While prevalent data augmentation approaches (Zhang et al., 2015; Kobayashi, 2018) aim to improve training performance by increasing the training set size, our work focuses on re-evaluating model performance by considering diverse surface forms in the evaluation set.

2.2 Aspect-based Sentiment Analysis

Current ABSA datasets are primarily derived from SemEval tasks, including aspect-sentiment triplet extraction (Pontiki et al., 2014, 2015, 2016) and target-aspect sentiment joint detection (Saias, 2015). As shown in Table 1, ABSA started as an aspect-based sentiment analysis but has evolved into different formats, such as target-oriented opinion words extraction (Fan et al., 2019), aspect-sentiment quad prediction (ASQP) (Zhang et al., 2021; Cai et al., 2021). In this paper, we focus on the ASQP task, where the output is a set of quadruples.

With the advent of pre-trained language models, BERT-based architectures (Devlin et al., 2019) have become dominant in ABSA research (Zhang et al., 2022; Cai et al., 2023; Gao et al., 2022; Wang et al., 2024). Recent ASQP approaches have adopted T5-style encoder-decoder architectures (Raffel et al., 2020), including Paraphrase (Zhang et al., 2021), which derives sentence structures fitting quadruple prediction; DLO (Hu et al., 2022), which augments datasets through element permutation; and MvP (Gou et al., 2023), which creates and ensembles quadruples of multiple element orders for majority-voted predictions. Including DLO and MvP, various studies (Li et al., 2020; Hsu et al., 2021) utilize data augmentation approaches to alter the training set and improve model performance. Our approach, however, differs significantly. We focus on establishing a fair ABSA evaluation framework by expanding the evaluation set to account for linguistic variability in aspect and opinion terms, without modifying the original input sentences or training dataset.

3 Method

Our pipeline aims to expand the original GT aspect and opinion terms to various equivalent expressions, to encompass the diversity of surface forms in the ABSA task. Based on Figure 2, we introduce Zoom In-N-Out and describe how the exact match measure is calculated on our GT sets.

3.1 Zoom In-N-Out

Let $D = \{(x_k, Y_k)\}_{k=1}^{N_{\text{data}}}$ denote test set, where N_{data} represents the total number of test examples. Consider $d = (x, Y) \in D$ where Y comprises one or more ground truth (GT) quadruples, such that $Y = \{y_i\}_{i=1}^{N}$ for $N \geq 1$. Each GT quadruple $y \in Y$, derived from x, is defined as (a, c, s, o), representing aspect, category, sentiment, and opinion, respectively. As illustrated in Figure 2, our pipeline consists of three steps separately applied to aspect and opinion terms before merging the results. In the following section, we will explain each step, focusing on the case of the opinion term, *i.e.*, "n't worth" from Figure 2.

Zoom-In Beginning with the original GT opinion term o in y and the given sentence x, the LLM (hereafter 'the model') investigates the inside of o, either by reshaping the given term or extracting meaningful components. From a given opinion term, the model processes "n't worth" to generate (i) "not worth" by resolving the contraction "n't" to "not" and (ii) "worth" as a partial component.

Zoom-Out Following the internal examination of o, we guide the model to explore the entire sentence x to identify and incorporate adjacent words. From a given sentence, the model extracts combinations such as "n't worth waiting", "not worth waiting", and "worth waiting" by integrating "waiting", which follows the original opinion term.

Judge and Filter Newly generated terms may not always be appropriate due to the hallucination. To exclude unsuitable terms, we establish four criteria: (i) relevance to aspect and category, (ii) consistency with opinion and sentiment, (iii) extractability from x, and (iv) independence from other GT terms, which is based on the SemEval annotation guideline (Pontiki et al., 2016). Following the LLM-as-a-judge (Zheng et al., 2023; Madaan et al., 2023), the model verifies if a new term meets all four criteria. We exclude the new term if any of the criteria are unmet. In Figure 2, "worth" and "worth waiting" are excluded for violating the second criterion by failing to maintain the negative sentiment of the original term.

Algorithm 1 Zoom In-N-Out

Input: $D = \{d_i\} = \{(x_k, Y_k)\}$
for $d = (x, Y = \{y\}) \in D$ **do**
  $Y' \leftarrow []$
  for $y = (a, c, s, o) \in Y$ **do**
    $A \leftarrow \{a\}, O \leftarrow \{o\}$
    for $E \in \{A, O\}$ **do**
      $E \leftarrow \text{Zoom-In}(x, y, E)$
      $E \leftarrow \text{Zoom-Out}(x, y, E)$
      $E \leftarrow \text{Judge}(x, y, E)$
    end for
    $y' \leftarrow \{(a', c, s, o') \mid a' \in A, o' \in O\}$
    $y' \leftarrow \text{Filter}(y')$
    $Y' \leftarrow Y' + y'$
  end for
  $d' \leftarrow (x, Y')$
end for
Output: $D^{\text{new}} = \{d'_i\} = \{(x_k, Y'_k)\}$

After completing the same procedure for both aspect and opinion terms, we obtain an aspect term set $A = \{a_n\}_{n=1}^{N_a}$ and an opinion term set $O = \{o_m\}_{m=1}^{M_o}$, where N_a and M_o represent the number of aspect and opinion terms, respectively. Each set includes the original term as its first. The new GT (y^{new}) becomes a list of (a_n, c, s, o_m) where $a_n \in A$ and $o_m \in O$, yielding $N_a \times M_o$ combinations for each aspect and opinion term. By applying the pipeline to all $y_i \in Y$, we obtain $Y^{\text{new}} = \{y_i^{\text{new}}\}_{i=1}^{N}$ where each y_i^{new} contains all new combinations for y_i. We perform simple post-processing to remove potential duplicates within Y^{new}. Repeating this procedure across all $d_k = (x_k, Y_k) \in D$, we achieve an expanded test set $D^{\text{new}} = \{(x_k, Y_k^{\text{new}})\}_{k=1}^{N_{\text{data}}}$.

Algorithm 1 formalizes our process, with newly added components denoted by ($'$). All three functions Zoom-In, Zoom-Out, and Judge correspond to the above steps, taking $E \in \{A, O\}$ as input and updating it based on the given x, y and the established criteria. While diverse prompting methods can be applied at each step of Zoom-In, Zoom-Out, and Judge, we adopt a straightforward approach: 5-shot ICL for Zoom-In and Zoom-Out steps and 5-shot chain-of-thought (CoT) for Judge. The specific details of our experimental design are elaborated in Section 4. Detailed prompts and examples are provided in Appendix D.

3.2 Measurement

For a given task T, here ABSA (focusing on the ASQP task), let a sentence x be an input sentence for a language model f that predicts one or more quadruples. Given the principles of T, each x has a ground truth (Y), a set of quadruples y_i where each y_i consists of (a, c, s, o). Let $\hat{Y} = f(x)$ be the prediction, a set of predicted quadruples $\{\hat{y}_j\}$. $\hat{y}_j$ is considered a true positive only if $\hat{y}_j$ in Y, following the conventional exact match (EM) criterion.

Given our process ZOOM IN-N-OUT as Z applied to D such that $D^{\text{new}} = Z(D)$, we calculate $F1(f, D^{\text{new}}; T, EM)$ instead of the original $F1(f, D; T, EM)$. Our method employs the conventional metrics, *i.e.*, EM and $F1$, in the same manner, with the key difference being the expansion of each ground truth $Y = y_i$ to $Y^{\text{new}} = y_i^{\text{new}}$, where $y_i \in y_i^{\text{new}}$ for every y_i and y_i^{new}. This expansion does not necessarily increase the number of effective GTs, but rather broadens the scope of quadruple candidates counted as a GT.

4 Experiments

In this section, we first assess the validity of our expanded GTs, show that our evaluation set aligns more closely with human evaluators than the original, and analyze diverse baseline models with the original and our GT sets.

4.1 Experimental Setting

Construction of Our Evaluation Set We expand the GT quadruples of the ACOS and ASQP test sets with ZOOM IN-N-OUT. The pipeline processes each aspect and opinion term independently before merging the results. All the expansion processes utilize *gpt-4o-2024-05-1*, with an average cost of $75 per extended dataset. ZOOM-IN and ZOOM-OUT steps employ 5-shot ICL to collect possible quadruples (examples in Table 18). To maximize the diversity, we perform three generations with a temperature of 0.3. JUDGE uses 5-shot CoT prompting to verify each quadruple (examples in Table 21).

Datasets ASQP (Zhang et al., 2021) and ACOS (Cai et al., 2021) datasets are utilized. For ACOS-Laptop, which contains 121 total categories, we focus on the 23 entity-level categories, since the category comprises two levels: entity, *e.g.*, "laptop", and attribute, *e.g.*, "price".

Base models We evaluate both T5 and LLM-based ABSA models in a few-shot setting to compare the original GT set Y with our extended set Y^{new} produced by ZOOM IN-N-OUT. All experiments use the same 20 examples per dataset,

Datasets	# Preds. MvP	# Preds. GPT 3.5	# GTs. Orig.	# GTs. New
ASQP-R15	137	147	128	418
ASQP-R16	120	141	110	292
ACOS-Laptop	108	131	114	255
ACOS-Rest	137	15	132	355

Table 2: Number of quadruples utilized in human evaluation study. The predicted quadruples for two models (#Preds.) and the number of evaluated GT quadruples (#GTs).

