

OASIS: An Open-vocabulary Benchmark for Aspect-Opinion-Sentiment Extraction with Implicit-aspect Specificity

Junyeop Im^{1*}, Jonghyun Park^{1*}, Sang-Chul Lee² and Dong-Kyu Chae¹

¹Hanyang University, Seoul, South Korea

²Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea

Correspondence: dongkyu@hanyang.ac.kr sangchul.lee@dgist.ac.kr

Abstract

In e-commerce product review mining, Aspect-Based Sentiment Analysis (ABSA) aims to extract fine-grained opinions toward specific aspects in reviews. Existing ABSA benchmarks have expanded from aspect-level sentiment classification to opinion-aware tuple extraction. However, they still provide limited support for evaluating *implicit aspects* with fine-grained specificity. In particular, implicit aspects are often omitted, collapsed into placeholders, or mapped to predefined categories, which can obscure what reviewers actually evaluate. We introduce **OASIS**, a review-level **Open-vocabulary** benchmark for Aspect-Opinion-Sentiment (AOS) extraction with **Implicit-aspect Specificity**. OASIS contains 6,775 Amazon reviews across eight product domains and 21,054 annotated tuples, where implicit aspects are directly represented as contextually inferred noun phrases rather than closed-set labels. This design yields an implicit-aspect-heavy evaluation setting, with 70.1% of tuples involving implicit aspects. We evaluate a range of open-source and commercial LLMs under fine-tuned and few-shot settings, showing that full AOS extraction remains challenging even for strong models. We further show that open-vocabulary aspect phrases in OASIS can provide task-useful information. Our benchmark and experimental code is available at: <https://github.com/DILAB-HYU/OASIS>.

1 Introduction

Online product reviews contain detailed evidence about what users value, dislike, and compare when making purchasing decisions (Rahayu et al., 2021). Extracting this evidence requires more than assigning a sentiment label to an entire review: a system must identify the product aspect being evaluated, the opinion expression that conveys the evaluation, and the polarity associated with that aspect. Aspect-Based Sentiment Analysis (ABSA) addresses this

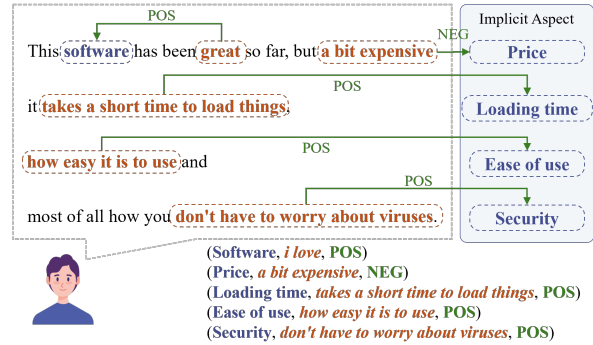


Figure 1: Our AOS triplet extraction examples.

fine-grained opinion mining problem (Liu, 2012), and the field has gradually moved from aspect-level sentiment classification to richer formulations such as *aspect-opinion* extraction (Fan et al., 2019) and *aspect-opinion-sentiment* (AOS) triplet extraction (Peng et al., 2020). Some studies further extend this structure to quadruples by adding predefined aspect categories and placeholders for *implicit aspects or opinions* (Cai et al., 2021).

Despite this progress, existing ABSA benchmarks remain limited for product-review settings where the evaluated aspect is often not explicitly named. Many datasets are built around *sentence-level* instances, which might be convenient for annotation and modeling, but insufficient for reviews whose opinions are distributed across multiple sentences. In such reviews, the opinion expression and the aspect it evaluates are not always located in the same sentence. Sentence-level annotation can therefore neither capture aspect-opinion pairs whose elements appear in different sentences, nor preserve the context needed to identify the intended product property. Recent resources such as DMASTE (Xu et al., 2023) and OATS (Chebolu et al., 2024) have moved toward longer or review-aware settings, but they do not directly address the representation problem that arises when the aspect itself is implicit.

A second limitation concerns how implicit as-

*Equal contribution.

Dataset	Open-Vocab Aspect	Domains	Reviews	Tuples	Avg token/ review	Avg tuple/ review	Implicit Aspect (%)
ACOS (Cai et al., 2021)	✗	2	6,362	9,416	15.51 [†]	1.48	22.63
MEMD-ABSA (Cai et al., 2025)	✗	5	18,094	28,194	16.13 [†]	1.56	24.44 [†]
M-ABSA (Wu et al., 2025)	✗	7	14,780	21,049	18.17 [†]	1.42	21.70
DMASTE (Xu et al., 2023)	✗	8	7,524	28,233	59.68	3.75	42.30
OASIS	✓	8	6,775	21,054	47.55	3.11	70.10

Table 1: Comparison of our benchmark to representative ABSA datasets. Values marked [†] are computed by us from publicly released raw data of each dataset.

pects are encoded. Prior datasets typically mark missing aspect terms with a placeholder such as NULL or [ia], sometimes together with a predefined aspect category (Cai et al., 2021, 2025; Xu et al., 2023; Chebolu et al., 2024). Such representations collapse fine-grained semantic information about what the reviewer actually evaluated. For example, (see Figure 1) in a review sentence stating that a product “*takes a short time to load things*,” the relevant target is not merely an absent (NULL) aspect term or an ambiguous category such as SOFTWARE#OPERATION_PERFORMANCE (the category schema of SemEval-2016 (Pontiki et al., 2016)); it is specifically the product’s *loading time*. Such distinctions are central to product-review analysis, yet they are difficult to evaluate when implicit aspects are reduced to closed-set labels.

To address this issue, we introduce **OASIS**, a review-level **O**pen-vocabulary benchmark for **A**OS triplet extraction with **I**mplicit-aspect **S**pecificity. It is constructed from Amazon Reviews 2023 (Hou et al., 2024) and covers 6,775 reviews across eight product categories, with 21,054 annotated AOS tuples. Each target instance is a full product review, where explicit aspects are annotated as they appear in the text, while implicit aspects are represented as contextually inferred noun phrases rather than placeholders or predefined categories. Figure 1 illustrates our annotation scheme, where a single review contains not only explicit aspects such as software, but also implicit aspects that do not appear in the surface text such as loading time, ease of use, and security, which are directly annotated as fine-grained noun phrases. Inspired by OOMB (Heo et al., 2025), the annotation follows a human-in-the-loop manner (see Figure 2), where an LLM first proposes candidate tuples, and human annotators independently verify, refine, or reject them under detailed guidelines. As a result, 70.1% of the tuples in OASIS involve implicit aspects, which makes our benchmark a challenging testbed for evaluating whether models can semantically cap-

ture the specific objects of evaluation that do not appear as surface spans in the review.

To examine the properties and utility of OASIS, we conduct extensive experiments and analyses using diverse LLMs under fine-tuned and few-shot settings. First, we observe that even the best-performing model (fine-tuned Llama-3.1-8B) reaches only 33.42 F1 score on AOS tuple extraction, indicating that jointly predicting implicit aspects, opinion evidence, and sentiment polarity remains challenging for current generative models. Second, we introduce a task that asks a model to retrieve the review/opinion evidence corresponding to a given aspect-sentiment pair, and highlight that our open-vocabulary aspect phrases carry more useful information than the existing placeholder-based representations. Third, cross-dataset transfer experiments reveal that existing ABSA datasets adopt largely incompatible annotation conventions for implicit aspects, limiting direct transfer across resources. Finally, beyond the traditional *Exact Match* (EM) evaluation, we additionally report *Relaxed Match* (RM) and *Contextual Match* (CM) analyses (Heo et al., 2025), showing how EM penalizes semantically valid predictions solely because of surface-form differences.

2 Related Work

Opinion mining and ABSA have been studied over a wide range of user-generated online content, including product reviews, social media posts, community forums, and video comments. Our work focuses on ABSA datasets dedicated to e-commerce product reviews.

ABSA datasets have evolved from the SemEval shared tasks (Pontiki et al., 2014) toward rich combinations of opinion-related elements. Early SemEval datasets provided simple annotations of aspects and sentiment polarities in the *Laptop* and *Restaurant* domains (Pontiki et al., 2015). SemEval-2016 further refined aspect categories into *Entity#Attribute* pairs and expanded the bench-

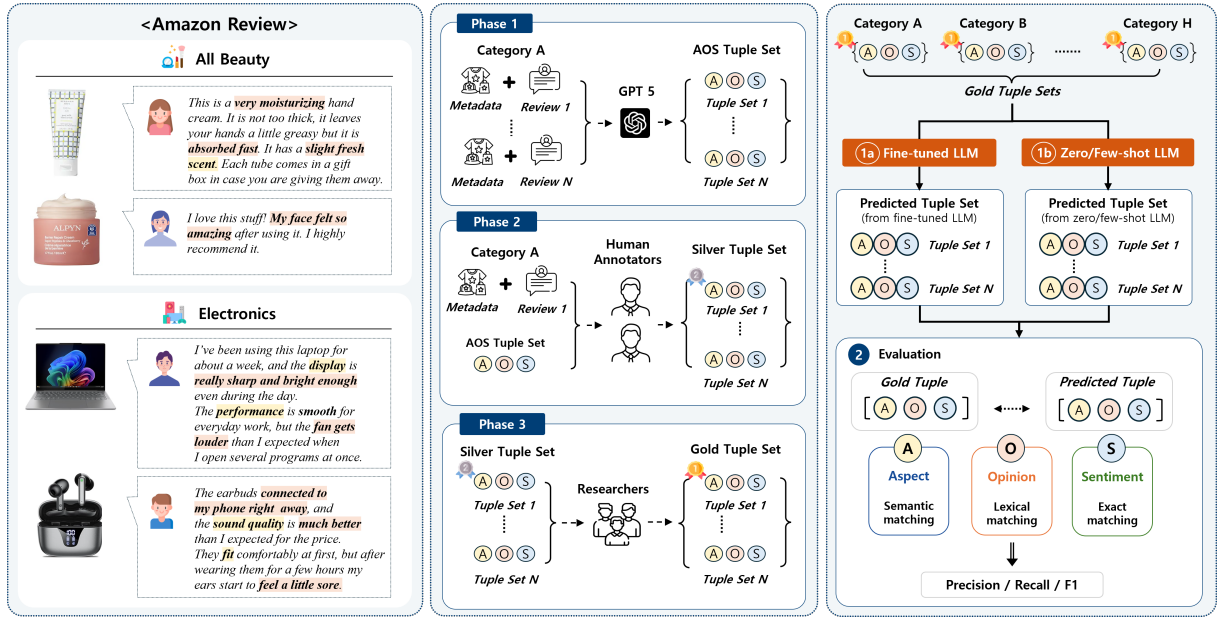


Figure 2: Overview of our benchmark construction framework.

mark coverage to multilingual and multi-domain settings (Pontiki et al., 2016). Building on SemEval annotations, Fan et al. (2019), Peng et al. (2020), and Wu et al. (2020) incorporated opinion spans to construct datasets for aspect-opinion pair and aspect sentiment triplet extraction. ACOS (Cai et al., 2021) and ASQP (Zhang et al., 2021) further extended this line of work to quadruple extraction by explicitly modeling *implicit aspects/opinions* within a four-element annotation schema.

Subsequent ABSA benchmarks have primarily focused on domain diversification and language expansion. For example, MEMD-ABSA (Cai et al., 2025) extends the ACOS schema to five domains and evaluates generalization under an open-domain setting; DMASTE (Xu et al., 2023) builds an AOS extraction dataset from Amazon product reviews across eight domains. M-ABSA (Wu et al., 2025) constructs a multilingual parallel dataset covering 21 languages and seven domains.

Beyond e-commerce, OATS (Chebolu et al., 2024) provides both sentence-level quadruple annotations and review-level tuples across Amazon FineFood, Coursera, and TripAdvisor domains. OOMB (Heo et al., 2025) is a recent benchmark for evaluating LLMs as opinion miners over diverse online content, including blogs, online communities, and YouTube comments. More related works are summarized in Appendix C.

3 OASIS Benchmark

This section presents our OASIS benchmark.

3.1 Schema

We treat the full review as the basic unit of analysis. For each review, we annotate a set of (*aspect*, *opinion-set*, *sentiment*) tuples, where each tuple captures what product property is evaluated, which opinion expression(s) support the evaluation, and what sentiment polarity is expressed.

Open-vocabulary Aspect. OASIS preserves *explicit aspects* as they appear in the review text (e.g., “battery” in “the battery is great”), while *implicit aspects* are directly annotated as fine-grained noun phrases inferred from the review context (e.g., “battery life” inferred from “run as a night-light all night”). This design preserves the semantic specificity of the target actually evaluated by the reviewer. To the best of our knowledge, OASIS is the first ABSA benchmark to represent implicit aspects with open-vocabulary inferred noun phrases rather than placeholders or closed-set category labels.

Product-level Satisfaction Marker. We observed that some reviews express overall satisfaction or dissatisfaction with the product without targeting a specific attribute (e.g., “I would still prefer this remover”, “highly recommend”). Forcing such cases into a fine-grained noun phrase (e.g., quality) would arbitrarily narrow the evaluation target and introduce semantic distortion. We therefore assign #satisfaction to the aspect field for those cases. We note that this is the only dedicated marker in our schema; all other implicit aspects are represented as open-vocabulary noun phrases.

Opinion Set. In many cases, aspects in a review are evaluated by a single opinion expression. However, sometimes aspects are supported by multiple opinion expressions across the review. For example, the aspect *texture* may be described by both “creamy” and “silky smooth”. To capture such cases without changing the tuple structure, our schema represents the opinion component as a *set* that may contain one or more spans, rather than forcing exactly one opinion per aspect. In the constructed dataset, 17.87% of tuples contain two or more opinion spans (see Table 18).

Sentiment Polarity. For each aspect–opinion–set pair, we assign one of three sentiment polarities: POS, NEG, or NEU. We assign NEU when the review expresses mixed evidence toward the same aspect, when the evaluation is weak, or when the polarity is ambiguous from the review context.

3.2 Annotation Pipeline

Inspired by OOMB (Heo et al., 2025), our annotation pipeline combines LLM-based candidate generation with human verification. An overview of the benchmark construction and evaluation framework is shown in Figure 2.