	ASQP-R15	ASQP-R16	ACOS-L	ACOS-R
Agr.	99.04	97.95	96.47	93.52

Table 3: Human agreement (%) on the validity of our expanded test set. ACOS-L and ACOS-R represent ACOS-Laptop and ACOS-Rest, respectively.

and the heuristic selection criteria can be found in Appendix D.3. Representative T5-based fine-tuning methodologies for ABSA include Paraphase (Zhang et al., 2021), DLO (Hu et al., 2022), and MVP (Gou et al., 2023). Following the low-resource learning protocol of MvP, each model undergoes initial fine-tuning on the ASTE dataset (Peng et al., 2020) for 20 epochs, followed by transfer learning with 20 shots on each dataset for 100 epochs. Each model is trained five times with different seeds to measure the average performance. For LLMs, we employ several LLMs with 20-shot ICL: GPT-3.5-Turbo in *gpt-3.5-turbo-0125*, GPT-4o-mini in *gpt-4o-mini-2024-07-18*, Gemini-v1.5-Pro (GeminiTeam, 2024), and Llama-3.1-70B-Instruct (LlamaTeam, 2024) [2]. System prompts are derived from the SemEval annotation guidelines (Pontiki et al., 2016). LLM-based predictions utilize greedy decoding with temperature set to 0. The few-shot examples are provided in Table 22.

4.2 Dataset Validity

To assess the validity of our expanded evaluation sets, we conduct human evaluations on our newly generated GTs. Specifically, we randomly sample 80 examples from each dataset. Table 2 shows the number of quadruples for prediction and GT sets. Three human evaluators are asked to determine whether our expanded GTs are valid considering the given sentence. Before the evaluation, the evaluators were trained on the task and ensured to have

[2]The experiments are done in April, May, September, and October 2024.

Models	ASQP-Rest15		ASQP-Rest16		ACOS-Laptop		ACOS-Rest	
	Orig.	Ours	Orig.	Ours	Orig.	Ours	Orig.	Ours
MVP	33.4 / 42.3	**39.5 / 47.4**	56.6 / 61.8	**58 / 64.1**	47.6 / 58.1	**52.7 / 59.8**	47.3 / 54.5	47.3 / 54.5
GPT-3.5-Turbo	8.3 / 18.8	**18 / 29.8**	26.3 / 38.9	**48.9 / 48.9**	12.8 / 22.5	**15 / 23.7**	23.5 / 33.9	**31.1 / 40.6**

Table 4: Cohen's Kappa and Kendall tau (κ / τ) between the human evaluators and the two GT sets when assessing the models' predictions. Scores are multiplied by 100.

Models	ASQP-Rest15			ASQP-Rest16			ACOS-Laptop			ACOS-Rest		
	Orig.	Ours	Δ	Orig.	Ours	Δ	Orig.	Ours	Δ	Orig.	Ours	Δ
Naive												
Paraphrase	22.55	25.93	3.38	31.27	33.62	2.35	8.97	9.53	0.56	22.91	24.17	1.26
DLO	31.17	35.7	4.53	37.84	42.52	4.68	12.44	13.32	0.88	34.3	35.82	1.52
MvP	33.95	36.97	3.02	42.26	45.05	2.79	16.3	17.5	1.2	30.53	31.93	1.4
GPT-3.5-Turbo	29.45	43.12	13.67	34.84	47.52	12.68	25.92	30.81	4.89	36.78	45.1	8.32
GPT-4o-mini	34.11	48.69	14.58	41.77	54.47	12.7	33.68	39.23	5.55	39.63	47.55	7.92
Gemini-v1.5-Pro	**36.61**	47.23	10.62	**45.46**	56.84	11.38	33.79	36.55	2.76	**42.73**	47.35	4.62
Llama-3.1-70B	35.63	**51.85**	16.22	44.6	**59.62**	15.02	**37.13**	**42.6**	5.47	40.93	**50.96**	10.03
MvP-full (SOTA)	50.49	55.97	5.48	59.31	63.98	4.67	61.44	63.91	2.47	57.05	62.63	5.58
Ensemble												
Paraphrase	24.86	27.85	2.99	34.75	36.46	1.71	8.5	9.07	0.57	24.6	25.91	1.31
DLO	34.58	38.78	4.2	41.31	46.22	4.91	12.52	13.36	0.84	37.6	38.99	1.39
MvP	35.77	38.11	2.34	44.85	47.25	2.4	16.36	17.49	1.13	34.27	35.55	1.28
GPT-3.5-Turbo	35.64	48.9	13.26	43.28	56.73	13.45	30.88	34.39	3.51	43.92	51.65	7.73
GPT-4o-mini	40.36	52.8	12.44	46.59	59.59	13.0	38.02	42.03	4.01	47.55	52.96	5.41
Gemini-v1.5-Pro	**40.88**	51.75	10.87	**49.44**	61.01	11.57	38.07	40.33	2.26	**49.61**	55.23	5.62
Llama-3.1-70B	40.41	**56.74**	16.33	48.99	**63.97**	14.98	**41.06**	**46.18**	5.12	46.0	**55.8**	9.8
MvP-full (SOTA)	51.74	57.03	5.29	59.75	64.15	4.4	62.81	65.28	2.47	59.02	64.67	5.65

Table 5: Comparison of average F1 scores between the original GTs (Orig.) and our expanded sets (Ours), highlighting the discrepancy between them (Δ). T5-based models, *i.e.*, Parapharse (Zhang et al., 2021), DLO (Hu et al., 2022), MvP (Gou et al., 2023), are transfer learned from ASTE to our 20-shot examples. LLMs, specifically GPT-3.5-Turbo, GPT-4o-mini, Gemini-v1.5-Pro, Llama-3.1-70B-Instruct, generate predictions based on the same 20-shot in-context learning. The **best scores** are bolded, and the second-best scores are underlined. The state-of-the-art (SOTA) model, *i.e.*, *full-finetuned MvP (MvP-full)*, is included for reference. Note that the gap between the scores of the SOTA model and others decreases when evaluated on our GT set compared to the original set.

a comprehensive understanding of the ABSA task. The final human validity result was determined by a majority vote. Details of the human evaluation, including the user interface sample, are provided in Appendix B.

As illustrated in Table 3, our human evaluation reveals a high level of validity, with all percentages exceeding 90%. We also calculate the difference in the model F1 score between our test set and the human-filtered test set for the sampled examples. Notably, there is an average 0.53% F1 score difference between the two GT sets, implying the high quality of our new evaluation set. Detailed score is described in Table 11.

4.3 Alignment with Human Evaluator

To demonstrate that evaluations with our expanded GTs align more closely with human judgments than those on the original GTs, we conduct human evaluations on model-predicted outputs and compare the alignment with our and the original GT sets. We randomly sample 80 test examples from each of the four datasets and extract predictions for two models: MvP and GPT-3.5-Turbo, which are consistent with the previous validity experiment. Three human annotators, provided with task descriptions and 20 examples identical to those given to the models, judge whether each prediction is appropriate for a given sentence and task. The final human decision is determined by a majority vote.

Inter-annotator agreement (IAA) is measured between human evaluators and the two GT sets (Origin vs. Humans and Ours vs. Humans) using Cohen's Kappa (κ) (Cohen, 1960) and Kendall Tau (τ) (Kendall, 1938) [3]. Table 4 shows that our new test set achieves higher IAA with human evaluators than the existing test set in most cases, demonstrating that our pipeline is more human-like than eval-

[3]Pearson correlation is omitted due to its identical score to Kendall Tau, given the limited sample size.

uations than the original set. MvP exhibits higher human agreement than other models, likely due to clearer distinctions between correct and incorrect predictions. Also, the Fleiss' kappa[4] among human annotators are 41.2, 44.3, 45.8, and 49.6 for ASQP-R15, R16, ACOS-Laptop, and Rest, respectively. This moderate agreement among annotators indicates a valid evaluation. Based on these results, our expanded GT sets demonstrate improved alignment with human judgments, supporting the effectiveness of our approach in enhancing ABSA evaluation.

4.4 Re-evaluate Model Performance

We evaluate diverse ABSA approaches, comprising three T5-based models and four LLMs, on our expanded GT set using test examples from four datasets: ASQP-R15, R16, ACOS-Laptop, and Rest. Data statistics are described in Table 9. The scores are F1 using the exact match measurement. Since the prediction order of the elements of each quadruple highly matters in accuracy (Hu et al., 2022; Gou et al., 2023), all models predict 24 different orders. Naive scores are summarized in their average values. We report the results using an ensemble alongside the naively predicted results. The prediction in the ensemble is chosen if the prediction appears at least three times in top-5 order or seed. The scores for all orders are in Appendix E.

As shown in Table 5, our experimental results reveal significant performance improvements in ABSA tasks when assessing LLMs using our GT sets. LLM-based models demonstrate an average F1 score improvement of 9.8 percentage points, compared to 2.3 for T5-based models. This substantial disparity suggests that conventional GT sets may have underestimated LLM performance in ABSA tasks due to their limited diversity of surface forms. Notably, Llama-3.1 advances from second to first place in three out of four datasets under our new evaluation method, surpassing the initially top-performing Gemini.

The ensemble approach consistently outperforms naive methods, with Llama-3.1 achieving results within 1%p of the SOTA model on the ASQP dataset. Interestingly, LLM performance varies between the ACOS and ASQP datasets, which differ in handling implicit opinion terms. The ASQP dataset, excluding "null" cases for implicit opinions, allows LLMs to demonstrate particular efficacy in extracting explicit opinion terms. This indicates that for tasks with explicit opinion expressions, LLMs can achieve performance levels comparable to specialized SOTA models. These findings collectively suggest that the potential of LLMs in ABSA tasks may have been previously underestimated, especially in scenarios with well-defined opinion expressions.

		Orig.	ZOOM-IN	ZOOM-OUT	Filter*	Ours
ASQP-R15	A	795	+212	+679	−313	+578
	O	1014	+217	+1480	−704	+993
ASQP-R16	A	799	+203	+732	−346	+589
	O	799	+208	+1034	−437	+805
ACOS-Lap	A	1161	+158	+1117	−477	+798
	O	1161	+163	+1415	−402	+1176
ACOS-Rest	A	916	+97	+791	−300	+588
	O	916	+115	+1393	−614	+894

Table 6: Number of terms added or removed at each step of ZOOM IN-N-OUT for aspect (A) and opinion (O). Our pipeline begins with the original terms (Orig.), generates new terms in ZOOM-IN and ZOOM-OUT, and eliminates terms through JUDGE and rule-based filtering (Filter*), resulting in a new GT set (Ours).

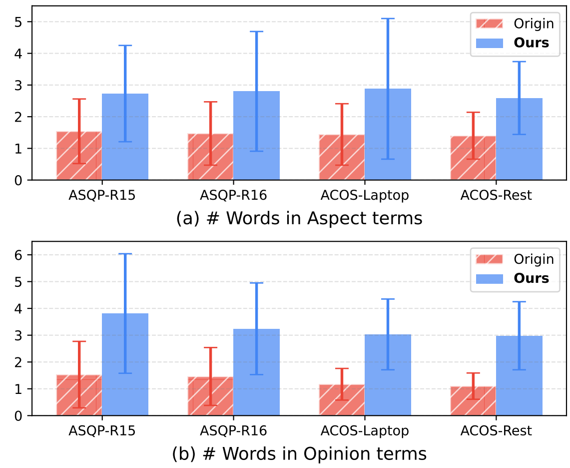

Figure 3: Average and standard deviation of word counts in explicit aspect and opinion terms across the four datasets.

5 Analysis

5.1 Statistics of Our GT set

Changes in each step Table 6 shows the number of terms added or removed at each pipeline step for aspect (A) and opinion (O) terms. Our final datasets contain, on average, 638 more aspect and 967 more opinion terms than the original GT set. ZOOM-OUT generates more terms in both aspect and opinion than ZOOM-IN step. This is natural because ZOOM-OUT extracts terms from anywhere

[4]Fleiss' kappa was used to measure agreement among three annotators, as it is more appropriate than Cohen's kappa for more than two annotators.

in the sentence, while ZOOM-IN is limited to the original GT term.

Number of words in a term We compute the number of words in aspect and opinion terms using the NLTK word tokenizer (Bird and Loper, 2004), excluding implicit cases where the term is “null.” Figure 3 illustrates the mean and standard deviation of the word counts for both the original and our test sets. Our terms exhibit higher word counts in all cases, suggesting a broader range of surface forms than the original dataset. The observed increases in standard deviation further suggest that our dataset selectively adds words where necessary, rather than indiscriminately increasing word count across all terms.

5.2 Ablation Study

We validate the role of ZOOM-IN, ZOOM-OUT, and JUDGE steps in expanding the existing GT set. Following the experimental setup in Section 4.3, we analyze the agreement with human evaluation for each step.

Table 7 shows the number of expanded quadruples for the sampled 80 examples with human ratings, along with the changes in IAA for each step. The ZOOM-IN and ZOOM-OUT steps generally improve IAA, except for the ZOOM-OUT step in ACOS-Laptop. This result implies that each of these processes helps align existing GT with human evaluation criteria. On the other hand, the IAA drop in the ZOOM-OUT process of ACOS-Laptop is attributed to an issue with categorical values, which our extension does not handle. This can be understood as a limitation of our work since our model does not consider the variability of category and sentiment. In the JUDGE process, while the number of quadruples decreases by approximately 20% on average, IAA improves or remains the same, indicating the judging process removes irrelevant quadruples that do not affect human alignment and fairly improves the dataset quality.

5.3 Comparison with SOTA Model

In Table 5, Llama-3.1-70B shows high accuracy, approaching the SOTA performance of MvP in our evaluation set. Especially in the ensemble setting, the difference is merely 0.2%p. To further analyze the capability of Llama, we increase the number of demonstration examples while maintaining the same experimental setup as in Table 5.

As illustrated in Table 8, Llama outperforms

Models	ASQP-R15		ASQP-R16		ACOS-Lap		ACOS-Rest	
	#Q	IAA	#Q	IAA	#Q	IAA	#Q	IAA
Orig.	128	8.3	110	26.3	114	12.8	132	23.5
+ZOOM-IN	215	**11.9**	185	**27.1**	145	**16.1**	166	**25**
+ZOOM-OUT	659	**18**	476	**37**	503	15	609	**31.1**
Ours	546	**18**	402	**48.9**	369	**15**	487	**31.1**

Table 7: Inter-annotator agreement (IAA) measured by Cohen’s kappa (κ) between human annotators and ground truth (GT) sets at each step of processing GPT-3.5-Turbo’s predictions. The number of GT quadruples (#Q) increases during generation steps and decreases after filtering. Scores are **bolded** if they equal or exceed those from the previous step.

Models	#Train	Orig.	Ours	Δ
Naive				
MvP-full (SOTA)	1,264	**59.31**	**63.98**	4.67
Llama-3.1-70B	20	44.6	59.62	15.02
	40	<u>46.33</u>	<u>62.14</u>	15.81
Ensemble				
MvP-full (SOTA)	1,264	**59.75**	<u>64.15</u>	4.4
Llama-3.1-70B	20	48.99	63.97	14.98
	40	<u>50.09</u>	**66.42**	16.33

Table 8: Comparison of full-finetuned MvP, *i.e.*, the SOTA model, and Llama-3.1-70B-Instruct in few-shot ICL setting on ASQP-Rest16 dataset. The scores are averaged F1 exact match scores in 24 orders.

MvP in the 40-shot ensemble setting when evaluated on our GT set. This finding implies that Llama has been apparently underestimated in the conventional single-answer exact match evaluation approach. Our observation aligns with previous research in RE tasks where human evaluation demonstrates the near-SOTA performance of LLMs (Wadhwa et al., 2023). Notably, these prediction results derive from a simple few-shot ICL setting. We expect that applying advanced techniques, such as example selection and chain-of-thoughts prompting, would yield results comparable to or exceeding current SOTA performance. We leave this exploration as our future work.

6 Conclusion

We present ZOOM IN-N-OUT, a novel approach for the ABSA task that addresses the limitations of traditional evaluation methods by accounting for the diversity of surface forms in aspect and opinion terms. By expanding the ground truth sets to include valid variations that preserve original meanings, our method aligns more closely with human judgments. Experimental results show that LLMs significantly outperform T5 models under the proposed evaluation scheme, suggesting that LLMs' capabilities in span extraction tasks have been underestimated due to restrictive evaluation practices. This underscores the importance of considering the diversity of expression in the LLM era for fair and accurate assessments. While our study focuses on ABSA, the principles underlying ZOOM IN-N-OUT are applicable to other NLP tasks involving span extraction, such as named entity recognition, relation extraction, and event extraction. Owing to the flexibility of our framework, we anticipate researchers in NLP tasks, especially for information extraction tasks such as named entity recognition or relation extraction, which include span extraction tasks.

Limitation

Our study performed the quadruple prediction task between the various tasks of ABSA, which included tuple or triplet predictions. However, such tasks exist as part of quadruples, so there is plenty of room for our approach to be utilized. Also, only explicit mention in the sentence is covered in our experiments. This is because the implicit cases tagged in "null" make it difficult to extract diverse surface forms where the meaning of "null" is hidden in the sentence.

Ethics Statement

All human evaluators voluntarily agreed to participate in this study. We informed them that their responses would be anonymized and securely protected. Evaluators were free to stop the survey at any time. Compensation was more than adequate. As our survey and dataset do not contain any sensitive or harmful content, we do not anticipate any negative ethical impacts from our research. Also, we utilize ChatGPT and Claude as a writing assistant and Copilot as a coding assistant.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)), Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea) & Gwangju Metropolitan City, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913). Lastly, we sincerely appreciate to our DAVIAN Lab people, sharing insightful feedback and discussions.

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*.

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 *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Hongjie Cai, Nan Song, Zengzhi Wang, Qiming Xie, Qiankun Zhao, Ke Li, Siwei Wu, Shijie Liu, Jianfei Yu, and Rui Xia. 2023. Memd-absa: A multi-element multi-domain dataset for aspect-based sentiment analysis.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling.

In *Proc. of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proc. of the 29th International Conference on Computational Linguistics*.

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

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proc. of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

LlamaTeam. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. Star: Boosting low-resource information extraction by structure-to-text data generation with large language models. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proc. of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. 2023. Guideline learning for in-context information extraction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*.

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)*.

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)*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 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. *The Journal of Machine Learning Research*, 21(1).