LLM Candidate Generation. We randomly sampled reviews from the Amazon Reviews 2023 (Hou et al., 2024) spanning eight product categories. We provided each review as input to GPT-5¹. We instructed the model to generate all extractable *aspect-opinion* (or *opinion set* if applicable)-*sentiment* tuples from the review. Aspects were required to be fine-grained noun phrases grounded in the review; when the aspect was not explicitly mentioned, the model was instructed to infer an appropriate phrase from the review context. This process yielded 22,622 raw candidate tuples from 6,955 reviews.

Human Annotation. Next, we employed four human annotators (NLP graduate students at Hanyang University) to verify the LLM generations. Before annotation, the annotators completed training on the annotation guidelines. They were divided into two groups, and each group was assigned four product categories. Within each category, two annotators independently reviewed all candidate tuples.

For each candidate tuple, annotators assigned one of three labels: (1) ACCEPT: the tuple is valid as-is; (2) HOLD: the tuple is not clearly incorrect

Metric	Value
Quadratic weighted Cohen’s κ	0.602
Raw Agreement	0.900
Cohen’s κ	0.460
Gwet’s AC1	0.889

Table 2: Inter-annotator agreement metrics.

but requires refinement; (3) REJECT: the tuple is not valid. Tuples labeled as HOLD, as well as tuples with annotator disagreement, were resolved through a discussion process involving an author (the authors did not participate in the initial annotation). Through this process, each candidate tuple was assigned one of three final outcomes: (i) retained without modification, (ii) revised in its span or polarity, or (iii) removed from the dataset. In the end, each annotator was paid a \$750 reward.

We finally conducted an author review stage. In this step, tuples missed by GPT-5 were identified and added to the gold set, and some existing candidates were removed after further inspection. Across these stages, 1,329 candidate tuples were discarded, while 2,191 tuples entered the gold set only after intervention: 2,162 were revised, in the sense that their opinion span was normalized to the review surface form, one or more fields were modified, or several candidates were merged or split, and a further 29 were newly added because no corresponding candidate had been generated. Revisions most often targeted the aspect field. More details of the entire annotation pipeline are in Appendix A.1 and Table 15.

Statistics. As shown in Table 1, OASIS exhibits an implicit aspect ratio of 70.1%, suggesting that the flexibility of our open-vocabulary schema allows implicit aspects to be captured as fine-grained noun phrases, rather than collapsed into placeholders or ambiguous category labels. In addition, because OASIS annotates full reviews rather than isolated sentences, it preserves aspect and sentiment evidence that depends on inter-sentence context. This makes OASIS a review-level benchmark tailored to e-commerce product reviews, where user opinions are sometimes distributed across multiple sentences rather than expressed within a single local sentence. OASIS has an average review length of 47.55 tokens and an average of 3.11 tuples per review. More detailed dataset statistics are reported in Appendix A.3.

¹gpt-5-2025-08-07

Judgment	Anchor	Individual	%
OASIS preferred	8	45	22.1
Independent preferred	5	22	11.2
Tie	21	133	64.2
Mixed	6	–	2.5
Total	40	200	100.0

Table 3: Blind human evaluation results on 240 sampled reviews. “Mixed” denotes anchor items where 4 raters did not converge on a single label.

3.3 Inter-Annotator Agreement Analysis

In our three-way labeling, tuples labeled as HOLD are not necessarily invalid candidates. After the subsequent discussion, approximately 70% of held tuples are retained in the gold set either unchanged or after modification, while only about 30% are removed. Thus, HOLD functions as a label for editable borderline candidates rather than for clearly invalid ones. As a result, ACCEPT and HOLD often lead to the same final gold-set decision, whereas ACCEPT and REJECT do not. We therefore aim to reflect this asymmetric distance among labels in the inter-annotator agreement (IAA) analysis.

Unlike prior ABSA datasets that report IAA at the span-level (Cai et al., 2021; Xu et al., 2023), we measure agreement over candidate-level verification labels. To ensure transparent reporting under this setting, we follow the suggestion from Klie et al. (2024) that multiple metrics should be reported rather than using a single metric. We employ four complementary metrics: *quadratic-weighted Cohen’s κ* (Cohen, 1968), *raw agreement*, *unweighted Cohen’s κ* , and *Gwet’s AC1* (Gwet, 2008). Among them, we consider *quadratic-weighted Cohen’s κ* as our primary metric, since it directly captures the asymmetric label distances discussed above by penalizing ACCEPT–HOLD disagreements less severely than ACCEPT–REJECT disagreements.

As shown in Table 2, across all candidate tuples, the assigned annotator pairs achieve a raw agreement above 90%. *Quadratic-weighted Cohen’s κ* is about 0.14 higher than *unweighted Cohen’s κ* , which indicates that most disagreements occur between adjacent labels. The significantly higher AC1 score compared to *Cohen’s κ* reflects the strong skew of the label distribution toward ACCEPT. These results indicate that approximately 85% of LLM-generated candidate tuples are accepted as valid by both annotators, suggesting that the LLM serves as an effective initial candidate generator in our annotation pipeline. More results of our IAA analysis

are reported in Appendix A.2.

3.4 Human Evaluation of Annotation Quality

To assess the annotation quality, we randomly sample 240 reviews from the constructed OASIS, with 30 reviews from each category. For these reviews, we conduct a new round of independent human annotation *without LLM assistance*, following the schema explained in Section 3.1. The annotators are not shown the OASIS gold tuples. We compare, for each identical input review, the tuple set produced by the OASIS construction pipeline against the separately produced human-only tuple set.

In our schema, both open-vocabulary aspect phrases and opinion span boundaries are affected by annotators’ lexical choices. As a result, the same underlying meaning may be expressed with different surface forms (Yang et al., 2025b). Therefore, rather than measuring a strict-match metric, we conduct a blind human preference evaluation. We hired four other NLP graduate students to compare OASIS and the independent annotation shown in randomized left–right order, and assign one of three judgments: {*OASIS preferred*, *independent annotation preferred*, *Tie*}. Among the 240 reviews, 40 anchor reviews are evaluated by all four evaluators, while the remaining 200 reviews are assigned to one evaluator each.

As shown in Table 3, 64.2% of the 240 reviews are judged as *Tie*, indicating that the two annotations are often not clearly distinguishable under human judgment. Among the 80 reviews for which a clear preference is assigned, OASIS is preferred in 53 cases, while the independent annotation is preferred in 27 cases. This suggests that OASIS is at least comparable to the independent human-only annotation. In the six Mixed anchor cases where the four evaluators do not converge on a single label, no evaluator judges the independent annotation to be superior. These findings support the use of an LLM-assisted human-in-the-loop pipeline for constructing an open-vocabulary ABSA benchmark by providing broader tuple coverage while maintaining annotation quality.

We also performed the systematic exact-match evaluation, and the results are in Appendix A.4.

4 Experiments

Through our OASIS benchmark, we conduct a comprehensive analysis to address the following four research questions: (RQ1) How well do current

Model		ASC	ATE									AOS								
		EM	EM			RM			CM			EM			RM			CM		
		Acc	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SFT	Llama-3.1-8B	<u>97.02</u>	63.97	<u>59.46</u>	61.63	66.34	61.67	<u>63.92</u>	71.80	66.75	69.18	<u>35.52</u>	33.01	<u>34.22</u>	<u>60.05</u>	<u>55.82</u>	<u>57.86</u>	76.39	<u>71.01</u>	73.60
	Qwen2.5-7B	97.33	59.90	53.91	56.75	63.31	56.97	59.97	67.73	60.95	64.16	33.23	29.90	31.48	55.80	50.22	52.86	72.20	64.97	68.40
	Gemma2-9B	96.78	40.53	58.12	47.76	42.10	60.37	49.61	45.81	65.69	53.98	23.15	<u>33.21</u>	27.28	37.96	54.43	44.72	48.81	70.00	57.52
	Qwen3.6-27B	96.99	66.63	65.26	65.94	67.95	66.55	67.25	85.57	83.80	84.68	37.72	36.94	37.33	61.99	60.71	61.34	<u>76.27</u>	74.70	75.48
0-shot	Llama-3.1-8B	78.43	4.35	16.44	6.88	5.10	19.26	8.07	5.56	20.99	8.79	1.33	5.03	2.11	3.43	12.94	5.42	6.07	22.90	9.59
	Qwen2.5-7B	<u>95.59</u>	11.95	13.03	12.47	14.89	16.24	15.54	19.24	20.99	20.07	5.40	5.89	5.64	11.46	12.50	11.96	28.02	30.57	29.24
	Gemma2-9B	78.04	3.71	18.11	6.16	4.61	22.52	7.66	5.16	25.20	8.57	0.91	4.46	1.52	2.79	13.61	4.63	5.24	25.59	8.70
	GPT-4o	95.52	42.32	<u>41.73</u>	<u>42.03</u>	46.40	<u>45.76</u>	<u>46.08</u>	<u>51.51</u>	<u>50.79</u>	<u>51.15</u>	<u>19.00</u>	<u>18.74</u>	<u>18.87</u>	39.02	<u>38.48</u>	<u>38.75</u>	59.33	<u>58.51</u>	<u>58.91</u>
	GPT-4o-mini	94.68	34.86	28.80	31.54	39.68	32.77	35.90	44.43	36.70	40.20	13.52	11.16	12.23	32.37	26.74	29.28	54.06	44.66	48.91
	GPT-5	95.94	<u>41.60</u>	53.14	46.67	<u>43.51</u>	55.58	48.81	64.82	82.80	72.71	21.08	26.93	23.65	<u>38.11</u>	48.68	42.75	<u>56.79</u>	72.52	63.71
1-shot	Llama-3.1-8B	87.85	10.08	22.47	13.91	10.33	23.05	14.27	11.00	24.53	15.19	3.80	8.48	5.25	7.22	16.10	9.97	13.94	31.10	19.25
	Qwen2.5-7B	92.23	19.96	28.99	23.64	21.61	31.38	25.60	23.49	34.12	27.82	8.25	11.98	9.77	17.19	24.96	20.36	28.24	41.02	33.45
	Gemma2-9B	87.24	19.60	20.30	20.38	14.56	37.33	20.95	15.74	40.34	22.64	5.44	13.94	7.83	11.10	28.46	15.97	16.90	43.32	24.31
	GPT-4o	95.04	49.89	<u>45.42</u>	<u>47.55</u>	53.11	<u>48.35</u>	50.61	<u>58.84</u>	<u>53.57</u>	<u>56.08</u>	25.79	<u>23.48</u>	<u>24.58</u>	46.00	<u>41.88</u>	<u>43.84</u>	68.32	<u>62.19</u>	<u>65.11</u>
	GPT-4o-mini	94.11	44.66	29.28	35.37	<u>48.90</u>	32.06	38.73	55.48	36.37	43.94	18.06	11.84	14.30	40.57	26.59	32.13	<u>63.60</u>	41.69	50.36
	GPT-5	94.98	<u>45.55</u>	51.56	48.37	47.42	53.67	<u>50.35</u>	71.30	80.69	75.70	<u>24.56</u>	27.79	26.07	<u>41.87</u>	47.44	44.46	62.79	71.06	66.67
3-shot	Llama-3.1-8B	91.93	14.59	39.20	21.27	15.41	41.40	22.46	17.27	46.38	25.17	6.03	16.20	8.79	11.92	32.01	17.37	19.23	51.65	28.03
	Qwen2.5-7B	89.70	15.33	36.75	21.63	17.01	40.78	24.00	19.50	46.77	27.53	6.24	14.95	8.80	13.25	31.77	18.70	22.12	53.04	31.22
	Gemma2-9B	88.54	23.75	24.70	24.57	17.47	45.90	25.31	19.60	51.51	28.40	7.00	18.40	10.14	13.79	36.22	19.97	21.94	57.64	31.78
	GPT-4o	95.74	52.62	<u>50.60</u>	51.59	55.26	<u>53.14</u>	54.18	<u>61.58</u>	<u>59.22</u>	<u>60.38</u>	28.90	27.79	28.33	49.03	<u>47.15</u>	48.07	70.60	<u>67.90</u>	69.22
	GPT-4o-mini	92.69	<u>47.12</u>	38.00	42.07	<u>49.73</u>	40.11	44.40	55.02	44.37	49.12	20.38	16.44	18.20	41.47	33.45	37.03	<u>63.70</u>	51.37	56.87
	GPT-5	<u>95.37</u>	45.95	52.80	<u>49.14</u>	47.42	54.72	<u>50.93</u>	70.52	81.03	75.41	<u>23.27</u>	<u>26.74</u>	<u>24.88</u>	<u>42.24</u>	48.54	<u>45.17</u>	62.55	71.87	<u>66.89</u>

Table 4: The results of benchmarking the LLMs on OASIS under various settings. All values are expressed as percentages (%).

large models perform on OASIS? **(RQ2)** Do the open-vocabulary aspect representations in OASIS provide useful information for identifying the target aspect of evaluation? **(RQ3)** To what extent are the annotations in OASIS compatible with existing ABSA resources? **(RQ4)** To what extent does the EM metric underestimate model performance on OASIS?

4.1 Experimental Settings

Tasks. Given a review R consisting of n sentences, the goal is to extract a set of tuples in the form of (aspect, opinion-set, sentiment):

$$T = \{(a_i, O_i, p_i)\}_{i=1}^{|T|},$$

where a_i denotes an aspect expression, $O_i = \{o_{i,1}, \dots, o_{i,k_i}\}$ is a set of opinion expressions associated with a_i , and $p_i \in \{\text{POS}, \text{NEG}, \text{NEU}\}$ denotes the sentiment polarity toward a_i .

Following standard ABSA evaluation practice, we evaluate the following three subtasks: (1) **Aspect Term Extraction (ATE)** identifies the set of aspect expressions $\{a_i\}$ from the review R ; (2) **Aspect Sentiment Classification (ASC)** predicts the sentiment polarity p_i given a gold aspect a_i ; (3) **Aspect-Opinion-Sentiment Extraction (AOS)** extracts the complete tuple (a_i, O_i, p_i) .

Models. We evaluate two types of LLMs. (1) **Open-source LLMs:** Llama-3.1-8B (Grattafiori

et al., 2024), Qwen2.5-7B (Yang et al., 2025a), Gemma2-9B (Riviere et al., 2024), and Qwen3.6-27B (Qwen Team, 2026). (2) **Proprietary LLMs:** GPT-4o and GPT-4o-mini (Hurst et al., 2024), and GPT-5 (Singh et al., 2026).