José Saias. 2015. Sentiue: Target and aspect based sentiment analysis in SemEval-2015 task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023a. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xingyao Wang, Sha Li, and Heng Ji. 2023b. Code4Struct: Code generation for few-shot event structure prediction. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zengzhi Wang, Rui Xia, and Jianfei Yu. 2024. Unified absa via annotation-decoupled multi-task instruction tuning. *IEEE Transactions on Knowledge and Data Engineering*.

Chenxiao Wu, Wenjun Ke, Peng Wang, Zhizhao Luo, Guozheng Li, and Wanyi Chen. 2024. Consistner: Towards instructive ner demonstrations for llms with the consistency of ontology and context. *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*, 38.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Self-improving for zero-shot named entity recognition with large language models. *Preprint*, arXiv:2311.08921.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? In *Proc. of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Appendix

A Experimental Details

A.1 Dataset Statistics

Table 9 shows statistics of dataset utilized in our experiments, *i.e.*, ASQP (Zhang et al., 2021) and ACOS (Cai et al., 2021).

Datasets	# Train	# Dev.	# Test
ASQP-R15	834	209	537
ASQP-R16	1,264	316	544
ACOS-Laptop	2,934	326	816
ACOS-Rest	1,530	171	583

Table 9: Statistics for each dataset. The number of examples for train, dev, and test sets.

A.2 Implementation Details

For Llama models, we utilize NVIDIA H100 80GB HBM3. To fine-tune T5 models, we used NVIDIA GeForce RTX 3090 and NVIDIA A100-PCIE-40GB. Table 10 shows the total cost to construct our GT set. The dataset was constructed in May 2024.

Dataset	ASQP-R15		ASQP-R16		ACOS-L		ACOS-R	
	A	O	A	O	A	O	A	O
ZOOM-IN	5.6	7.1	5.9	6.9	8.6	7.9	6.6	6.9
ZOOM-OUT	7.4	9.7	7.6	9.6	10.3	8.8	7.7	7.6
JUDGE	16.3	30.4	16.3	21.5	21.7	27.2	16.1	26.4
Sum. element	29.3	47.2	29.8	38	40.6	43.9	30.4	40.9
Sum. dataset	76.5		67.8		84.5		71.3	

Table 10: Total cost ($) to construct expanded GT sets for each step, element (aspect (A) and opinion (O)), and dataset.

B Details of Human Evaluation

In this section, we describe the full results and the experimental details of Section 4.2 and Section 4.3, which include human annotations. Three authors volunteered to participate in this study. We used the four ASQP datasets and randomly selected 80 samples in test sets where the seed was 42.

B.1 Dataset Validity

In this experiment, we asked human annotators to validate whether the new GT quadruple is correct with the given sentence. The user interface is shown in Figure 4. Also, Table 11 describes the model performance difference between our GT set and the human annotator's oracle GT.

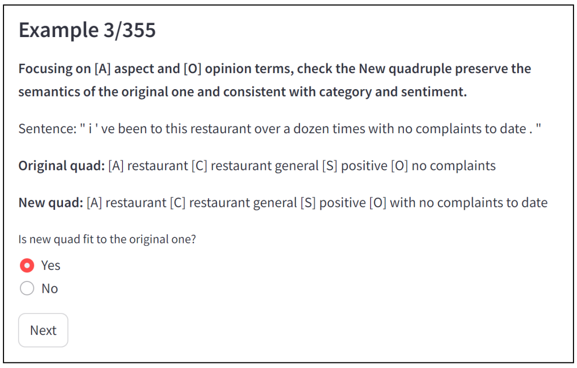

Figure 4: Dataset validity study UI example on ACOS-Rest dataset in Section 4.2.

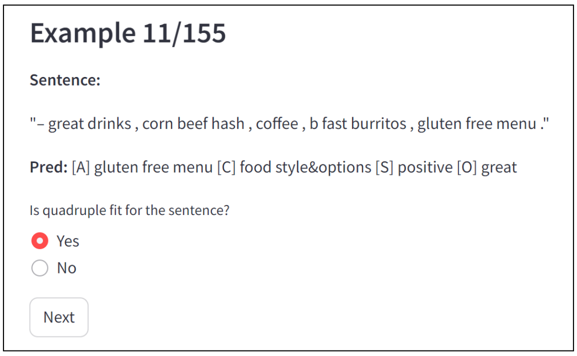

Figure 5: Human evaluation study UI example on ACOS-Rest dataset.

B.2 Human evaluation

Human annotators evaluated the predicted quadruples from two models: MvP and GPT-3.5-Turbo. The annotators were provided with a sentence and a predicted quadruple and asked to tag whether the predicted quadruple was correct or not. The user interface is shown in Figure 5.

C Further Analysis

C.1 Examples of New GT set

Table 12 illustrates the aspect and opinion terms deemed acceptable in true positive cases within our evaluation set. In the first example, the prediction that "sake" is "successfully easing" is assessed as correct in our revised ground truth (GT) set, whereas the original GT would have classified it as incorrect. This example demonstrates that the newly accepted term aligns semantically with the original annotation, justifying its inclusion as a valid prediction.

Datasets / Models	ASQP-R15			ASQP-R16			ACOS-Lap			ACOS-Rest		
	Ours	Oracle	Diff	Ours	Oracle	Diff	Ours	Oracle	Diff	Ours	Oracle	Diff
MvP-full	39.25	38.49	0.76	46.96	46.96	0	19.82	19.82	0	25.28	25.28	0
GPT-3.5-Turbo	39.27	37.82	1.45	41.43	41.43	0	31.02	30.2	0.82	39.02	38.33	0.69
GPT-4o-mini	46.32	44.85	1.47	59.41	59.41	0	39.15	38.3	0.85	44.12	42.65	1.47
Gemini-pro	49.06	49.06	0	62.07	62.07	0	45.81	45.81	0	43.08	42.31	0.77
Llama-3.1-70B-Instruct	49.25	47.76	1.49	57.63	57.63	0	39.82	39.82	0	46.87	46.09	0.78
Average	44.63	43.60	1.03	53.50	53.50	0.00	35.12	34.79	0.33	39.67	38.93	0.74

Table 11: Difference in the model F1 score (%) between our test set (**Ours**) and the human-filtered test set (**Oracle**) for the sampled examples.

C.2 Ablation study

As described in Section 5.2, we measure IAA between humans and variants of our GT sets for each step. Table 13 shows the full result.

D Prompts

D.1 System Prompt

Here, we show the full system prompts for our dataset expansion pipeline and model prediction. For dataset construction, three steps are done independently for each aspect and opinion term: Table 14 for ZOOM-IN step, Table 15 for ZOOM-OUT step, Table 16 for JUDGE step in aspect term, Table 17 for JUDGE step in opinion term. On the other hand, to get the prediction of diverse LLMs, we input the same 20-shot, as shown in Table 19.

D.2 Examples in ZOOM IN-N-OUT

In ZOOM-IN and ZOOM-OUT, 5-shot ICL is used. The example is described in Table 18. In JUDGE step, to exercise caution in judgment, we utilize 5-shot CoT prompting. Table 20 shows the demonstration example, and Table 21 is the real decision of LLM classifying the validness of the new term.

D.3 Demonstration Examples

20 examples are selected by heuristics considering the balance of implicit cases, category, and sentiment. Table 22 shows the selected examples for the ACOS-Rest dataset.

E LLMs Performance

Full experimental results in Table 5 are shown in three tables for three different LLMs as follows: Table 23 for GPT-3.5-Turbo, Table 24 for GPT-4o-mini, Table 25 for Gemini-v1.5-Pro, and Table 26 for Llama-3.1-70B-Instruct.

Original	→	Ours	Generated Step
Sentence: the sake ' s complimented the courses very well and is successfully easing me into the sake world .			
[A] sake ' s **[O]** successfully	→	**[A]** sake **[O]** successfully easing	**[A]** ZOOM-IN **[O]** ZOOM-OUT
Sentence: the lemon chicken tasted like sticky sweet donuts and the honey walnut prawns , the few they actually give you ... were not good .			
[A] lemon chicken **[O]** sticky sweet	→	**[A]** lemon chicken **[O]** tasted like sticky sweet donuts	**[A]** ORIGINAL **[O]** ZOOM-OUT
Sentence: thus far , i ' ve loaded a number of games , my password vault , several productivity apps , skype , spotify and some network scanning android apps and they ' ve all worked very well .			
[A] android apps **[O]** well	→	**[A]** network scanning android apps **[O]** very well	**[A]** ZOOM-OUT **[O]** ZOOM-OUT
Sentence: the side of potatoes is to die for , as is the labne (yogurt dip) .			
[A] labne (yogurt dip) **[O]** die for	→	**[A]** labne **[O]** die for	**[A]** ZOOM-IN **[O]** ORIGINAL

Table 12: Four examples have been included as true positives in our GT set. Each row contains examples sourced from the ASQP-R15, ASQP-R16, ACOS-Laptop, and ACOS-Rest datasets. The generated step for each aspect and opinion term is explained separately.