For fine-tuning, we use LoRA (Hu et al., 2022). We split OASIS at the review level into train/validation/test sets with an 80%/10%/10% ratio (see Appendix Table 8). All models are evaluated with task-specific prompt templates (see Table 34–Table 37 in Appendix). Detailed experimental settings are provided in Appendix B.

Metrics. To evaluate the ability of LLMs to perform structured opinion extraction, we use three tuple-matching criteria inspired by Heo et al. (2025): (1) **Exact Match (EM)** judges that a predicted tuple is counted as correct only when every element exactly matches the corresponding element in a gold tuple; (2) **Relaxed Match (RM)** complements the strict EM evaluation by allowing span-level variation through a similarity-based criterion that combines semantic matching for aspects and set-level lexical matching for opinions; (3) **Contextual Match (CM)** also aims to recover semantically faithful predictions rejected by EM, but unlike RM, it does not require a similarity function or threshold; instead, it uses an LLM-as-a-judge based on

Claude Sonnet 4.6². For all three matching criteria, sentiment polarity is evaluated by exact match.

Due to space limitations, we explain more details of the evaluation metrics in Appendix B.3.

4.2 Benchmarking Current LLMs on OASIS

To answer RQ1, we evaluate the LLMs under various settings introduced in the previous section. Table 4 reports the results. Under EM, even the best fine-tuned model, Qwen3.6-27B, reaches only 37.33 F1 on AOS, indicating its challenging nature for current models. Across fine-tuned models, ATE F1 ranges from 47.76 to 65.94, while AOS F1 remains lower, ranging from 27.28 to 37.33. There is also a large gap between fine-tuned and few-shot performance, suggesting that the benchmark is not readily solved by instruction following alone. Although GPT-4o is a considerably stronger proprietary model than the 7B-9B open-source baselines, its 3-shot performance does not consistently surpass task-specific fine-tuning.

The RM and CM results suggest that EM misses a substantial number of predictions that are acceptable under more flexible matching criteria. Under RM, performance already increases notably, and the scores increase further under CM. For example, the AOS F1 of fine-tuned Qwen3.6-27B rises from 37.33 under EM to 61.34 under RM, and reaches 75.48 under CM. A similar pattern appears for ATE. Many predictions rejected by EM differ from the gold annotation in surface form, span boundary, or lexical realization, rather than being entirely incorrect. Section 4.5 analyzes this issue in detail.

Since GPT-5 generated the initial candidate tuples (Section 3.2), the gold annotations might be biased toward the specific wording that GPT-5 produced. If this were the case, GPT-5’s own predictions would be accepted at a higher rate than those of other models under EM. However, its EM F1 stays within roughly five points of GPT-4o in every prompted setting, and GPT-4o is in fact higher at 3-shot on both ATE (51.59 vs. 49.14) and AOS (28.33 vs. 24.88), suggesting that the gold set does not simply reward GPT-5’s own wording.

Overall, the results show that OASIS (1) benefits largely from task-specific fine-tuning, (2) is primarily challenging at the level of full (*aspect-opinion set-sentiment*) extraction, and (3) remains far from saturated even for strong current models.

	Ablations	Rev.-EM	Rev.&Opi.-EM	Rev.&Opi.-RM
Llama	OASIS	84.28	38.09	49.52
	Normalized	77.84	31.70	40.75
	[ia]	73.84	21.21	32.89
	NULL	74.77	24.17	35.93
Qwen	OASIS	86.40	37.67	49.21
	Normalized	79.26	32.28	42.33
	[ia]	75.85	23.86	34.63
	NULL	76.53	22.83	34.24
Gemma	OASIS	60.82	20.39	20.62
	Normalized	61.67	15.77	17.23
	[ia]	58.02	10.78	12.26
	NULL	58.02	11.03	14.20

Table 5: ASOR F1 results for the fine-tuned LLMs across aspect representation variants.

4.3 Diagnosing Open-Vocabulary Aspects with Opinion Retrieval

To answer RQ2, we define a simple task, (Aspect-Sentiment-conditioned Opinion Retrieval). Given an (*aspect, sentiment*) pair, ASOR retrieves the review in which the pair is expressed and extracts the opinion span in that review that supports the sentiment toward the given aspect. It thus highlights how well an *aspect* representation identifies the target of evaluation and whether an (*aspect-sentiment*) query is specific enough to retrieve the corresponding evidence.

To isolate the effect of aspect representation, we fine-tune separate models on four variants that differ only in how the aspect is represented: (1) **OASIS** uses the open-vocabulary aspect phrase from OASIS without modification. (2) **Normalized** embeds open-vocabulary aspects and clusters them with HDBSCAN³, replacing each aspect with the most frequent expression in its cluster; this variant approximates a coarser normalization of long-tail aspect expressions. (3) **[ia]** and (4) **NULL** replace implicit aspects with the placeholder tokens [ia] and NULL, respectively, mimicking the closed-vocabulary treatment commonly used in existing ABSA datasets.

We report three metrics. Rev.-EM evaluates whether the model correctly identifies the review containing the given (*aspect-sentiment*) pair. Rev.&Opi.-EM further requires the model to extract the corresponding *opinion span* from the identified review, using EM for opinion evaluation. Unlike Rev.&Opi.-EM, Rev.&Opi.-RM allows partially overlapping opinion spans to be credited. The RM evaluation for opinion matching is detailed in Appendix B.3. The prompt used for ASOR is pro-

²claude-sonnet-4-6, <https://www.anthropic.com>

³<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html>

Model	Train	ATE									AOS					
		OASIS			DMASTE			SemEval			OASIS			DMASTE		
		EM	RM	CM	EM	RM	CM	EM	RM	CM	EM	RM	CM	EM	RM	CM
Llama	OASIS	61.63	63.92	69.18	15.55	18.73	49.97	30.90	<u>35.33</u>	<u>53.26</u>	34.22	57.86	73.60	3.38	12.47	41.24
	DMASTE	12.71	16.26	<u>50.67</u>	62.28	63.48	69.81	25.90	26.54	39.59	3.11	11.65	38.68	28.33	54.59	53.66
	SemEval	<u>15.66</u>	<u>20.63</u>	34.94	<u>25.79</u>	<u>27.63</u>	33.32	80.27	82.60	86.01	–	–	–	–	–	–
Qwen	OASIS	56.75	59.97	64.16	15.82	18.90	48.73	29.04	<u>34.42</u>	<u>52.15</u>	31.48	52.86	68.40	3.04	11.75	39.67
	DMASTE	<u>13.37</u>	<u>17.38</u>	<u>56.18</u>	59.51	60.33	67.90	<u>29.66</u>	31.43	45.02	3.09	11.75	43.77	25.09	50.54	49.92
	SemEval	10.86	14.45	26.17	<u>17.23</u>	<u>18.45</u>	21.98	68.93	71.48	74.35	–	–	–	–	–	–
Gemma	OASIS	47.76	49.61	53.98	12.09	13.95	<u>36.68</u>	23.16	27.13	39.28	27.28	44.72	57.52	3.13	9.24	30.78
	DMASTE	<u>13.68</u>	<u>18.21</u>	<u>55.20</u>	64.31	65.37	72.00	<u>44.57</u>	<u>46.69</u>	<u>53.97</u>	3.09	12.18	42.72	30.18	56.27	55.92
	SemEval	9.90	13.10	21.34	<u>14.83</u>	<u>15.54</u>	17.36	25.01	25.54	25.96	–	–	–	–	–	–

Table 6: Cross-dataset F1 results. SemEval2014 does not include opinion annotations, so AOS is not reported. **Bold** indicates in-domain performance, and underline indicates the best cross-domain result for each model and column.

vided in Appendix Table 37.

Table 5 shows that OASIS achieves the best performance on most metrics, and that performance decreases as the aspect representation is reduced. This pattern is especially clear in the metrics that require opinion extraction. For example, the Rev.&Opi.-EM F1 of Llama-3.1-8B drops from 38.09 with OASIS to 31.70 with Normalized, and further to 21.21 and 24.17 when implicit aspects are replaced with [ia] and NULL, respectively. A similar trend appears under the more permissive Rev.&Opi.-RM criterion. The performance drop under Normalized indicates that long-tail fine-grained aspect phrases are not merely surface variants, but often contain information that helps specify the intended evaluation target. The larger drop under [ia] and NULL suggests that collapsing implicit aspects into placeholders removes information needed to retrieve the relevant opinion evidence. Overall, the results support the design choice of preserving implicit aspects as open-vocabulary phrases.

4.4 Cross-dataset Transfer Analysis

To answer RQ3, we examine whether knowledge learned from existing ABSA datasets transfers to OASIS, and whether the reverse transfer is possible. We fine-tune the three open-source LLMs on each of SemEval2014 (Pontiki et al., 2014), DMASTE (Xu et al., 2023), and our OASIS, and evaluate each fine-tuned model on each test set of the three datasets. The results are reported in Table 6. Naturally, in-domain performance (Source = Target) substantially exceeds cross-dataset transfer performance. But both ATE and AOS performance degrade sharply under cross-dataset transfer, even when transferring between OASIS and DMASTE that

Aspect near-miss	
Gold	“battery life”
Prediction	“battery longevity”
Gold	“durability”
Prediction	“sturdiness”
Opinion near-miss	
Gold	“they lasted around a week in a half”
Prediction	“lasted around a week in a half”
Gold	“not too thick and not watered down”
Prediction	“not too thick”

Table 7: Representative near-miss cases.

share a closely related *aspect–opinion–sentiment* extraction formulation; for example, AOS on EM F1 remains around 3% in both directions.

Such degradation under cross-dataset transfer is likely driven by differences in annotation schema. Unlike OASIS, DMASTE replaces all implicit aspects with the placeholder token [ia]; SemEval2014 only annotates explicit aspect terms that appear in the review. Given these annotation differences, predictions from a model trained on one dataset can be semantically faithful to the review, yet penalized by EM when evaluated on another dataset. As shown by the metric gaps in Table 6, RM and CM can recognize non-exact but acceptable matches that EM treats as errors. We provide a detailed case study of this mismatch in Appendix B.5 and Table 31. These results suggest schema-level incompatibility among ABSA datasets.

4.5 Limitations of Exact Match

To answer RQ4, we analyze *near-miss* predictions in depth, i.e., cases rejected by EM but accepted by RM or CM. We demonstrate both their prevalence and linguistic characteristics.

Table 7 presents representative near-miss cases. Aspect near-misses mainly arise from semantic variants that refer to the same product

property. For example, “battery life” and “battery longevity”, or “durability” and “sturdiness”, differ in surface form but denote the same underlying aspect. Opinion near-misses show two major patterns. First, the predicted span may differ slightly in boundary from the gold span. For instance, “*they lasted around a week in a half*” and “*lasted around a week in a half*” convey the same core evaluative content, but differ in whether the preceding pronoun is included. Second, when multiple opinion spans are annotated for a single aspect, models may extract only a subset of the gold opinion set. Such cases, where the gold answer consists of multiple spans, have also been recognized as requiring specialized evaluation designs in multi-span question answering (Li et al., 2022). For example, if the gold opinion is “*not too thick and not watered down*” and the model extracts only “*not too thick*”, EM treats the prediction as incorrect, even though it captures part of the correct evaluative evidence. In such multiple-span settings, neither exact string matching nor token-level partial matching fully captures the structure of correctness (Yang et al., 2025b). More details can be found in Appendix B.6.

5 Conclusion

We introduced OASIS, an open-vocabulary benchmark for AOS extraction in e-commerce reviews. The annotation schema for OASIS represents implicit aspects as fine-grained noun phrases inferred from review context, rather than collapsing them into placeholders or coarse categories. Our experiments show that this setting remains challenging for current LLMs and provides a complementary evaluation signal for implicit-aspect understanding beyond existing ABSA resources. We hope OASIS will support future work on models that can more accurately recover what reviewers evaluate, how they express their opinions, and how such opinions should be interpreted in context.

Limitations

Language and Modality. OASIS focuses only on English, text-only e-commerce product reviews. Multilingual ABSA (Wu et al., 2025) and multimodal reviews involving images or videos are beyond the scope of this paper. We leave these promising and important directions for future work.

Trade-off in Open-Vocabulary Aspect Annotation. Our open-vocabulary annotation schema

represents implicit aspects as fine-grained noun phrases, which helps preserve the semantic specificity of what reviewers evaluate. At the same time, this design introduces a tradeoff: annotators’ lexical choices are directly reflected in the aspect labels. For example, the same implicit aspect may be annotated with surface-form variants such as loading time, loading speed, or load time. Although RM and CM mitigate this lexical variability at evaluation time, the annotation layer itself does not canonicalize such variants.

Evaluation based on LLM-as-a-judge. Our Contextual Match (CM) metric relies on Claude Sonnet 4.6 as the judge model. Therefore, if the model is deprecated or undergoes model drift, the CM scores reported in this paper may not be exactly reproducible in future work. In addition, we cannot completely rule out the possibility that biases of the judge model, such as preferences for particular surface forms or its own criteria for semantic equivalence, affect CM decisions. To mitigate this dependence, we report CM together with EM and RM, allowing the evaluation results to be examined from multiple perspectives.

Sentiment Polarity Imbalance. OASIS reflects the positive skew commonly observed in online product reviews: 77.97% of its tuples are positive, and the positive ratio remains between 75.37% and 80.56% across all eight domains (see Appendix Table 9). While this distribution is natural for e-commerce reviews and is also observed in many review-based ABSA benchmarks, it may make aggregate scores less informative about model robustness on minority polarities, especially neutral tuples.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00345398).

References

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. *ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2069–2079, Online. Association for Computational Linguistics.