	ASQP-R15			ASQP-R16			ACOS-Lap			ACOS-Rest		
	#GT	MvP	GPT	#GT	MvP	GPT	#GT	MVP	GPT	#GT	MvP	GPT
Orig.	128	33.4/42.3	8.3/18.8	110	56.6/61.8	26.3/38.9	114	50.5/58.1	12.8/22.5	132	47.3/54.5	23.5/33.9
+ZOOM-IN	215	**39.5/47.4**	**11.9/23.4**	185	**60.9/65.3**	**27.1/39.6**	145	50.5/58.1	**16.1/26.2**	166	47.3/54.5	**25/35.3**
+ZOOM-OUT	659	**40.6/48.2**	**18/29.8**	476	60.9/65.3	**37/46.5**	503	**52.7/59.8**	15/23.7	609	47.3/54.5	**31.1/40.6**
Ours	546	39.5/47.4	18/29.8	402	59.5/64.1	**38.7/48.9**	369	52.7/59.8	15/23.7	487	47.3/54.5	31.1/40.6

Table 13: Agreement scores of Cohen's kappa and Kendal tau (κ / τ). Each row shows the accumulation of each module. The models are MvP and GPT-3.5-Turbo, respectively. Scores are **bolded** when the score increases to the previous step.

Given an input sentence, Sentiment, Opinion, Categroy, and **target Aspect terms**, extract expressions that narrow the span of the aspect term. The new expressions must be confined within the original aspect term and adhere to the following criteria:

1. Remain relevant to the given aspect term without altering its original meaning.
2. Exclude any unnecessary words or spaces.
3. Correct any typos if present and resolve contraction if present.
4. Revert to the original expression if narrowing proves challenging.
5. Ensure the expression exists exactly as it appears in the given sentence.
6. Keep the aspect term and opinion term distinct and independent.

Given an input sentence, Category, Aspect, Sentiment, and **target Opinion terms**, extract expressions that narrow the span of the opinion term. The new expressions must be confined within the original opinion term and adhere to the following criteria:

1. Be related to both the aspect and opinion terms. Maintain the original sentiment polarity; changes in sentiment by narrowing the span are not allowed.
2. Correct any typos and resolve contractions if present.
3. Exclude any unnecessary words or spaces.
4. Return the original expression if reducing it proves difficult.
5. Ensure the expression exists verbatim in the given sentence.
6. Keep the aspect and opinion terms distinct and independent.

Table 14: ZOOM-IN step system prompts for Aspect and Opinion, respectively.

Given an input sentence, Sentiment, Opinion, Categroy, and **target Aspect terms**, extract various expressions that extend the span of the aspect term. The new expressions should be formed by adding surrounding words and must meet the following criteria:

1. Be related to the given aspect term.
2. Include neighboring words around the aspect term to form a new expression.
3. Should not overlap with the original opinion term [O].
4. While expanding the aspect term, avoid incorporating the entire sentence.
5. If it is not feasible to expand the expression, return only the original aspect term.
6. Ensure the new expression matches exactly as it appears in the input sentence.

Given an input sentence, Category, Aspect, Sentiment, and **target Opinion terms**, extract various expressions that extend the span of the opinion term. The new expressions should adhere to the following criteria:

1. Be related to the given opinion term.
2. Should not overlap with the original aspect term [A].
3. Include neighboring words around the opinion term to form a new expression, while ensuring the new expression does not encompass the entire sentence.
4. If it is challenging to expand the expression, return only the original opinion term.
5. Ensure the sentiment polarity remains consistent; expanding the expression should not alter the given sentiment polarity.

Table 15: ZOOM-OUT step system prompts for Aspect and Opinion, respectively.

You are tasked with assessing whether a newly created aspect term aligns with a given Ground Truth (GT) quadruple in aspect-based sentiment analysis (ABSA). Here's how to do it:

1. Review the provided sentence and the GT quadruple, which includes:
   - Aspect Term (A): The specific word or phrase referring to an aspect in the sentence.
   - Opinion Term (O): The word or phrase expressing an opinion about the aspect.
   - Aspect Category (C): The category to which the aspect term belongs. Categories include:
     – Location General
     – Food Prices
     – Food Quality
     – Food General
     – Food Style&Options
     – Ambience General
     – Service General
     – Restaurant General
     – Restaurant Prices
     – Restaurant Miscellaneous
     – Drinks Prices
     – Drinks Quality
     – Drinks Style&Options
   - Sentiment Polarity (S): The sentiment associated with the opinion, chosen from:
     – Positive
     – Neutral
     – Negative

2. Determine the alignment based on the following criteria:
   1. Aspect and Category Consistency:
      - The new aspect term must maintain the target object of the [A] aspect and the [C] category in the GT.
   2. Sentiment and Opinion Relevance:
      - The new aspect term must directly relate to the [S] sentiment and [O] opinion as the GT.
   3. Extractability:
      - The new aspect term must be directly taken from the sentence without adding new words or significantly rearranging existing ones. - Minor adjustments like unwinding contractions or fixing typos are allowed.
   4. Independency:
      - Each aspect and opinion term must be independent and not overlap.
      - The new aspect term must not contain the GT [O] opinion term.

3. Determining Validity:
   - If all criteria are met, the new term is "valid."
   - If any criterion is not met, the new term is "invalid."

4. Providing Feedback:
   - Explain why a term was deemed valid or invalid based on the above criteria.
   - Specific feedback helps in understanding the decision.

Table 16: System prompt of Judge step for Aspect terms, especially on ACOS-Rest dataset.

You are tasked with assessing whether a newly created opinion term aligns with a given Ground Truth (GT) quadruple in aspect-based sentiment analysis (ABSA). Here's how to do it:

1. Review the provided sentence and the GT quadruple, which includes:
    - Aspect Term (A): The specific word or phrase referring to an aspect in the sentence.
    - Opinion Term (O): The word or phrase expressing an opinion about the aspect.
    - Aspect Category (C): The category to which the aspect term belongs. Categories include:
        – Location General
        – Food Prices
        – Food Quality
        – Food General
        – Food Style&Options
        – Ambience General
        – Service General
        – Restaurant General
        – Restaurant Prices
        – Restaurant Miscellaneous
        – Drinks Prices
        – Drinks Quality
        – Drinks Style&Options
    - Sentiment Polarity (S): The sentiment associated with the opinion, chosen from:
        – Positive
        – Neutral
        – Negative
2. Determine the alignment based on the following criteria:
    1. Aspect and Category Relevance:
        - The new opinion term must directly relate to the [A] aspect and the [C] category in the GT.
    2. Sentiment and Opinion Consistency:
        - The new opinion term should maintain the same [S] sentiment polarity and [O] opinion as the GT.
    3. Extractability:
        - The new opinion term must be directly taken from the sentence without adding new words or significantly rearranging existing ones.
        - Minor adjustments like unwinding contractions or fixing typos are allowed.
    4. Independency:
        - Each aspect and opinion term must be independent and not overlap.
        - The new opinion term must not contain the GT [A] aspect term.
3. Determining Validity:
    - If all criteria are met, the new term is "valid."
    - If any criterion is not met, the new term is "invalid."
4. Providing Feedback:
    - Explain why a term was deemed valid or invalid based on the above criteria.
    - Specific feedback helps in understanding the decision.

Table 17: System prompt of Judge step for Opinion terms, especially on ACOS-Rest dataset.

System Prompt
Given an input sentence, Category, Aspect, Sentiment, and **target Opinion terms**, extract expressions that narrow the span of the opinion term. The new expressions must be confined within the original opinion term and adhere to the following criteria:

1. Be related to both the aspect and opinion terms. Maintain the original sentiment polarity; changes in sentiment by narrowing the span are not allowed.
2. Correct any typos and resolve contractions if present.
3. Exclude any unnecessary words or spaces.
4. Return the original expression if reducing it proves difficult.
5. Ensure the expression exists verbatim in the given sentence.
6. Keep the aspect and opinion terms distinct and independent.

Demonstration
Input sentence: "the pizza was delivered cold and the cheese was n ́t even fully melted !"

- Category term: "food quality"
- Aspect term: "cheese"
- Sentiment term: "negative"
- Target Opinion term: "was n ́t even fully melted"

Test Sample
Input sentence:

System Prompt
Given an input sentence, Sentiment, Opinion, Categroy, and **target Aspect terms**, extract various expressions that extend the span of the aspect term. The new expressions should be formed by adding surrounding words and must meet the following criteria:

1. Be related to the given aspect term.
2. Include neighboring words around the aspect term to form a new expression.
3. Should not overlap with the original opinion term [O].
4. While expanding the aspect term, avoid incorporating the entire sentence.
5. If it is not feasible to expand the expression, return only the original aspect term.
6. Ensure the new expression matches exactly as it appears in the input sentence.

Demonstration
Input sentence: "quacamole at pacifico is yummy , as are the wings with chimmichuri ."

- Sentiment term: "positive"
- Opinion term: "yummy"
- Category term: "food quality"
- Target Aspect term: "quacamole"

Test Sample
Input sentence:

Table 18: Full prompt for Opinion terms in ZOOM-IN and Aspect terms in ZOOM-OUT, respectively.

Aspect-based sentiment analysis aims to identify the aspects of given target entities and the sentiment expressed towards each aspect.
For example, from an example sentence: “This restaurant is rude, but the food is delicious”, we can extract the negative sentiment that the restaurant is (1) “rude” in terms of “service general” and (2) “delicious” in terms of “food quality”. As such, the complex task of categorizing the aspect terms and their corresponding categories and the sentiment expressed for the aspect in the sentence into one of three classes [positive, negative, neutral] is the Aspect-based sentiment analysis (ABSA) task.
Each element that is extracted is called an element, and the characteristics of each element can be described as follows.

1. Aspect: The aspect covered by the sentence, such as restaurant, food name, or service.
   - Any phrase, verb, or noun that mentions a particular aspect can be an aspect.
   - Aspects can be extracted with or without quotation marks.
   - Determiners are excluded unless they are part of a noun phrase.
   - Subjectivity indicators that indicate opinion are not included.
   - Specific product names are not aspect terms.
   - Even if pronouns refer to an aspect, they are not aspect terms.
   - Pronouns (e.g., “it”, “they”, “this”) even if they refer to an aspect.
   - If they appear in the sentence, we extract their span as an aspect; if they do not appear directly in the sentence, we define ‘null’ as the aspect term.
2. Category: Predefined categories to categorize aspects. Categories are divided into two levels: six entity categories and five attribute labels, each characterized by the following features:
   1) Entity
      a. FOOD for opinions focusing on the food in general or in terms of specific dishes, dining options etc.
      b. DRINKS for opinions focusing on the drinks in general or in terms of specific drinks, drinking options etc.
      c. SERVICE for opinions focusing on the (customer/kitchen/counter) service, on the promptness and quality of the restaurant’s service in general, the food preparation, the staff’s attitude and professionalism, the wait time, the options offered (e.g. takeout), etc
      d. AMBIENCE for opinions focusing on the atmosphere or the environment of the restaurant’s interior or exterior space (e.g. terrace, yard, garden), the décor, entertainment options, etc.
      e. LOCATION for opinions focusing on the location of the reviewed restaurant in terms of its position, the surroundings, the view, etc.
      f. RESTAURANT for opinions expressed about the (specific) evaluated restaurant as whole not focusing on any of the above five entity types.
   2) Attribute
      a. GENERAL. This attribute label is assigned to sentences that express general positive or negative sentiment about an entity type.
      b. PRICES for opinions that refer to the prices of the food, the drinks or the restaurant in general. e.g.
      c. QUALITY for opinions focusing on the taste, the freshness, the texture, the consistency, the temperature, the preparation, the authenticity, the cooking or general quality of the food and the drinks served in the restaurant.
      d . STYLE&OPTIONS for opinions referring to the presentation, the serving style, the portions size, the food/menu options or variety (e.g. innovative dishes/drinks, vegetarian options) of the food and of the drinks served in the restaurant.
      e. MISCELLANEOUS for attributes that do not fall into any of the aforementioned cases.
   3) Final Entity Attribute pairs We end up with 13 Entity Attribute category pairs: [‘location general’, ‘food prices’, ‘food quality’, ‘food general’, ‘food style&options’, ‘ambience general’, ‘service general’, ‘restaurant general’, ‘restaurant prices’, ‘restaurant miscellaneous’, ‘drinks prices’, ‘drinks quality’, ‘drinks style&options’].
3. Opinion: An opinion term that expresses a sentiment about an aspect. If it appears directly in the sentence as a single word or phrase, we extract it. However, if no specific phrase can be extracted, and the sentiment about the aspect can be gleaned from the nuances of the sentence as a whole, we define ‘null’ as the opinion term.
4. Sentiment: The sentiment expressed by the customer about an aspect, divided into three classes: [Positive, Neutral, Negative]. The neutral label applies for mildly positive or negative sentiment, thus it does not indicate objectivity

To summarize, we want to extract one or more quadruples of (aspect, category, opinion, sentiment) from a given review. As mentioned before, the aspect term and opinion term can be extracted as ‘null’ if they are not evident in the sentence, while category and sentiment must be selected from the predefined classes “2. Category” and “4. Sentiment”.
It is up to the model to decide in which order to predict each element of the quadruple. The model is given 20 examples, as shown below, and the sentence you want to test.
Place the aspect term after [A], the opinion term after [O], the sentiment polarity after [S], and the aspect category after [C] in the format ‘ORDER’. If multiple quadruples are predicted, insert ‘####’ to separate the quadruples.

Table 19: System prompt of ACOS-Rest datasets, which LLMs utilize to make the prediction, in Section 4.4. ‘ORDER’ is a hyperparameter that guides the model in a specific element order. The variants of score according to order are described in Table 23, and Table 25.

User:

- Input Sentence: the pizza was delivered cold and the cheese was n' t even fully melted !
- GT: ['cheese', 'food quality', 'negative', 'n' t even fully melted'],
- New Opinion Term: even fully melted

AI:

The input sentence details a quality issue with the pizza, explicitly noting that the [A] "cheese" was not adequately melted, articulated through the opinion [O] "n ' t even fully melted.' The new opinion term under evaluation is "even fully melted." Let's consider the criteria step by step:

1. Aspect and Category Relevance:
   - Reasoning: The new opinion term "even fully melted" is derived from the original opinion term [O] "n ' t even fully melted" related to [A] "cheese" and pertains directly to [C] "food quality," focusing specifically on the melting quality of the cheese. The new term aligns with the original aspect and category.
   - Decision: True
2. Sentiment and Opinion Consistency:
   - Reasoning: The new term "even fully melted" captures part of the original opinion term [O] "n t́ even fully melted." However, it omits the critical negation "n ' t," which is essential to convey the negative sentiment [S] "negative" associated with the GT. Without this negation, the sentiment changes to positive. Thus, the new term does not maintain the sentiment polarity.
   - Decision: False
3. Extractability:
   - Reasoning: "Even fully melted" is a phrase directly extracted from the end of the sentence without adding or significantly altering the words.
   - Decision: True
4. Independence:
   - Reasoning: The new opinion term is "even fully melted," and the aspect term is [A] "cheese." The new term and aspect term are independent, with no overlap.
   - Decision: True

Judgment: Since the new term "even fully melted" fails to meet the Sentiment and Opinion Consistency criterion, it is deemed invalid.]

Table 20: Demonstration example of JUDGE.

User:
- Input sentence: great place to relax and enjoy your dinner
- GT: [A] place [C] ambience general [S] positive [O] great
- New Aspect term: great place

AI:
The input sentence describes a “great place to relax and enjoy your dinner.” The new aspect term under evaluation is “great place”. Let’s consider the criteria step by step:

1. Aspect and Category Consistency:
   - Reasoning: The original GT aspect term [A] "place" falls under the category [C] "ambience general." The new term "great place" still refers to the same aspect of the place, maintaining the focus on the ambiance. The new term aligns with the original aspect and category.
   - Decision: True
2. Sentiment and Opinion Relevance:
   - Reasoning: The original opinion term [O] “great” describes the positive sentiment towards the “place.” The new term “great place” combines the aspect and opinion into one phrase, which is not allowed as it should be independent. The sentiment and opinion relevance are not maintained correctly.
   - Decision: False
3. Extractability
   - Reasoning: The new term “great place” is directly taken from the sentence without adding new words or significantly rearranging existing ones. However, it combines the aspect and opinion into one phrase, which is not allowed.
   - Decision: False
4. Independence
   - Reasoning: The new aspect term “great place” incorporates the opinion term [O] “great,” which violates the independence criterion. The aspect term should not include the opinion term.
   - Decision: False

Judgment: Since the new term “great place” fails to meet the criteria for Sentiment and Opinion Relevance, Extractability, and Independence, it is deemed invalid.

User:
- Input sentence: i ordered one , the touch pad failed to work consistently .
- GT: [A] touch pad [C] hardware [S] negative [O] failed
- New Opinon Term: failed to work

AI:
The input sentence mentions that the touch pad failed to work consistently. The new opinion term under evaluation is “failed to work”. Let’s consider the criteria step by step:

1. Aspect and Category Consistency:
   - Reasoning: The original GT opinion term is [O] “failed,” which describes the [A] “touch pad” in the [C] “hardware” category. The new term "failed to work" still describes the touch pad’s functionality, aligning with the original aspect and category.
   - Decision: True
2. Sentiment and Opinion Relevance:
   - Reasoning: The original GT opinion term is [O] “failed,” which conveys a negative sentiment. The new term “failed to work” maintains the same negative sentiment and expands on the original opinion by specifying the failure to work. The sentiment and opinion consistency are preserved.
   - Decision: True
3. Extractability
   - Reasoning: The new term “failed to work” can be directly extracted from the sentence without adding new words or significantly rearranging existing ones.
   - Decision: True
4. Independence
   - Reasoning: The new opinion term “failed to work” and the GT aspect term [A] “touch pad” are independent of each other. The opinion term does not include any part of the aspect term, ensuring no overlap.
   - Decision: True

Judgment: The new opinion term “failed to work” meets all the criteria and is deemed valid.

Table 21: Output example of JUDGE: invalid aspect and valid opinion terms.