- Hongjie Cai, Nan Song, Zengzhi Wang, Qiming Xie, Qiankun Zhao, Ke Li, Siwei Wu, Shijie Liu, Heqing Ma, Jianfei Yu, and Rui Xia. 2025. [Memd-absa: a multi-element multi-domain dataset for aspect-based sentiment analysis](#). *Lang. Resour. Eval.*, 59(3):2501–2529.
- Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. [Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.
- 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.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.
- Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. [Adaptive recursive neural network for target-dependent Twitter sentiment classification](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.
- 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.
- Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. [Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning](#). In *Proceedings of the 2018 International Conference on Digital Health, DH ’18*, page 121–125, New York, NY, USA. Association for Computing Machinery.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Kilem Li Gwet. 2008. [Computing inter-rater reliability and its variance in the presence of high agreement](#). *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.
- Felix Hamborg and Karsten Donnay. 2021. [NewsMTSC: A dataset for \(multi-\)target-dependent sentiment classification in political news articles](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online. Association for Computational Linguistics.
- Nils Constantin Hellwig, Jakob Fehle, Markus Bink, and Christian Wolff. 2024. [GERestaurant: A German dataset of annotated restaurant reviews for aspect-based sentiment analysis](#). In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 123–133, Vienna, Austria. Association for Computational Linguistics.
- Ryang Heo, Yongsik Seo, Junseong Lee, and Dongha Lee. 2025. [Can large language models be effective online opinion miners?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23097–23136, Suzhou, China. Association for Computational Linguistics.
- Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. [A challenge dataset and effective models for aspect-based sentiment analysis](#). 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 6280–6285, Hong Kong, China. Association for Computational Linguistics.
- Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. [Analyzing dataset annotation quality management in the wild](#). *Computational Linguistics*, 50(3):817–866.
- Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. [MultiSpanQA: A dataset for multi-span question answering](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260,

- Seattle, United States. Association for Computational Linguistics.
- Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. *G-Eval: NLG evaluation using GPT-4 with better human alignment*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. *Visual attention model for name tagging in multimodal social media*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.
- Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. *SemEval-2016 task 4: Sentiment analysis in Twitter*. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California. Association for Computational Linguistics.
- Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. *SemEval-2013 task 2: Sentiment analysis in Twitter*. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. *Knowing what, how and why: A near complete solution for aspect-based sentiment analysis*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.
- 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.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. *SemEval-2014 task 4: Aspect based sentiment analysis*. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.
- Qwen Team. 2026. Qwen3.6-27b. <https://qwen.ai/blog?id=qwen3.6-27b>.
- Agus Rahayu, Dian Herdiana Utama, and Ririe Novianty. 2021. *The impact of online customer reviews on purchase intention in online marketplace*. In *Proceedings of the 5th Global Conference on Business, Management and Entrepreneurship (GCBME 2020)*, pages 471–477. Atlantis Press.
- Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 178 others. 2024. *Gemma 2: Improving open language models at a practical size*. *Preprint*, arXiv:2408.00118.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. *SemEval-2017 task 4: Sentiment analysis in Twitter*. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. *Openai gpt-5 system card*. *Preprint*, arXiv:2601.03267.
- Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. *Sentfin 1.0: Entity-aware sentiment analysis for financial news*. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335.
- ChengYan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Yun Xue, and Barbara Plank. 2025. *M-ABSA: A multilingual dataset for aspect-based sentiment analysis*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2530–2557, Suzhou, China. Association for Computational Linguistics.
- Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. *Grid tagging scheme for aspect-oriented fine-grained opinion extraction*. In *Findings of the Association for Computational*

Linguistics: EMNLP 2020, pages 2576–2585, Online. Association for Computational Linguistics.

Ting Xu, Huiyun Yang, Zhen Wu, Jiaze Chen, Fei Zhao, and Xinyu Dai. 2023. [Measuring your ASTE models in the wild: A diversified multi-domain dataset for aspect sentiment triplet extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2837–2853, Toronto, Canada. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2025a. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.

Soyoung Yang, Hojun Cho, Jiyoung Lee, Sohee Yoon, Edward Choi, Jaegul Choo, and Won Ik Cho. 2025b. [Single ground truth is not enough: Adding flexibility to aspect-based sentiment analysis evaluation](#). In *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, Albuquerque, New Mexico. Association for Computational Linguistics.

Jianfei Yu and Jing Jiang. 2019. [Adapting bert for target-oriented multimodal sentiment classification](#). *IJCAI*.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. [Adaptive co-attention network for named entity recognition in tweets](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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, 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*, 35(11):11019–11038.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*.

A Dataset Construction

A.1 Annotation Details

Annotation Pipeline. We adopt a human-in-the-loop annotation pipeline that combines LLM-based

candidate generation with human verification, following Heo et al. (2025). The GPT-5 prompt used to generate the initial candidate tuples is provided in Table 32. Before annotation, all annotators completed a *three-day* training session. The training process included studying the annotation guidelines in Table 33, practicing labels on example candidates, comparing practice results, and resolving disagreements through discussion.

Annotation Interface. We provided annotators with a dedicated web interface for the verification task. As shown in Figure 4, the left side of the interface displays the original review text and product information, and the right side presents the LLM-generated candidate tuples together with {ACCEPT, HOLD, REJECT} decision buttons for each tuple. Annotators selected a verification label for each candidate tuple by jointly referring to the review and the product information.

When assigning a HOLD label, annotators additionally selected one of the predefined hold sub-types, indicating which element requires revision, such as the aspect phrase, opinion span, or sentiment polarity. When the sub-type *aspect word incorrect* was selected, the interface further required the annotator to attach a replacement suggestion before the label could be submitted, so that the direction of each aspect correction originated from the annotator rather than from the authors. The annotation work was assigned by product category, and two annotators independently verified all reviews within the same category.

Verification Examples. Table 11 shows the candidate-level labels assigned by the two annotators for a single review, together with the hold sub-types they selected and the resulting gold aspects. Table 12 presents examples of modification, merging, and removal, showing the tuples proposed by GPT-5, the corresponding gold tuples, and how each discrepancy was resolved.

A.2 Inter-Annotator Agreement Analysis

Table 13 reports inter-annotator agreement (IAA) metrics for each of the eight product categories. Raw agreement remains consistently high across categories, ranging from 0.84 to 0.95. However, *quadratic-weighted Cohen’s κ* varies substantially across domains, from 0.205 in All Beauty to 0.801 in Tools and Home Improvement. The All Beauty category, which shows the lowest κ , consists of substantially longer and denser reviews

than the other categories, with an average review length of approximately 102 tokens and 5.36 tuples per review (see Table 17). This is consistent with the observation that annotators are more likely to diverge when judging the same candidate tuple in longer, information-rich reviews. In addition, *quadratic-weighted* κ is consistently higher than *unweighted* κ across all categories, indicating that most disagreements occur between adjacent labels such as ACCEPT–HOLD, rather than polar disagreements such as ACCEPT–REJECT.

Table 14 presents the distribution of label pairs assigned by the two annotators for each category. Overall, ACCEPT–ACCEPT accounts for 85.1% of all cases, whereas ACCEPT–REJECT accounts for only 2.0%, showing that polar disagreements are relatively rare. The most notable category-level difference appears in Toys and Games, where REJECT–REJECT occurs 223 times, substantially more often than in the other categories (16–103 cases). This suggests that GPT-5 produced invalid or less appropriate candidate tuples relatively more frequently in the Toys and Games category.

Table 15 and Table 16 report, respectively, the category-wise outcome of verifying every GPT-5 candidate and the distribution of hold reasons. The as-is rate is stable across domains, ranging from 86.8% in Toys_and_Games to 93.5% in Tools_and_Home_Improvement. Removals, in contrast, are unevenly distributed: Toys_and_Games alone accounts for 379 of them, more than twice the per-category average. Among the modified tuples, 1,030 cases (82.3%) involve an aspect edit, indicating that LLM-generated candidates are often broadly plausible but frequently require refinement before entering the gold set. The distribution of hold reasons, measured over a different population, points in the same direction. The label *aspect word incorrect* (59.2%) and *aspect refers to product itself* (24.2%) together account for approximately 83% of all hold reasons, whereas *opinion span incorrect* (3.4%) and *sentiment label incorrect* (2.5%) together account for less than 6%. This indicates that the LLM generally identifies where an evaluation occurs in the review and what polarity it conveys, but requires more frequent refinement when assigning the aspect that the evaluation targets.

At the category level, *aspect word incorrect* is especially frequent in Beauty_and_Personal_Care (74.9%) and Toys_and_Games (80.4%). One possible explanation is that products in

Category	Train	Dev	Test	Total
All_beauty	534	66	68	668
Art_crafts_and_sewing	736	92	92	920
Beauty_and_personal_care	672	84	84	840
Electronics	697	87	88	872
Office_products	696	87	87	870
Sports_and_outdoors	725	90	92	907
Tools_and_home_improvement	716	89	91	896
Toys_and_games	641	80	81	802
Total	5,417	675	683	6,775

Table 8: Train/dev/test split of our benchmark by product category. Numbers indicate review counts.

Domain	Pos %	Neg %	Neu %
All_Beauty	75.37	18.77	5.86
Arts_Crafts_and_Sewing	80.56	15.67	3.77
Beauty_and_Personal_Care	76.77	19.80	3.43
Electronics	76.61	20.89	2.51
Office_Products	79.64	17.40	2.96
Sports_and_Outdoors	78.89	16.93	4.18
Tools_and_Home_Improvement	78.87	17.55	3.58
Toys_and_Games	78.28	17.33	4.39
TOTAL	77.97	18.10	3.93

Table 9: Sentiment polarity distribution per category (% of total tuples).

these domains are often evaluated through fine-grained and subjective properties that can be expressed in many different ways. These two domains also have relatively high implicit-aspect ratios (Beauty_and_Personal_Care: 75.9%; Toys_and_Games: 72.3%; see Table 17), meaning that a larger share of aspects must be inferred from context rather than directly extracted; the LLM thus has correspondingly broad latitude in choosing a noun phrase to name an inferred aspect, and its assigned expressions are more often flagged for refinement than accepted as-is.

A.3 Dataset Statistics

OASIS provides an 80/10/10 train/dev/test split at the review level (Table 8), supporting in-domain evaluation across all eight product categories. The number of reviews per category ranges from 668 to 920.

Table 9 reports the sentiment polarity distribution by product category. Overall, positive tuples are dominant, accounting for 77.97% of all annotations, followed by negative tuples (18.10%) and neutral tuples (3.93%). This distribution is consistent with the general tendency of e-commerce platform reviews to be positively skewed, and the variation across categories is relatively small (Pos: 75.37%–80.56%).

Table 17 presents the number of reviews, average token length, number of tuples, and aspect-type distribution for each of the eight product categories. All_Beauty contains reviews that are approximately two to three times longer and more tuple-dense than those in other categories; accordingly, it includes fewer reviews than most other categories. Across all categories, implicit aspects consistently outnumber explicit aspects, with implicit-aspect ratios ranging from 66.29% to 75.85%, compared with explicit-aspect ratios of 24.15% to 33.71%. Among implicit aspects, the dedicated #satisfaction label accounts for 10.03% (All_Beauty) to 25.07% (Toys_and_Games) of all tuples, with an overall average of 18.19%.

Figure 3 shows a heatmap of the top 20 aspects in OASIS, excluding #satisfaction. Aspects such as price, quality, appearance, size, and durability appear frequently across product categories, indicating category-general evaluation targets. By contrast, aspects such as scent and installation are concentrated in specific categories. This pattern suggests that OASIS captures both broadly shared product-review aspects and fine-grained aspects that are specific to particular product categories.

A.4 Annotation Quality Assessment with Exact-Match Evaluation

Using the same 240 reviews and independent annotations described in Section 3.4, we additionally perform a strict-match comparison against the OASIS gold set. For this analysis, we treat the independent annotation as the prediction and the OASIS gold set as the reference.

The strict-match evaluation over all 240 reviews yields $F1 = 0.602$ ($P = 0.633$, $R = 0.574$). The category-wise precision, recall, and F1 results are reported in Table 19. Precision is consistently higher than recall across all eight categories, which is consistent with the observation that the OASIS gold set contains more tuples than the independent annotation.

However, these strict-match scores only quantify surface-level overlap between the two annotations; they do not determine which annotation is semantically more appropriate. Prior work has shown that, even for explicit span annotation, the same meaning can be realized with different surface forms in ABSA evaluation (Yang et al., 2025b). Our open-vocabulary schema naturally extends this variability to implicit aspects, where annotators may

infer semantically similar but lexically different aspect phrases. For this reason, we conducted the blind human preference evaluation in Section 3.4 as the main quality assessment, since it can evaluate semantic appropriateness beyond surface-form agreement.

B Experimental Details

B.1 Proprietary and Open-source LLMs

Proprietary LLMs. We evaluate proprietary LLMs through the OpenAI API⁴. The models used in our experiments are as follows:

- **GPT-4o-mini:** gpt-4o-mini-2024-07-18
- **GPT-4o:** gpt-4o-2024-11-20
- **GPT-5:** gpt-5-2025-08-07

Open-source LLMs. We evaluate open-source LLMs from their Hugging Face model repositories and conduct all the experiments using two NVIDIA A100 80GB GPUs. The models used in our experiments are as follows:

- **Llama-3.1-8B:** meta-llama/Llama-3.1-8B
- **Qwen2.5-7B:** Qwen/Qwen2.5-7B
- **Gemma2-9B:** google/gemma-2-9b
- **Qwen3.6-27B:** Qwen/Qwen-3.6-27B

B.2 Fine-tuning Details

Hyperparameter	Value
LoRA rank (r)	16
LoRA α	32
LoRA dropout	0.05
Target modules	q_proj, k_proj, v_proj, o_proj
Learning rate	2e-4
Batch size	4
Training epochs	3
Max seq. length	1024

Table 10: LoRA fine-tuning hyperparameters used for all open-source LLMs.

We fine-tune the open-source LLMs using LoRA (Hu et al., 2022). The same hyperparameter configuration is applied to all open-source models, and the detailed values are summarized in Table 10.

⁴<https://platform.openai.com/docs/models>

B.3 Details of Evaluation Metrics

We use three tuple-matching criteria to evaluate the ability of LLMs to perform structured opinion extraction.

(1) Exact Match (EM): Following prior work on opinion extraction (Zhang et al., 2022; Cai et al., 2021), a predicted tuple is counted as correct only when every element exactly matches the corresponding element in a gold tuple. For the opinion element, a match is defined only when the gold opinion set O_{gt} and the predicted opinion set O_{pred} are identical.

(2) Relaxed Match (RM): EM requires exact surface-form agreement, and therefore treats a prediction as incorrect even when it differs only slightly in span boundary or lexical realization. To reduce this rigidity, we design the similarity-based Relaxed Match (RM) criterion inspired by Heo et al. (2025). Given two spans x and y , we regard them as matching if their similarity $\text{sim}(x, y)$ is greater than or equal to a threshold τ . However, rather than applying the same RM criterion uniformly to all tuple elements, we distinguish between aspect and opinion matching according to their linguistic properties. For aspects, we use cosine similarity computed with SentenceBERT⁵ to capture semantic similarity. We refer to this criterion as **Asp-RM**. A predicted aspect is matched to a gold aspect if the cosine similarity between their embeddings is greater than or equal to the aspect threshold τ_a .

Since OASIS allows multiple opinion spans for a single aspect ($|O_i| \geq 1$), opinion matching is defined as a **set-level** comparison between the gold opinion set O_{gt} and the predicted opinion set O_{pred} . We refer to this criterion as **Opi-RM**. Given a gold opinion set $O_{gt} = \{o_1, \dots, o_n\}$ and a predicted opinion set $O_{pred} = \{p_1, \dots, p_m\}$, we first compute the lexical similarity $\text{lex}(p_j, o_i)$ for every predicted-gold pair and construct a matching matrix M :

$$M[j][i] = \begin{cases} 1 & \text{if } \text{lex}(p_j, o_i) \geq \tau_{\text{lex}} \\ 0 & \text{otherwise.} \end{cases}$$

Here, τ_{lex} denotes the pairwise lexical-similarity threshold for determining whether two opinion phrases match. From the matching matrix, we define set-level precision P_{op} and recall R_{op} as follows. P_{op} is the proportion of predicted opinion spans that match at least one gold opinion span,

and R_{op} is the proportion of gold opinion spans that are matched by at least one predicted span:

$$P_{op} = \frac{|\{j : \exists i, M[j][i] = 1\}|}{|O_{pred}|},$$

$$R_{op} = \frac{|\{i : \exists j, M[j][i] = 1\}|}{|O_{gt}|},$$

$$F1_{op} = \frac{2P_{op}R_{op}}{P_{op} + R_{op}}.$$

The predicted and gold opinion sets are finally considered matched if the set-level $F1_{op}$ is greater than or equal to the opinion-set threshold τ_{op} :

$$\text{Opi-RM}(O_{pred}, O_{gt}) = \begin{cases} 1 & \text{if } F1_{op} \geq \tau_{op} \\ 0 & \text{otherwise.} \end{cases}$$

(3) Contextual Match (CM). To further complement EM, we introduce an LLM-as-a-judge evaluation criterion (Zheng et al., 2023; Liu et al., 2023), which we call Contextual Match (CM). CM takes as input a predicted tuple $\hat{t} = (\hat{a}, \hat{O}, \hat{p})$, a gold tuple $t = (a, O, p)$, and determines whether the prediction is semantically equivalent to the gold tuple. This allows predictions that refer to the same aspect, opinion evidence, and sentiment polarity to be counted as correct even when their surface forms differ. We use Claude Sonnet 4.6⁶ as the judge model. The prompt and detailed settings for CM are provided in Table 36.

For all three matching criteria, sentiment polarity is evaluated by exact match. ATE is evaluated by matching only the aspect element, whereas AOS is evaluated as full-triple matching, requiring the aspect, opinion set, and sentiment to match simultaneously. ASC is measured as sentiment classification accuracy under the condition that the gold aspect is given.

B.4 Threshold Sensitivity Analysis and Adopted Values

To examine how threshold choices affect Opi-RM-based evaluation, we perform 3-shot AOS evaluation over combinations of the aspect relaxed-match threshold $\tau_a \in \{0.5, 0.6, 0.7, 0.8\}$ and the opinion-side thresholds $\tau_{\text{lex}}, \tau_{op} \in \{0.5, 0.6, 0.7, 0.8\}$ (Table 27–Table 30).

The results show that increasing the thresholds leads to a gradual decrease in F1 for all models. For example, when $\tau_a = 0.5$ is fixed, GPT-4o’s

⁵<https://www.sbert.net>

⁶claude-sonnet-4-6, <https://www.anthropic.com>

F1 decreases from $56.13 \rightarrow 53.05 \rightarrow 43.48 \rightarrow 40.55$ as the opinion threshold increases from 0.5 to 0.8. Similarly, when the opinion threshold is fixed at 0.5, increasing the aspect threshold from $\tau_a = 0.5$ to $\tau_a = 0.8$ decreases GPT-4o’s F1 from $56.13 \rightarrow 51.54 \rightarrow 49.63 \rightarrow 48.07$. Importantly, the relative ranking among models remains stable across threshold settings. GPT-4o achieves the highest performance and Llama-3.1-8B the lowest performance under all threshold combinations, suggesting that the main comparison among models is not overly sensitive to the specific threshold choice.

We also analyze qualitative examples where EM, RM, and CM produce different decisions on the same review (Table 31). Case 1, the predicted and gold aspect spans are near-identical (Asp-RM=1.00) and are accepted across all opinion thresholds. By contrast, Cases 3, the predicted aspect is semantically different from the gold aspect (e.g., *softness* vs. *afterfeel*, *shine* vs. *appearance*) and receives a low Asp-RM score (0.15 and 0.4); but CM accepts them as semantically equivalent. In Case 2, a valid opinion variant is accepted at lower thresholds ($\tau_{op} = 0.5, 0.6$) but rejected once τ_{op} is raised to 0.7 or higher. Case 4, where the sentiment polarity is wrong (predicted NEG vs. gold NEU), and Case 5, where the aspect–opinion pair is mismatched (*exfoliation* vs. *scent*, Asp-RM = 0.23), are both rejected.

Based on these observations, we adopt $\tau_a = 0.8$ for aspect semantic matching and $\tau_{lex} = \tau_{op} = 0.5$ for opinion lexical/set-level matching.

B.5 Annotation-Mismatch Case Study

Table 6 shows substantial performance degradation in cross-dataset transfer. To better understand how task-definition incompatibility arises, we compare predictions generated by a model fine-tuned on OASIS against the gold annotations of DMASTE and SemEval2014 evaluation examples (Table 22).

Case 1: Treatment of implicit aspects in DMASTE. DMASTE collapses all implicit aspects into a single placeholder token, [ia], regardless of their semantic content. In contrast, OASIS preserves each implicit aspect as a free-form textual label, such as *price* or *longevity*. As a result, a model trained on OASIS generates semantically faithful predictions such as *price*, *longevity*, and *color*, but these predictions are counted as incorrect because the DMASTE gold annotation expects the placeholder form [ia].

Case 2: Fine-grained aspect separation in OASIS. The review in Case 2 mentions multiple distinct product properties, such as pores, hydra-

tion, and sunscreen. OASIS captures these targets as separate aspects, annotating them as *pores*, *hydration*, and *sunscreen*. Accordingly, a model trained on OASIS also extracts these targets as fine-grained aspect units. However, if the target dataset groups these evaluations under a single broad aspect such as *stuff*, fine-grained predictions that are semantically more specific can still be penalized under Exact Match.

Case 3: Exclusion of implicit aspects in SemEval2014. The review in Case 3 describes the laptop’s loading time, ease of use, and security without explicitly naming these aspects. Since SemEval2014 annotates only aspect terms that appear as explicit surface spans in the input text, its gold annotation includes only the explicit aspect *software*, while the three implicit aspects are omitted. OASIS annotates all four aspects for the same input, and a model trained on OASIS extracts all four. However, the three implicit-aspect predictions that have no corresponding tuples in the SemEval2014 gold annotation are penalized as false positives.

These cases suggest that the low cross-dataset scores reflect not only model limitations, but also schema incompatibility arising from different annotation philosophies. In particular, schemas that do not preserve the semantic content of implicit aspects can limit the evaluation of implicit aspect understanding. By explicitly preserving fine-grained implicit aspects, OASIS provides a complementary resource for evaluating capabilities that are not fully captured by existing ABSA datasets.

B.6 Ablation Study on Exact Match

EM can underestimate a model’s extraction ability by treating semantically acceptable predictions as mismatches, and this issue becomes more pronounced for AOS, where the evaluation unit is a structured tuple rather than a single span. Table 23 reports, for ATE and AOS, the proportion of prediction–gold pairs that are accepted by CM or RM but fail under EM. Under CM, this ratio is approximately 12% for ATE and approximately 53% for AOS across the three fine-tuned models. This gap reflects the different evaluation structures of the two tasks. ATE requires matching only a single aspect element, whereas AOS requires the aspect, opinion, and sentiment to match jointly; therefore, a minor variation in any one element can cause the entire tuple to fail under EM. This effect is especially strong for opinion spans, which exhibit

greater lexical and boundary variation than aspects. As a result, a substantial portion of semantically acceptable predictions, and more than half at the tuple level under CM, are missed by EM. This suggests that the low EM scores observed in §4.2 partly arise from surface-form mismatch rather than complete extraction failure.

Overall, EM failures arise from different sources depending on the tuple element: aspects mainly fail due to semantic variation among expressions referring to the same property, whereas opinions mainly fail due to boundary mismatch and partial extraction. This explains why element-specific Relaxed Match, which applies different relaxation criteria according to the linguistic properties of each element, consistently recovers performance relative to EM. The ablation results for applying Asp-RM, Opi-RM, and their combination are reported in Table 20. CM, which judges correctness based on contextual semantic equivalence rather than surface-form identity, also accepts many of these near-miss cases and provides a complementary way to diagnose the underestimation caused by EM.

C ABSA Benchmarks Across Diverse Domains

Research on ABSA and related benchmark construction has been extended beyond e-commerce product reviews to a wide range of opinion-bearing domains.

In the hospitality domain, MAMS (Jiang et al., 2019) is built from Citysearch New York restaurant reviews and annotates only instances in which multiple aspects with different sentiment polarities co-occur within the same sentence. This design mitigates the simplification caused by single-aspect sentence distributions. ASAP (Bu et al., 2021) collects 46,730 restaurant reviews from a Chinese online-to-offline (O2O) e-commerce platform and provides both sentiment labels for 18 fine-grained aspect categories and 5-star review-level ratings, supporting a multi-task setting for aspect-category sentiment analysis (ACSA) and rating prediction. GERestaurant (Hellwig et al., 2024) annotates 3,078 sentences extracted from German TripAdvisor restaurant reviews following the SemEval-2015/2016 schema. By reporting baselines for ACD · ACSA · E2E-ABSA · TASD, it fills a resource gap for German ABSA.

In the social media domain, the SemEval Twitter task series (Nakov et al., 2013, 2016; Rosen-

thal et al., 2017) has provided shared-task benchmarks for message-level and topic-based Twitter sentiment analysis. Dong et al. (2014) annotate 6,940 tweets collected through the Twitter API with negative/neutral/positive sentiment polarity for each target entity. Twitter-15 (Zhang et al., 2018) and Twitter-17 (Lu et al., 2018) were originally released for multimodal named entity recognition, and were later repurposed as multimodal target/aspect-level sentiment resources when Yu and Jiang (2019) added 3-way sentiment polarity annotations.

In the news domain, NewsMTSC (Hamborg and Donnay, 2021) annotates instances extracted from U.S. political news articles and provides a separate multi-target test set containing sentences with two or more semantically distinct targets, thereby extending target-dependent sentiment classification to political news. SEntFiN 1.0 (Sinha et al., 2022) annotates 10,753 financial news headlines from *The Economic Times* with more than 1,000 company and sector entities, together with their associated sentiment labels.

Beyond these domains, ABSA-style resources have also been developed for healthcare. Gräßer et al. (2018) construct a dataset from patient drug reviews on *Drugs.com* and *Druglib.com*, annotating sentiment toward three predefined aspects: overall satisfaction, side effects, and effectiveness.

<i>Review</i> : “This seems to be in general a nice face serum, which moisturise my skin very well... I actually can notice the improvements so you can use it from time to time to keep the results. But the biggest problem is the glass packaging, you can actually cut yourself...”				
Candidate aspect	Opinion span (excerpt)	Labels	Hold reason	Gold aspect
serum	“nice face serum”	HLD/HLD	aspect refers to product itself	#satisfaction
moisturization	“moisturise my skin very well”	ACC/ACC	—	moisturization
results	“can notice the improvements”	ACC/HLD	aspect word incorrect	functionality
price	“a little expensive for a one week set”	ACC/ACC	—	price
packaging	“glass packaging, you can actually cut yourself”	ACC/HLD	opinion span incorrect	packaging

Table 11: Candidate-level verification labels assigned by the two annotators for a single All Beauty review, with disagreements shown in bold. The two hold sub-types lead to different outcomes: *serum* names the reviewed product rather than an evaluated property and is replaced by the #satisfaction marker, whereas *results* is a valid but underspecified aspect name that is refined to *functionality*. In the remaining cases the gold aspect is unchanged.

Case 1: Modified — the candidate names the product, not the evaluated property	
<i>Review</i> : “Great little tools that really work! One star off for smushed tips and cracked packaging.”	
GPT-5 candidate	OASIS gold
Tuple 1 (<u>tools</u> , “really work”, POS)	(effectiveness, “really work”, POS)
<i>Resolution</i> : Both annotators assigned a hold label with the sub-type <i>aspect refers to product itself</i> and suggested <i>effectiveness</i> as the evaluated property. The suggestion was adopted after discussion.	
Case 2: Merged — two candidates evaluate a single underlying property	
<i>Review</i> : “I’m not sure what happened, but this is a photo of everything that was included with my kit. It only included one set of 12 colors, and without the others, this actually isn’t very useful to me. I can’t give this more than two stars when it doesn’t have what it said it would. I hope they fixed this, which it appears they have from the other reviews. Shame. It was otherwise fine.”	
GPT-5 candidate	OASIS gold
Tuple 1 (<u>contents</u> , “only included one set of 12 colors”, NEG)	(set composition, “only included one set of 12 colors” \\ “without the others, this actually isn’t very useful to me”, NEG)
Tuple 2 (<u>usefulness</u> , “without the others, this actually isn’t very useful to me”, NEG)	—
<i>Resolution</i> : An annotator assigned a hold label on the grounds that the second opinion restates the consequence of the first. The two candidates were consolidated into one gold tuple with a merged opinion set.	
Case 3: Removed — the candidate refers to a non-target product	
<i>Review</i> : “My youngest daughter has Eczema and struggles with dry itchy skin. This lotion worked well at keeping her skin hydrated, without causing any irritation. We have also ordered the Gryph & IvyRose Shampoo & Conditioner and it works well for her also. Love this brand!!”	
GPT-5 candidate	OASIS gold
Tuple 1 (<u>shampoo & conditioner</u> , “works well for her”, POS)	—
<i>Resolution</i> : Both annotators rejected the candidate because the aspect belongs to a different product mentioned in passing rather than to the reviewed item.	