1. **User:** although the tables may be closely situated , the candle - light , food - quality and service overcompensate .
   AI: [A] tables [C] ambience general [S] negative [O] closely situated #### [A] candle - light [C] ambience general [S] positive [O] overcompensate #### [A] food [C] food quality [S] positive [O] overcompensate #### [A] service [C] service general [S] positive [O] overcompensate
2. **User:** these are overpriced and you can get better just around the corner :
   AI: [A] null [C] food prices [S] negative [O] overpriced #### [A] null [C] food quality [S] negative [O] null
3. **User:** they have authentic indian at amazing prices .
   AI: [A] indian [C] food quality [S] positive [O] authentic #### [A] null [C] food prices [S] positive [O] amazing
4. **User:** prices too high for this cramped and unappealing resturant .
   AI: [A] resturant [C] restaurant prices [S] negative [O] high #### [A] resturant [C] ambience general [S] negative [O] cramped #### [A] resturant [C] ambience general [S] negative [O] unappealing
5. **User:** do n ' t get me started on the margaritas , either .
   AI: [A] margaritas [C] drinks quality [S] negative [O] null
6. **User:** $ 6 and there is much tasty food , all of it fresh and continually refilled .
   AI: [A] food [C] food style&options [S] positive [O] refilled #### [A] food [C] food quality [S] positive [O] tasty #### [A] food [C] food quality [S] positive [O] fresh #### [A] food [C] food prices [S] positive [O] null
7. **User:** located at the end of a magnificent block .
   AI: [A] null [C] location general [S] positive [O] magnificent
8. **User:** the staff has been nice , but they seemed really stressed and the unisex bathroom needs to be cleaned more often .
   AI: [A] staff [C] service general [S] positive [O] nice #### [A] staff [C] service general [S] positive [O] stressed #### [A] unisex bathroom [C] ambience general [S] negative [O] null
9. **User:** and amazingly cheap .
   AI: [A] null [C] food prices [S] positive [O] amazingly
10. **User:** cheese plate is a varied delight and great bargain at $ 10 .
    AI: [A] cheese plate [C] food quality [S] positive [O] varied delight #### [A] cheese plate [C] food style&options [S] positive [O] varied delight #### [A] cheese plate [C] food prices [S] positive [O] great bargain
11. **User:** not what i would expect for the price and prestige of this location .
    AI: [A] location [C] restaurant prices [S] neutral [O] null #### [A] location [C] restaurant miscellaneous [S] neutral [O] null #### [A] null [C] service general [S] negative [O] null
12. **User:** terrible waste of money . . scammers
    AI: [A] null [C] restaurant general [S] negative [O] scammers #### [A] null [C] restaurant prices [S] negative [O] terrible
13. **User:** small servings for main entree , i had salmon (wasnt impressed) girlfriend had chicken , it was good .
    AI: [A] salmon [C] food quality [S] negative [O] wasnt impressed #### [A] chicken [C] food quality [S] positive [O] good #### [A] servings for main entree [C] food general [S] negative [O] small
14. **User:** first it took us a long time to find the place .
    AI: [A] place [C] restaurant miscellaneous [S] negative [O] null
15. **User:** they were such a rip - off ($ 8 . 95 for four small meat patties in steamed buns) and not worth trying .
    AI: [A] null [C] food quality [S] negative [O] not worth #### [A] null [C] food style&options [S] negative [O] small #### [A] null [C] food prices [S] negative [O] rip - off
16. **User:** restaurant with a view
    AI: [A] view [C] location general [S] neutral [O] null
17. **User:** the lobster sandwich is $ 24 and although it was good it was not nearly enough to warrant that price .
    AI: [A] lobster sandwich [C] food quality [S] positive [O] good #### [A] lobster sandwich [C] food prices [S] negative [O] not nearly enough
18. **User:** my husbands birthday and my sons was not as it was intended . . . and we drove two hours to spend too much money to be treated terribly !
    AI: [A] null [C] restaurant general [S] negative [O] null #### [A] null [C] restaurant prices [S] negative [O] null #### [A] null [C] service general [S] negative [O] terribly
19. **User:** the bar is very well stocked with interesting beers and well priced wines .
    AI: [A] bar [C] drinks style&options [S] positive [O] well stocked #### [A] beers [C] drinks style&options [S] positive [O] interesting #### [A] wines [C] drinks prices [S] positive [O] well priced
20. **User:** the decor however seems to be the distraction so you wo n ' t notice that you just payed 300 bucks for some cold eggplant that took 2 frickin hours to come ! ! ! !
    AI: [A] decor [C] ambience general [S] neutral [O] distraction #### [A] eggplant [C] food quality [S] negative [O] cold #### [A] eggplant [C] food prices [S] negative [O] null #### [A] null [C] service general [S] negative [O] null

Table 22: 20-shot examples of ACOS-Rest dataset utilized in Section 4.4. "####" is a separate token that distinguish the quadruples.

GPT-3.5	ASQP						ACOS					
	Rest15			Rest16			Laptop			Rest		
Orders	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**
AOSC	28.85	42.87	14.02	37.81	50.53	12.72	25.66	30.5	4.84	42.34	50.18	7.84
OCSA	32.44	43.88	11.44	31.61	41.17	9.56	27.35	31.18	3.83	36.58	44.31	7.73
OSAC	32.89	43.7	10.81	35.97	48.18	12.21	28.18	32.19	4.01	41.32	49.21	7.89
OSCA	32.09	44.01	11.92	34.62	47.25	12.63	25.17	28.03	2.86	34.61	41.46	6.85
OACS	32.27	44.11	11.84	36.61	48.45	11.84	29.06	32.91	3.85	39.96	49.0	9.04
AOCS	29.95	45.5	15.55	35.36	48.62	13.26	23.39	28.55	5.16	37.86	45.78	7.92
COAS	32.48	44.48	12.0	35.31	46.48	11.17	24.94	28.91	3.97	35.95	44.96	9.01
SAOC	29.1	43.59	14.49	36.54	50.31	13.77	25.22	31.22	6.0	40.78	49.5	8.72
OASC	**34.33**	44.72	10.39	**39.64**	**51.22**	11.58	26.69	30.52	3.83	**43.4**	**51.74**	8.34
SOAC	30.49	42.46	11.97	36.26	48.57	12.31	23.7	27.58	3.88	39.62	48.16	8.54
SOCA	31.19	44.52	13.33	30.98	40.95	9.97	**40.13**	**48.57**	**8.44**	38.14	45.67	7.53
ASOC	28.81	43.68	14.87	36.79	50.11	13.32	24.41	29.4	4.99	38.43	46.71	8.28
CAOS	28.23	44.09	15.86	36.91	49.66	12.75	22.89	29.3	6.41	36.94	45.32	8.38
SCAO	26.55	43.03	16.48	34.41	49.26	14.85	26.86	33.37	6.51	35.09	43.91	8.82
OCAS	33.95	**46.83**	12.88	36.14	47.43	11.29	30.61	34.01	3.4	39.83	48.78	8.95
COSA	27.85	40.04	12.19	33.68	44.25	10.57	24.62	27.93	3.31	29.69	36.79	7.1
CASO	26.92	41.05	14.13	33.65	48.7	**15.05**	24.12	28.9	4.78	33.89	41.86	7.97
CSAO	26.34	43.33	**16.99**	33.59	48.42	14.83	25.29	31.75	6.46	36.33	45.76	**9.43**
ACOS	26.06	41.74	15.68	33.32	46.4	13.08	24.39	30.73	6.34	35.88	44.14	8.26
ACSO	25.79	40.7	14.91	32.03	46.24	14.21	23.19	27.89	4.7	31.85	41.25	9.4
SCOA	28.95	42.2	13.25	33.7	46.15	12.45	23.25	27.59	4.34	34.28	42.93	8.65
CSOA	28.84	40.97	12.13	34.21	46.73	12.52	23.2	27.79	4.59	34.22	42.95	8.73
SACO	26.49	41.74	15.25	34.29	47.75	13.46	26.43	31.76	5.33	33.63	42.14	8.51
ASCO	25.91	41.71	15.8	32.84	47.61	14.77	23.31	28.76	5.45	31.99	39.94	7.95
Avg.	29.45	43.12	13.67	34.84	47.52	12.67	25.92	30.81	4.89	36.78	45.1	8.33
Std.	2.71	1.63	1.91	2.06	2.57	1.51	3.64	4.25	1.31	3.5	3.63	0.66

Table 23: F1 scores of GPT-3.5-Turbo. The **best score** for each column is bold.