Table 12: Representative candidate-to-gold outcomes. Each case shows the tuples proposed by GPT-5, the corresponding tuples in the OASIS gold set, and how the discrepancy was resolved during verification. Underlining marks the aspect field, where the two annotations differ in all three cases. Opinion sets containing multiple spans are separated by \\.

Category	N	Quadratic weighted Cohen’s κ	Raw Agreement	Cohen’s κ	Gwet’s AC1
All Beauty	3,752	0.205	0.843	0.172	0.826
Arts Crafts and Sewing	2,512	0.673	0.926	0.524	0.920
Beauty and Personal Care	2,639	0.455	0.895	0.405	0.885
Electronics	2,725	0.592	0.915	0.438	0.908
Office Products	2,669	0.525	0.878	0.441	0.864
Sports and Outdoors	2,923	0.779	0.950	0.633	0.947
Tools and Home Improvement	2,765	0.801	0.951	0.617	0.947
Toys and Games	2,637	0.673	0.857	0.552	0.830
Total	22,622	0.602	0.900	0.460	0.889

Table 13: Per-category inter-annotator agreement metrics. The Total row reports values computed on the concatenated data of all categories.

Category	acc-acc	hld-hld	rej-rej	acc-hld	hld-rej	acc-rej	Total
All Beauty	3,074	71	17	490	10	90	3,752
Arts Crafts and Sewing	2,216	58	53	160	5	20	2,512
Beauty and Personal Care	2,254	74	33	210	9	59	2,639
Electronics	2,391	35	66	180	5	48	2,725
Office Products	2,192	94	58	259	5	61	2,669
Sports and Outdoors	2,648	33	97	104	12	29	2,923
Tools and Home Improvement	2,509	16	103	117	0	20	2,765
Toys and Games	1,970	67	223	243	11	123	2,637
Total	19,254	448	650	1,763	57	450	22,622
% of total	85.1	2.0	2.9	7.8	0.3	2.0	100.0

Table 14: Per-category breakdown of label pair distributions between the two annotators. *acc-hld*: accept $\leftrightarrow$ hold; *hld-rej*: hold $\leftrightarrow$ reject; *acc-rej*: accept $\leftrightarrow$ reject.

Category	Cand.	Removed	Gold	as-is	(%)	norm.	modified	merged	split	added
All Beauty	3,752	157	3,581	3,151	88.0	122	268	25	6	9
Arts Crafts and Sewing	2,512	98	2,386	2,154	90.3	58	129	35	4	6
Beauty and Personal Care	2,639	108	2,505	2,198	87.7	69	206	27	4	1
Electronics	2,725	138	2,551	2,329	91.3	88	94	37	0	3
Office Products	2,669	153	2,471	2,126	86.0	69	219	49	4	4
Sports and Outdoors	2,923	153	2,752	2,558	93.0	91	83	19	0	1
Tools and Home Improvement	2,765	143	2,598	2,428	93.5	68	72	26	2	2
Toys and Games	2,637	379	2,210	1,919	86.8	54	180	52	2	3
Total	22,622	1,329	21,054	18,863	89.6	619	1,251	270	22	29

Table 15: Category-wise candidate-to-gold outcomes of the verification pipeline. *Cand.* is the number of GPT-5 candidate tuples and *Gold* the number of tuples in the final OASIS gold set; the six rightmost outcome columns partition *Gold* and sum to it, and *(%)* gives *as-is* as a share of *Gold*. *removed*: a candidate discarded, e.g. hallucinated, unsupported by the review, or referring to a non-target product. *as-is*: entered the gold set unchanged after verification. *normalized*: the semantic tuple was preserved while the opinion expression was corrected to the original review surface form (e.g. fixing an LLM typo or replacing a paraphrase with the exact span). *modified*: at least one of the aspect, opinion, or sentiment fields was revised. *merged*: several candidates were consolidated into one gold tuple, so that the gold count is lower than the number of candidates involved. *split*: one candidate was divided into several gold tuples. *added*: a tuple absent from the candidate set was added during the author review stage. Representative cases for each outcome are given in Appendix A.1.

Category	Aspect word incorrect	Aspect refers to product itself	Opinion span incorrect	Sentiment label incorrect	Same aspect across tuples	Others	Total
All Beauty	68.69	17.76	5.76	5.14	0.31	2.34	642
Arts Crafts and Sewing	37.37	40.57	1.42	1.42	3.20	16.01	281
Beauty and Personal Care	74.93	21.80	1.63	0.82	0.00	0.82	367
Electronics	26.27	13.73	5.88	4.71	14.12	35.29	255
Office Products	72.79	22.79	1.55	1.77	0.00	1.11	452
Sports and Outdoors	26.37	41.21	8.24	2.20	5.49	16.48	182
Tools and Home Improvement	21.48	44.30	2.01	2.01	2.01	28.19	149
Toys and Games	80.41	18.30	1.03	0.00	0.00	0.26	388
Total %	59.24	24.23	3.35	2.47	2.21	8.51	2,716

Table 16: Per-category distribution of hold reasons (% within category, except the rightmost column showing absolute hold counts).

Category	Reviews	Avg Tok/Rev	Tuples	Avg Tup/Rev	Exp %	Imp %	#sat %
All_Beauty	668	102.48	3,581	5.36	33.71	66.29	10.03
Arts_Crafts_and_Sewing	920	30.32	2,386	2.59	30.55	69.45	21.54
Beauty_and_Personal_Care	840	41.17	2,505	2.98	24.15	75.85	20.60
Electronics	872	47.53	2,551	2.93	32.34	67.66	18.35
Office_Products	870	40.82	2,471	2.84	29.02	70.98	21.29
Sports_and_Outdoors	907	45.86	2,752	3.03	31.94	68.06	18.39
Tools_and_Home_Improvement	896	44.86	2,598	2.90	27.79	72.21	14.90
Toys_and_Games	802	40.47	2,210	2.76	27.69	72.31	25.07
TOTAL	6,775	47.55	21,054	3.11	29.90	70.10	18.19

Table 17: Per-category statistics of OASIS. **Exp %** and **Imp %** are reported as percentages of total tuples and sum to 100%. **Exp**: aspect appears as a substring of the review text. **Imp**: aspect does not appear in surface text, including both inferred noun phrases and the #satisfaction reserved label (§ 3.1). The rightmost column reports the share of the #satisfaction label, which is a subset of **Imp**.

Category	Tuples	Opinion spans	Avg Opi/tuple	# opinions per tuple					Multi (≥2) %
				1	2	3	≥4		
all_beauty	3,581	4,393	1.227	2,894	573	105	9	19.18	
art_crafts_and_sewing	2,386	2,783	1.166	2,042	299	38	7	14.42	
beauty_and_personal_care	2,505	3,091	1.234	2,028	388	72	17	19.04	
electronics	2,551	3,179	1.246	2,045	406	84	16	19.84	
office_products	2,471	3,008	1.217	2,037	350	68	16	17.56	
sports_and_outdoors	2,752	3,306	1.201	2,285	393	63	11	16.97	
tools_and_home_improvement	2,598	3,106	1.196	2,172	355	62	9	16.40	
toys_and_games	2,210	2,739	1.239	1,788	332	76	14	19.10	
Total	21,054	25,605	1.216	17,291	3,096	568	99	17.87	

Table 18: Per-category opinion-set cardinality distribution in OASIS. “Multi (≥ 2) %” denotes the share of tuples with two or more opinion spans.

Category	Matched	Ind.	Gold	P	R	F1
all_beauty	93	155	166	0.600	0.560	0.579
art_crafts_and_sewing	51	73	85	0.699	0.600	0.646
beauty_and_personal_care	51	74	88	0.689	0.580	0.630
electronics	40	71	77	0.563	0.519	0.540
office_products	54	70	76	0.771	0.711	0.740
sports_and_outdoors	56	79	85	0.709	0.659	0.683
tools_and_home_improvement	41	79	86	0.519	0.477	0.497
toys_and_games	42	75	83	0.560	0.506	0.532
Overall	428	676	746	0.633	0.574	0.602

Table 19: Strict-match precision, recall, and F1 of independent annotation evaluated against the OASIS gold set, per product category. *Matched* denotes the number of tuples shared between the independent annotation and OASIS; *|Ind.|* and *|Gold|* denote the total tuple counts of independent annotation and OASIS gold set, respectively.

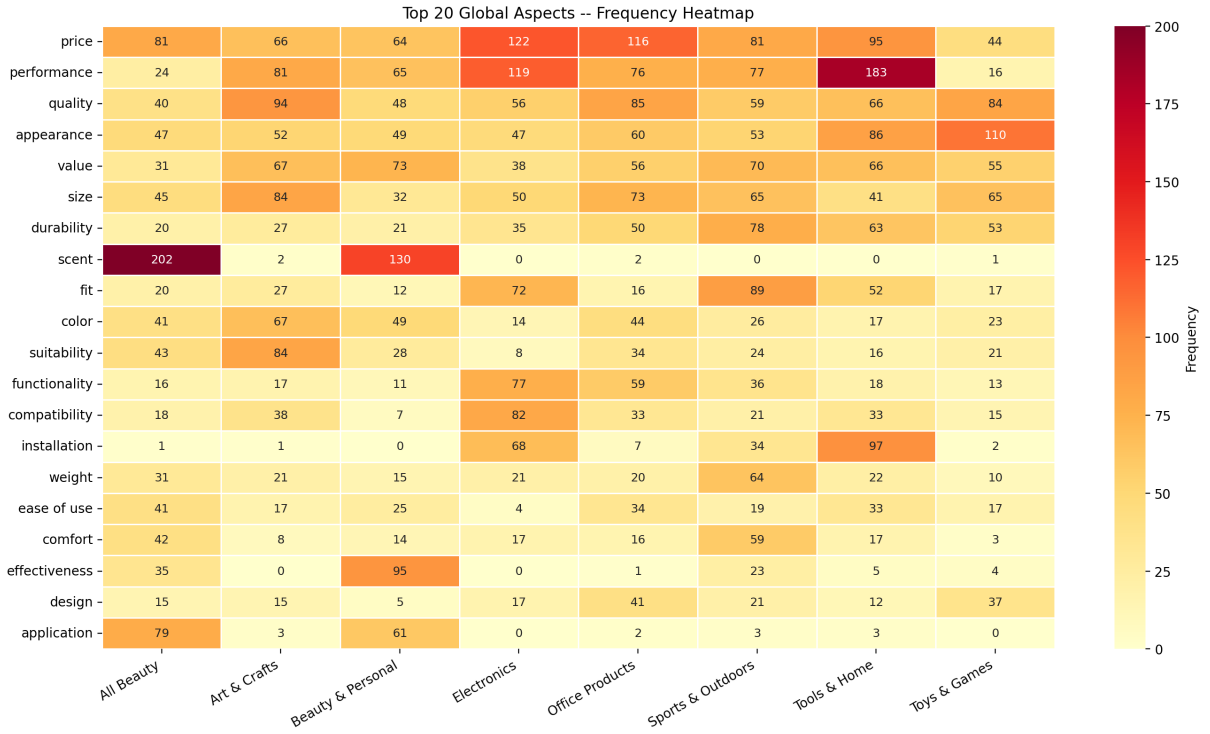


Figure 3: Frequency heatmap of the top 20 aspects across 8 product categories in our benchmark (#satisfaction markers excluded).

Setting	Model	ATE				AOS											
		EM F1	Asp-RM			EM F1	Asp-RM			Opi-RM			Asp-RM + Opi-RM				
			P	R	F1		P	R	F1	P	R	F1	P	R	F1		
Fine-tuned	Llama-3.1-8B	61.63	66.55	61.67	64.01	34.22	36.48	34.07	35.23	58.14	54.05	56.02	60.05	55.82	57.86		
	Qwen2.5-7B	56.75	63.38	56.97	60.00	31.48	36.36	33.95	35.11	52.82	47.53	50.04	55.80	50.22	52.86		
	Gemma2-9B	47.76	42.10	60.37	49.61	27.28	23.99	34.40	28.27	36.55	52.42	43.07	37.96	54.43	44.72		
Zero-shot	Llama-3.1-8B	6.88	5.10	19.26	8.07	2.11	1.49	5.61	2.35	3.07	11.60	4.86	3.43	12.94	5.42		
	Qwen2.5-7B	12.47	14.89	16.24	15.54	5.64	6.28	6.85	6.55	9.62	10.49	10.04	11.46	12.50	11.96		
	Gemma2-9B	6.16	4.61	22.52	7.66	1.52	1.07	5.22	1.78	2.32	11.31	3.85	2.79	13.61	4.63		
	GPT-4o	42.03	46.40	45.76	46.08	18.87	20.94	20.65	20.80	35.81	35.31	35.56	39.02	38.48	38.75		
	GPT-4o-mini	31.54	39.68	32.77	35.90	12.23	15.02	12.41	13.59	28.65	23.67	25.92	32.37	26.74	29.28		
1-shot	Llama-3.1-8B	13.91	10.33	23.05	14.27	5.25	3.89	8.67	5.37	7.65	17.06	10.56	7.22	16.10	9.97		
	Qwen2.5-7B	23.64	21.61	31.38	25.60	9.77	8.87	12.89	10.51	15.87	23.05	18.80	17.19	24.96	20.36		
	Gemma2-9B	20.38	14.56	37.33	20.95	7.83	5.78	14.81	8.31	10.28	26.35	14.79	11.10	28.46	15.97		
	GPT-4o	47.55	53.11	48.35	50.61	24.58	27.58	25.11	26.29	43.26	39.39	41.23	46.00	41.88	43.84		
	GPT-4o-mini	35.37	48.90	32.06	38.73	14.30	20.03	13.13	15.86	37.06	24.29	29.35	40.57	26.59	32.13		
3-shot	Llama-3.1-8B	21.27	15.41	41.40	22.46	8.79	6.37	17.11	9.28	11.31	30.38	16.48	11.92	32.01	17.37		
	Qwen2.5-7B	21.63	17.01	40.78	24.00	8.80	6.73	16.15	9.51	12.01	28.80	16.95	13.25	31.77	18.70		
	Gemma2-9B	24.57	17.47	45.90	25.31	10.14	7.40	19.45	10.73	12.86	33.78	18.62	13.79	36.22	19.97		
	GPT-4o	51.59	55.26	53.14	54.18	28.33	30.19	29.04	29.60	46.89	45.09	45.97	49.03	47.15	48.07		
	GPT-4o-mini	42.07	49.73	40.11	44.40	18.20	21.57	17.39	19.26	39.22	31.62	35.01	41.47	33.45	37.03		

Table 20: Effect of element-specific relaxed-match criteria on ATE and AOS (%). For each task, **EM** reports F1 under Exact Match. **Asp-RM** denotes aspect relaxed matching, **Opi-RM** denotes opinion relaxed matching, and **Asp-RM + Opi-RM** applies both simultaneously. Sentiment polarity is always evaluated by exact match.

Model	Variant	Review-EM			Opinion-EM			Opinion-RM		
		P	R	F1	P	R	F1	P	R	F1
Llama-3.1-8B	OASIS	84.81	83.76	84.28	38.33	37.85	38.09	49.83	49.21	49.52
	Normalized	74.92	81.00	77.84	30.50	32.99	31.70	39.20	42.42	40.75
	[ia]	76.26	71.56	73.84	21.91	20.56	21.21	33.97	31.88	32.89
	NULL	75.79	73.78	74.77	24.50	23.85	24.17	36.42	35.46	35.93
Qwen2.5-7B	OASIS	87.78	85.05	86.40	38.28	37.09	37.67	50.00	48.44	49.21
	Normalized	77.81	80.76	79.26	31.68	32.90	32.28	41.53	43.15	42.33
	[ia]	76.57	75.14	75.85	24.09	23.64	23.86	34.96	34.31	34.63
	NULL	76.69	76.36	76.53	22.88	22.78	22.83	34.32	34.17	34.24
Gemma2-9B	OASIS	44.53	95.93	60.82	13.72	39.67	20.39	12.83	52.42	20.62
	Normalized	45.03	97.81	61.67	10.34	33.24	15.77	10.71	43.97	17.23
	[ia]	41.89	94.34	58.02	7.14	21.99	10.78	7.60	31.73	12.26
	NULL	41.67	95.49	58.02	7.27	22.92	11.03	9.21	30.95	14.20

Table 21: ASOR evaluation results for fine-tuned models across aspect representation variants. **OASIS** denotes the original open-vocabulary aspect representation, and **Normalized** denotes the aspect-normalized version. **[ia]** and **NULL** denote variants where implicit aspects are replaced with [ia] and NULL, respectively. ASOR is evaluated using **Review-EM**, **Opinion-EM**, and **Opinion-RM**. All scores are reported as percentages (%).

Case 1: DMASTE — Implicit aspects collapsed into [ia]		
<i>Input:</i> “[ia] This nail polish is the best! Yes, it is kind of expensive, but it does last and the colors are always fresh and beautiful.”		
	Our prediction	DMASTE gold
Tuple 1	(#satisfaction, “the best”, POS)	(nail polish, “the best”, POS)
Tuple 2	(price, “kind of expensive”, NEG)	([ia], “expensive”, NEG)
Tuple 3	(longevity, “does last”, POS)	([ia], “last”, POS)
Tuple 4	(color, “always fresh and beautiful”, POS)	(colors, “always fresh \\ <i>beautiful</i> ”, POS)
Case 2: DMASTE — Coarse and underspecified aspect annotation		
<i>Input:</i> “[ia] Amazing stuff , does not clog my pores , keeps my skin hydrated through the day and the sunscreen is + + + Would buy again .)”		
	Our prediction	DMASTE gold
Tuple 1	(#satisfaction, “Amazing stuff\\ <i>Would buy again</i> ”, POS)	(stuff, “Amazing \\ <i>does not clog my pores</i> \\ <i>keeps my skin hydrated through the day</i> ”, POS)
Tuple 2	(pores, “ <i>does not clog my pores</i> ”, NEG)	([ia], “ <i>Would buy again</i> ”, POS)
Tuple 3	(hydration, “ <i>keeps my skin hydrated through the day</i> ”, POS)	—
Tuple 4	(sunscreen, “+ + +”, POS)	—
Case 3: SemEval2014 — Only explicit aspects are annotated		
<i>Input:</i> “I use my mac all the time, I love the software, the way it takes a short time to load things, how easy it is to use and most of all how you don’t have to worry about viruses.”		
	Our prediction	SemEval2014 gold
Tuple 1	(software, “ <i>I love the software</i> ”, POS)	(software, —, POS)
Tuple 2	(loading time, “ <i>takes a short time to load things</i> ”, POS)	—
Tuple 3	(ease of use, “ <i>how easy it is to use</i> ”, POS)	—
Tuple 4	(security, “ <i>don’t have to worry about viruses</i> ”, POS)	—

Table 22: Examples of schema mismatch: DMASTE collapses implicit aspects into [ia] and underspecified aspect annotation, whereas SemEval2014 annotates only explicit aspect terms.

Task	Metric	Llama-3.1-8B			Qwen2.5-7B			Gemma2-9B		
		Matched	EM-failed	Ratio	Matched	EM-failed	Ratio	Matched	EM-failed	Ratio
ATE	CM	1,393	161	11.56%	1,272	153	12.03%	1,371	164	11.96%
	RM	1,267	26	2.05%	1,159	34	2.93%	1,239	26	2.10%
AOS	CM	1,482	793	53.51%	1,356	734	54.13%	1,461	768	52.57%
	RM	1165	476	40.86%	1048	424	40.46%	1136	443	38.99%

Table 23: EM-failure rates among prediction–gold pairs matched by CM or RM. **Matched** denotes the number of pairs accepted by each metric, and **EM-Fail** denotes the subset rejected by exact match.

Model	Train	ASC			ATE			AOS		
		OASIS	DMASTE	SemEval	OASIS	DMASTE	SemEval	OASIS	DMASTE	SemEval
Llama-3.1-8B	OASIS	97.02	<u>93.18</u>	89.15	61.63	15.55	<u>30.90</u>	34.22	<u>3.38</u>	–
	DMASTE	<u>94.32</u>	91.67	<u>91.74</u>	12.71	62.28	25.90	<u>3.11</u>	28.33	–
	SemEval	91.67	90.23	79.78	<u>15.66</u>	<u>25.79</u>	80.27	–	–	–
Qwen2.5-7B	OASIS	97.33	<u>91.64</u>	<u>87.66</u>	56.75	15.82	29.04	31.48	<u>3.04</u>	–
	DMASTE	<u>92.42</u>	91.77	85.38	<u>13.37</u>	59.51	<u>29.66</u>	<u>3.09</u>	25.09	–
	SemEval	91.83	90.98	84.61	10.86	<u>17.23</u>	68.93	–	–	–
Gemma2-9B	OASIS	96.78	<u>93.02</u>	85.91	47.76	12.09	23.16	27.28	<u>3.13</u>	–
	DMASTE	<u>92.94</u>	91.59	<u>87.07</u>	<u>13.68</u>	64.31	<u>44.57</u>	<u>3.09</u>	30.18	–
	SemEval	90.29	90.71	82.52	9.90	<u>14.83</u>	25.01	–	–	–

Table 24: Cross-dataset evaluation under the fine-tuned setting, measured with **EM**. Each model is fine-tuned on a source dataset (**Train**) and evaluated on three target datasets (**OASIS**, **DMASTE**, **SemEval2014**). ASC reports accuracy (%), while ATE and AOS report F1 (%) under EM, where every element must match the gold span exactly. **Bold** denotes in-domain (Train = Test) performance, and underline denotes the best cross-domain performance for each model in each column.

Model	Train	ATE			AOS		
		OASIS	DMASTE	SemEval	OASIS	DMASTE	SemEval
Llama-3.1-8B	OASIS	63.92	18.73	<u>35.33</u>	57.86	<u>12.47</u>	–
	DMASTE	16.26	63.48	26.54	<u>11.65</u>	54.59	–
	SemEval	<u>20.63</u>	<u>27.63</u>	82.60	–	–	–
Qwen2.5-7B	OASIS	59.97	<u>18.90</u>	<u>34.42</u>	52.86	<u>11.75</u>	–
	DMASTE	17.38	60.33	31.43	<u>11.75</u>	50.54	–
	SemEval	14.45	18.45	71.48	–	–	–
Gemma2-9B	OASIS	49.61	13.95	27.13	44.72	9.24	–
	DMASTE	<u>18.21</u>	65.37	46.69	<u>12.18</u>	56.27	–
	SemEval	13.10	<u>15.54</u>	25.54	–	–	–

Table 25: Cross-dataset evaluation under the fine-tuned setting, measured with **RM**. Each model is fine-tuned on a source dataset (**Train**) and evaluated on three target datasets (**OASIS**, **DMASTE**, **SemEval2014**). ATE and AOS report F1 (%) under RM, which counts a span as correct when its aspect semantic threshold is at least $\tau_a = 0.8$ and opinion lexical/set-level thresholds $\tau_{\text{lex}}, \tau_{\text{op}} = 0.5$; ATE applies Asp-RM and AOS applies Asp-RM and Opi-RM jointly. **Bold** denotes in-domain (Train = Test) performance, and underline denotes the best cross-domain performance for each model in each column.

Model	Train	ATE			AOS		
		OASIS	DMASTE	SemEval	OASIS	DMASTE	SemEval
Llama-3.1-8B	OASIS	69.18	<u>49.97</u>	<u>53.26</u>	73.60	<u>41.24</u>	–
	DMASTE	<u>50.67</u>	69.81	39.59	<u>38.68</u>	53.66	–
	SemEval	34.94	33.32	86.01	–	–	–
Qwen2.5-7B	OASIS	64.16	<u>48.73</u>	<u>52.15</u>	68.40	<u>39.67</u>	–
	DMASTE	<u>56.18</u>	67.90	45.02	<u>43.77</u>	49.92	–
	SemEval	26.17	21.98	74.35	–	–	–
Gemma2-9B	OASIS	53.98	<u>36.68</u>	39.28	57.52	<u>30.78</u>	–
	DMASTE	<u>55.20</u>	72.00	<u>53.97</u>	<u>42.72</u>	55.92	–
	SemEval	21.34	17.36	25.96	–	–	–

Table 26: Cross-dataset evaluation under the fine-tuned setting, measured with **CM**. Each model is fine-tuned on a source dataset (**Train**) and evaluated on three target datasets (**OASIS**, **DMASTE**, **SemEval2014**). ATE and AOS report F1 (%) under CM, where an LLM judge determines whether a prediction is semantically equivalent to the gold tuple given the review context. **Bold** denotes in-domain (Train = Test) performance, and underline denotes the best cross-domain performance for each model in each column.

Model	EM			Opi-RM ≥ 0.5			Opi-RM ≥ 0.6			Opi-RM ≥ 0.7			Opi-RM ≥ 0.8			CM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama-3.1-8B	6.03	16.20	8.79	14.86	39.91	21.66	13.47	36.18	19.63	9.97	26.78	14.53	9.13	24.53	13.31	19.23	51.65	28.03
Qwen2.5-7B	6.24	14.95	8.80	16.15	38.72	22.79	14.59	34.98	20.59	11.17	26.78	15.77	10.11	24.25	14.27	22.12	53.04	31.22
Gemma2-9B	7.00	18.40	10.14	17.51	46.00	25.36	15.99	42.02	23.17	12.29	32.30	17.80	11.40	29.95	16.51	21.94	57.64	31.78
GPT-4o	28.90	27.79	28.33	57.25	55.06	56.13	54.11	52.04	53.05	44.34	42.64	43.48	41.36	39.77	40.55	70.60	67.90	69.22
GPT-4o-mini	20.38	16.44	18.20	49.32	39.77	44.03	45.04	36.32	40.21	34.76	28.03	31.03	32.20	25.97	28.75	63.70	51.37	56.87

Table 27: AOS evaluation results when the aspect relaxed-match threshold is fixed to $\tau_a = 0.5$.

Model	EM			Opi-RM ≥ 0.5			Opi-RM ≥ 0.6			Opi-RM ≥ 0.7			Opi-RM ≥ 0.8			CM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama-3.1-8B	6.03	16.20	8.79	12.77	34.31	18.62	11.70	31.43	17.06	8.71	23.38	12.69	8.01	21.51	11.67	19.23	51.65	28.03
Qwen2.5-7B	6.24	14.95	8.80	14.47	34.69	20.42	13.13	31.48	18.53	10.15	24.34	14.33	9.19	22.04	12.97	22.12	53.04	31.22
Gemma2-9B	7.00	18.40	10.14	14.95	39.29	21.66	13.68	35.94	19.81	10.65	27.98	15.43	9.83	25.83	14.24	21.94	57.64	31.78
GPT-4o	28.90	27.79	28.33	52.57	50.55	51.54	49.48	47.58	48.51	40.51	38.96	39.72	37.52	36.08	36.79	70.60	67.90	69.22
GPT-4o-mini	20.38	16.44	18.20	44.39	35.79	39.63	40.40	32.58	36.07	31.25	25.20	27.90	28.94	23.33	25.84	63.70	51.37	56.87

Table 28: AOS evaluation results when the aspect relaxed-match threshold is fixed to $\tau_a = 0.6$.

Model	EM			Opi-RM ≥ 0.5			Opi-RM ≥ 0.6			Opi-RM ≥ 0.7			Opi-RM ≥ 0.8			CM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama-3.1-8B	6.03	16.20	8.79	12.26	32.92	17.86	11.22	30.14	16.35	8.42	22.62	12.27	7.74	20.80	11.28	19.23	51.65	28.03
Qwen2.5-7B	6.24	14.95	8.80	13.81	33.11	19.49	12.53	30.04	17.68	9.73	23.33	13.74	8.79	21.08	12.41	22.12	53.04	31.22
Gemma2-9B	7.00	18.40	10.14	14.33	37.66	20.76	13.11	34.45	18.99	10.28	27.02	14.90	9.52	25.01	13.79	21.94	57.64	31.78
GPT-4o	28.90	27.79	28.33	50.62	48.68	49.63	47.68	45.86	46.75	38.96	37.47	38.20	36.17	34.79	35.47	70.60	67.90	69.22
GPT-4o-mini	20.38	16.44	18.20	42.78	34.50	38.20	38.92	31.38	34.75	30.01	24.20	26.79	27.69	22.33	24.72	63.70	51.37	56.87

Table 29: AOS evaluation results when the aspect relaxed-match threshold is fixed to $\tau_a = 0.7$.