GPT-4o-mini	ASQP						ACOS					
	Rest15			Rest16			Laptop			Rest		
Orders	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**
AOSC	33.28	48.91	15.63	42.54	54.59	12.05	32.3	37.45	5.15	37.09	44.72	7.63
OCSA	**38.45**	49.89	11.44	43.41	54.31	10.9	34.98	39.42	4.44	42.16	50.08	7.92
OSAC	37.25	49.86	12.61	43.7	54.68	10.98	35.39	39.93	4.54	41.99	49.69	7.7
OSCA	36.12	48.01	11.89	42.86	53.46	10.6	34.84	39.69	4.85	41.08	48.7	7.62
OACS	36.9	48.71	11.81	41.46	53.3	11.84	**37.38**	**41.97**	4.59	40.23	47.67	7.44
AOCS	32.7	49.38	16.68	42.55	56.06	13.51	32.69	38.74	6.05	39.23	46.39	7.16
COAS	37.1	50.68	13.58	43.26	56.51	13.25	33.43	39.23	5.8	41.1	49.32	8.22
SAOC	32.18	49.01	16.83	41.05	54.7	13.65	32.65	37.89	5.24	39.14	46.91	7.77
OASC	37.23	49.29	12.06	42.14	53.73	11.59	34.86	38.69	3.83	41.77	48.89	7.12
SOAC	35.82	48.7	12.88	41.02	53.22	12.2	34.75	39.77	5.02	40.27	47.22	6.95
SOCA	36.82	48.99	12.17	42.63	53.68	11.05	34.48	39.97	5.49	38.87	45.75	6.88
ASOC	31.5	47.59	16.09	41.32	53.99	12.67	31.51	36.3	4.79	38.41	45.26	6.85
CAOS	33.12	50.09	16.97	43.08	**57.32**	14.24	34.79	41.03	6.24	39.96	48.76	8.8
SCAO	32.17	49.8	**17.63**	40.81	56.48	15.67	31.87	39.16	**7.29**	38.66	48.28	**9.62**
OCAS	37.94	49.69	11.75	**44.36**	55.36	11.0	36.32	41.42	5.1	40.52	47.42	6.9
COSA	37.39	**51.01**	13.62	44.11	55.78	11.67	33.24	38.75	5.51	42.67	**51.0**	8.33
CASO	31.15	48.49	17.34	41.73	56.69	14.96	33.4	39.72	6.32	38.2	46.05	7.85
CSAO	31.1	48.71	17.61	39.87	56.17	**16.3**	32.45	39.65	7.2	41.59	50.99	9.4
ACOS	28.92	43.95	15.03	40.51	53.64	13.13	34.64	40.36	5.72	36.25	45.21	8.96
ACSO	29.22	45.37	16.15	39.51	51.94	12.43	32.39	38.07	5.68	36.84	45.09	8.25
SCOA	35.61	49.52	13.91	42.71	54.56	11.85	31.39	37.17	5.78	40.0	47.5	7.5
CSOA	37.51	50.06	12.55	41.79	53.67	11.88	33.31	39.15	5.84	**42.71**	50.66	7.95
SACO	30.37	46.69	16.32	38.1	52.4	14.3	32.67	39.04	6.37	34.96	43.78	8.82
ASCO	28.9	46.12	17.22	37.93	51.09	13.16	32.61	38.92	6.31	37.4	45.78	8.38
Avg.	34.11	48.69	14.57	41.77	54.47	12.7	33.68	39.23	5.55	39.63	47.55	7.92
Std.	3.21	1.69	2.24	1.71	1.59	1.55	1.55	1.3	0.85	2.12	2.11	0.79

Table 24: F1 scores of GPT-4o-mini. The **best score** for each column is bold.

Gemini-v1.5	ASQP						ACOS					
	Rest15			Rest16			Laptop			Rest		
Orders	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**
AOSC	36.56	46.73	10.17	44.36	56.44	12.08	34.52	37.88	3.36	41.0	45.47	4.47
OCSA	36.74	45.98	9.24	46.81	58.07	11.26	36.78	39.19	2.41	47.51	52.38	4.87
OSAC	35.16	44.66	9.5	44.93	56.73	11.8	33.77	36.74	2.97	47.29	52.87	**5.58**
OSCA	36.31	46.45	10.14	46.5	58.3	11.8	36.07	38.3	2.23	46.3	51.05	4.75
OACS	36.28	45.58	9.3	45.81	57.09	11.28	34.46	36.97	2.51	**48.17**	**53.15**	4.98
AOCS	38.9	48.62	9.72	45.89	57.26	11.37	36.78	40.08	3.3	44.65	48.74	4.09
COAS	35.35	44.08	8.73	46.26	55.42	9.16	35.17	37.48	2.31	36.78	42.0	5.22
SAOC	37.83	48.67	10.84	45.94	57.65	11.71	33.67	35.94	2.27	46.1	50.59	4.49
OASC	36.14	46.01	9.87	45.71	57.8	12.09	32.2	35.05	2.85	47.23	52.72	5.49
SOAC	36.42	46.53	10.11	45.42	56.18	10.76	32.27	35.42	3.15	47.63	53.06	5.43
SOCA	36.47	46.3	9.83	46.21	56.9	10.69	32.46	34.36	1.9	43.95	48.9	4.95
ASOC	38.14	48.35	10.21	45.22	57.86	12.64	33.4	35.57	2.17	46.3	50.24	3.94
CAOS	38.87	49.83	10.96	46.9	59.06	12.16	35.24	38.52	3.28	39.6	43.53	3.93
SCAO	35.9	49.42	**13.52**	45.3	57.91	12.61	33.15	36.08	2.93	38.0	42.08	4.08
OCAS	**39.13**	48.54	9.41	46.42	58.76	12.34	**37.64**	**40.28**	2.64	47.3	52.01	4.71
COSA	36.4	46.29	9.89	44.69	53.88	9.19	31.57	34.11	2.54	38.26	42.56	4.3
CASO	38.04	**50.14**	12.1	46.76	**59.14**	12.38	33.15	36.69	3.54	39.57	43.74	4.17
CSAO	36.3	48.63	12.33	45.38	58.06	**12.68**	32.41	35.97	**3.56**	36.74	40.92	4.18
ACOS	34.65	44.54	9.89	44.46	56.38	11.92	35.38	38.3	2.92	40.63	45.04	4.41
ACSO	34.92	46.53	11.61	43.65	54.48	10.83	32.1	35.09	2.99	43.55	48.22	4.67
SCOA	37.56	48.04	10.48	45.8	55.3	9.5	31.34	33.43	2.09	36.56	40.75	4.19
CSOA	37.93	48.53	10.6	**47.42**	57.04	9.62	30.36	32.42	2.06	38.3	42.83	4.53
SACO	34.51	47.77	13.26	41.9	53.79	11.89	33.67	36.77	3.1	42.08	46.97	4.89
ASCO	34.19	47.23	13.04	43.37	54.59	11.22	33.42	36.53	3.11	42.0	46.68	4.68
Avg.	36.61	47.23	10.61	45.46	56.84	11.37	33.79	36.55	2.76	42.73	47.35	4.63
Std.	1.43	1.68	1.34	1.27	1.57	1.08	1.87	1.99	0.5	4.07	4.33	0.48

Table 25: F1 scores of Gemini-v1.5-Pro. The **best score** for each column is bold.

Llama-3.1-70B	ASQP						ACOS					
	Rest15			Rest16			Laptop			Rest		
Orders	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**	**Origin**	**Ours**	**Δ**
AOSC	34.97	50.5	15.53	42.87	57.54	14.67	36.19	41.51	5.32	42.54	51.28	8.74
OCSA	38.35	54.35	16.0	46.25	60.88	14.63	40.11	45.1	4.99	41.82	51.78	9.96
OSAC	35.53	49.33	13.8	45.33	58.25	12.92	38.2	43.54	5.34	44.11	54.1	9.99
OSCA	**40.61**	**57.29**	16.68	45.65	60.32	14.67	41.02	**46.04**	5.02	41.79	51.31	9.52
OACS	38.01	51.38	13.37	46.63	60.47	13.84	37.79	42.65	4.86	**46.13**	**56.45**	10.32
AOCS	36.36	53.0	16.64	43.83	58.56	14.73	34.37	40.07	5.7	43.07	51.68	8.61
COAS	35.82	50.5	14.68	46.88	61.01	14.13	39.36	44.36	5.0	40.6	50.52	9.92
SAOC	34.53	50.62	16.09	41.47	57.83	16.36	33.55	38.82	5.27	42.48	51.77	9.29
OASC	36.14	50.5	14.36	45.47	59.13	13.66	37.75	42.63	4.88	45.15	55.18	10.03
SOAC	34.08	48.41	14.33	44.96	58.75	13.79	35.92	41.35	5.43	45.0	54.83	9.83
SOCA	37.88	55.1	17.22	46.01	60.91	14.9	38.69	44.35	5.66	41.56	51.54	9.98
ASOC	33.1	49.76	16.66	42.73	58.49	15.76	35.7	41.8	6.1	42.54	52.93	10.39
CAOS	38.03	54.15	16.12	45.98	60.8	14.82	36.45	42.51	6.06	37.22	46.49	9.27
SCAO	33.79	51.74	17.95	44.56	60.65	16.09	36.58	41.98	5.4	36.45	46.96	10.51
OCAS	38.65	53.32	14.67	**47.46**	**63.6**	16.14	**41.28**	45.98	4.7	41.86	51.94	10.08
COSA	37.24	52.92	15.68	46.24	59.76	13.52	39.66	45.0	5.34	38.14	47.56	9.42
CASO	34.22	53.64	**19.42**	43.61	61.08	**17.47**	36.08	42.47	**6.39**	35.66	46.08	10.42
CSAO	33.23	51.04	17.81	42.56	59.65	17.09	33.27	39.26	5.99	35.16	45.89	10.73
ACOS	35.09	51.22	16.13	43.63	58.56	14.93	36.67	42.4	5.73	40.2	50.49	10.29
ACSO	32.52	51.39	18.87	41.95	58.63	16.68	34.82	40.94	6.12	38.97	50.87	**11.9**
SCOA	36.19	51.45	15.26	44.59	58.05	13.46	39.23	44.31	5.08	41.42	50.71	9.29
CSOA	36.3	51.74	15.44	44.67	59.2	14.53	38.03	43.86	5.83	40.13	49.72	9.59
SACO	31.68	50.0	18.32	42.66	58.84	16.18	34.81	40.61	5.8	38.83	50.17	11.34
ASCO	32.84	51.04	18.2	44.51	59.88	15.37	35.5	40.79	5.29	41.51	52.68	11.17
Avg.	35.63	51.85	16.22	44.6	59.62	15.01	37.13	42.6	5.47	40.93	50.96	10.02
Std.	2.25	2.03	1.63	1.66	1.4	1.23	2.26	2.01	0.46	2.94	2.82	0.78

Table 26: F1 scores of Llama-3.1-70B-Instruct. The **best score** for each column is bold.