Model	EM			Opi-RM ≥ 0.5			Opi-RM ≥ 0.6			Opi-RM ≥ 0.7			Opi-RM ≥ 0.8			CM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama-3.1-8B	6.03	16.20	8.79	11.92	32.01	17.37	10.94	29.37	15.94	8.24	22.14	12.01	7.58	20.36	11.05	19.23	51.65	28.03
Qwen2.5-7B	6.24	14.95	8.80	13.25	31.77	18.70	12.01	28.80	16.95	9.29	22.28	13.12	8.41	20.17	11.87	22.12	53.04	31.22
Gemma2-9B	7.00	18.40	10.14	13.79	36.22	19.97	12.60	33.11	18.25	9.92	26.07	14.37	9.17	24.10	13.29	21.94	57.64	31.78
GPT-4o	28.90	27.79	28.33	49.03	47.15	48.07	46.19	44.42	45.29	37.92	36.46	37.18	35.18	33.83	34.49	70.60	67.90	69.22
GPT-4o-mini	20.38	16.44	18.20	41.47	33.45	37.03	37.79	30.47	33.74	29.17	23.53	26.05	27.04	21.80	24.14	63.70	51.37	56.87

Table 30: AOS evaluation results when the aspect relaxed-match threshold is fixed to $\tau_a = 0.8$.

Office_Products

Logout

Open Product Page

ASIN: B0008BAJW - Rating: 5 - helpful: 0 - Verified: true

Title: Filing Bulky Items



Filing Bulky Items

Just what I needed to file away Warranty books with CD's etc that are too bulky for a regular file folder. Works well.

Annotation Results

Live mode

UNDECIDED

Tuple #0

Aspect: capacity

Opinion: Just what I needed to file away Warranty books with CD's etc that are too bulky for a regular file folder

Sentiment: positive

Accept

Hold

Reject

Reset

Hold reason

Aspect span incorrect

Aspect refers to product itself

Opinion span incorrect

Sentiment label incorrect

Same aspect across tuples

Other

Hold reason is required. Suggestion is optional for "Aspect span incorrect".

Submit Hold

UNDECIDED

Tuple #1

Aspect: #satisfaction

Opinion: Works well.

Sentiment: positive

Accept

Hold

Reject

Reset

Revise

OK

Figure 4: Annotation interface used for human verification of LLM-generated candidate tuples.

16224

Case 1: EM fails, RM and CM succeed

Review: Roller does not glide as smoothly as I had hoped but it really does a great job at blotting the oily shine from my face. Small and compact about as tall as a tube of chap stick just wider. I keep it with me to use on my tzone and it really does the trick. Quicker and easier than blotting with a papertowel. The real test will come this summer. Comes with a second roller which is nice.

Predicted tuple

(oil control, "does a great job at blotting the oily shine from my face \\ really does the trick", POS)

Gold tuple

(oil control, "really does a great job at blotting the oily shine from my face \\ really does the trick", POS)

EM	Asp-RM	O@.5	O@.6	O@.7	O@.8	CM
X	1.00	✓	✓	✓	✓	✓

Case 2: Opi-RM succeeds only under lower thresholds

Review: It is my intention to try out every lotion, body butter and cream out there, so when I saw this was available from Vine, I requested it and here we are! This lotion is decadent-feeling. It's thicker than I thought it would be, but absorbs easily and moisturizes very well. The scent is citrus'y but it's a light scent, it's not overpowering. My only issue with this product is the bottle. I have arthritis in my fingers/hands/wrists and this bottle is super hard to squeeze. Maybe it's difficult for me due to the arthritis, but I don't think so. When I asked my husband to put some in my hand, even he said that the bottle is hard to squeeze. I absolutely love this lotion, but I doubt I would buy more, because it's a real bummer to have a lotion that you can't dispense yourself.

Predicted tuple

(scent, "citrus'y \\ light scent \\ not overpowering", POS)

Gold tuple

(scent, "citrus'y \\ light \\ not overpowering", POS)

EM	Asp-RM	O@.5	O@.6	O@.7	O@.8	CM
X	1.00	✓	✓	X	X	✓

Case 3: Asp-RM fails, CM succeed

Review: (Reviewed by my wife): This is a very interesting cleanser - I've never tried anything like it. It does not foam/lather at all. Instead, it is so thick it has the consistency of lotion. It is gentle but cleans effectively. It leaves my skin feeling clean and soft.

Predicted tuple

(softness, "leaves my skin feeling clean and soft", POS)

Gold tuple

(afterfeel, "leaves my skin feeling clean and soft", POS)

EM	Asp-RM	O@.5	O@.6	O@.7	O@.8	CM
X	0.15	✓	✓	✓	✓	✓

Case 3: Asp-RM fails, CM succeed

Review: This gel polish has a nice collection of glitter gels. They are shiny and sparkly. I worked with mine while they were at room temperature and did not have any trouble with thickness. I found them to be easy to apply and to cure quickly. I do wish that they had come with a base and a top coat. While there is no colored label on the bottles, they do provide a set of stickers to make your own.

Predicted tuple

(shine, "shiny and sparkly", POS)

Gold tuple

(appearance, "shiny and sparkly", POS)

EM	Asp-RM	O@.5	O@.6	O@.7	O@.8	CM
X	0.40	✓	✓	✓	✓	✓

Case 4: Sentiment mismatch

Review: This did a good job on my hair. The shampoo was a bit difficult to get out as it produces more suds than I care for. The scent was wonderful, and I wasn't allergic to it which was nice. The shampoo did a nice job of moisturizing the length. It seems a bit pricey to see since it isn't a salon type shampoo. It probably isn't though since the price includes conditioner. I rarely buy conditioner because I can't use it on my scalp, and my hair is oily. I only use it in the middle of winter when my hair gets a bit dry..It did do a great job on my hair.

Predicted tuple

(price, "a bit pricey", NEG)

Gold tuple

(price, "a bit pricey \\ price includes conditioner", NEU)

EM	Asp-RM	O@.5	O@.6	O@.7	O@.8	CM
X	1.00	✓	✓	X	X	✓

Case 5: Wrong aspect-opinion pairing

Review: This face wash is so exfoliating and smells great. It works really well on the beard area and preps the skin for a nice, super smooth shave. It doesnt seem to strip the face of natural oils, which is great. Overall, it just leaves a really soft, clean, oil-free (the bad oils) look and feel. I would definitely recommend this for a nice clean face, especially for men with bushy beards.

Predicted tuple

(exfoliation, "so exfoliating", POS)

Gold tuple

(scent, "smells great", POS)

EM	Asp-RM	O@.5	O@.6	O@.7	O@.8	CM
X	0.23	X	X	X	X	X

Table 31: Qualitative examples comparing EM, RM, and CM for AOS tuple evaluation. Asp-RM denotes similarity scores for aspect spans. O@.5–O@.8 indicate whether the prediction is counted as correct when the opinion threshold $\tau_{\text{lex}}, \tau_{\text{op}}$ varies from 0.5 to 0.8.

LLM Annotation Prompt for Aspect–Opinion–Sentiment Extraction
Task You are an expert annotator for Aspect-Based Sentiment Analysis (ABSA), specializing in real-world e-commerce product reviews. You are given a product review and its Amazon product category. Your task is to extract all valid (<i>aspect</i>, <i>opinion</i>, <i>sentiment polarity</i>) tuples from the review. Sentiment polarity must be one of: <i>positive</i> <i>negative</i> <i>neutral</i>. Aspect An aspect MUST be a concrete, fine-grained noun referring to a specific component, property, or usage-related target (e.g., nib, weight, size, fit, balance, coverage, color). Prefer a single, singular noun whenever possible. Do NOT use generic placeholders such as “product,” “item,” “this,” or “it.” If a feature is implicit, infer the aspect from the review context with the help of the category information. Object-level product nouns (e.g., stand, pen, toy, castle) MAY be used as aspects when the opinion evaluates a concrete physical, functional, or usage-related property of the product. Use <i>#satisfaction</i> only for overall sentiment not tied to a specific feature. Opinion Preserve the original surface form of the opinion as much as possible. Use the minimal span that captures the evaluation. Do NOT paraphrase explicit sentiment expressions. Light abstraction is allowed only when the evaluation is clearly implied but not directly expressed. Aspect–Opinion Alignment Rules 1. Experience-Priority Rule Annotate ONLY the reviewer’s own experience. Do NOT annotate hypothetical, conditional, or third-person evaluations if a self-experienced evaluation exists for the same aspect. 2. Consequence Exclusion Rule Do NOT annotate consequences or outcomes as separate aspects when they are only used to explain another feature. 3. Same-Aspect Consolidation Rule Each aspect must appear only once. Merge all related opinions into one tuple and assign a single overall sentiment. When multiple opinion spans are merged, concatenate them using a double backslash (\\) in their original surface form. Output Format Return the results as a list of tuples in the following format: (<i>aspect</i>: "<aspect>", <i>opinion</i>: "<opinion>", <i>sentiment polarity</i>: "<positive negative neutral>")

Table 32: Prompt used for LLM-based Aspect–Opinion–Sentiment tuple extraction.

Human Annotation Guideline
Task Description This study constructs an Aspect–Opinion–Sentiment (AOS) annotated dataset from user-generated product reviews. Each annotation consists of a triple (<i>Aspect</i>, <i>Opinion</i>, <i>Sentiment</i>) defined as follows. Aspect An Aspect refers to a product’s function, characteristic, or component that is being evaluated in the review. An Aspect must be naturally interpretable within the review context and must be grounded in the reviewer’s actual usage experience. An Aspect may appear explicitly as a word or phrase directly mentioned in the text, or implicitly when the evaluation target can be reasonably inferred from contextual evidence. However, implicit Aspects must not be assigned without textual justification. Opinion An Opinion is the textual span that expresses the reviewer’s subjective evaluation toward a specific Aspect. The Opinion must be present explicitly in the original text and based on the reviewer’s actual experience or judgment rather than speculation. Sentiment Sentiment represents the polarity of the Opinion toward the Aspect. Each (Aspect, Opinion) pair is assigned one of three labels: Positive (POS), Negative (NEG), or Neutral (NEU). POS indicates a strong positive evaluation of the Aspect, NEG indicates a strong negative evaluation, and NEU indicates a weak, mixed, or neutral evaluation. NEU is also used when conflicting evaluations about the same Aspect appear in the review. During verification, annotators evaluate each tuple and assign one of three decisions: Accept, Hold, or Reject. Accept is assigned when the tuple fully satisfies all annotation criteria. Hold is selected when the tuple violates one or more rules but can be corrected through modification. Reject is reserved for cases where the tuple cannot be corrected and should be removed entirely. Reject is used only when the Opinion does not exist in the text, when the statement does not express an evaluation, or when the Opinion refers to something unrelated to the product’s features, functions, or components. Opinion Existence Check - Does the Opinion span appear explicitly in the original text? Aspect Existence & Aspect–Opinion Alignment Check - Is the Aspect word contextually natural and semantically appropriate? - Does the Aspect accurately refer to the feature, function, or component being evaluated? - Does the Opinion describe the Aspect’s function, characteristic, or usage experience? - Is the evaluation grounded in the reviewer’s actual experience? Non-Product Aspect Exclusion Rule - Do NOT annotate evaluations related to <i>shipping, delivery, seller service, customer support, or other post-purchase logistics</i>. - Only product-intrinsic features, functions, or components should be considered valid Aspects. Tuple Deduplication and Consolidation Check - Are there duplicate tuples? - Cases where a single conceptual Aspect is split into multiple tuples - Multiple tuples describing the same Aspect separately Sentiment Check - Is the assigned Sentiment supported by explicit evidence in the Opinion? - If conflicting evaluations appear for the same Aspect, were they merged into NEU? - Was a weak or mixed evaluation incorrectly labeled as strongly POS or NEG?

Table 33: Human Annotation Guideline.

Fine-tuning Prompt for Aspect-Based Sentiment Analysis
Task You are an expert in Aspect-Based Sentiment Analysis (ABSA). Given a review, extract all sentiment triplets in the format: <i>(Aspect, Opinion, Sentiment)</i>. Aspect A word or short phrase representing a property, feature, or characteristic of the product that the reviewer is evaluating (e.g., <i>battery life, price, screen, build quality</i>). Aspects can be either: Explicit: the aspect is directly named in the review (e.g., “<i>The screen is bright</i>” → aspect: <i>screen</i>; “<i>The color is beautiful</i>” → aspect: <i>color</i>). Implicit: the aspect is not literally stated but clearly implied by the opinion (e.g., “<i>It broke after a week</i>” → aspect: <i>durability</i>; “<i>Fits perfectly in my hand</i>” → aspect: <i>size</i>; “<i>I love the golden finish</i>” → aspect: <i>color</i>). If the sentiment expresses overall satisfaction or dissatisfaction with the product rather than a specific attribute, use <i>#satisfaction</i> as the aspect. Opinion The opinion expression in the review that describes the aspect. Use the exact phrase or span from the review text. If multiple opinion phrases describe the same aspect, combine them with “ \ ” (e.g., “<i>great screen \ very bright</i>”). Sentiment The polarity of the opinion toward the aspect. Must be one of: POS (positive), NEG (negative), NEU (neutral or mixed). Output Format Return a Python list of tuples: <code>[('aspect', 'opinion', 'SENTIMENT'), ...]</code> If no aspect-sentiment triplets can be extracted, return an empty list: <code>[]</code>. Do not include any explanation, only the list. Input Review: {review} Extract all triplets:

Table 34: Prompt used for fine-tuning open-source LLMs on Aspect-Opinion-Sentiment triplet extraction.

3-shot Prompt for Aspect-Opinion-Sentiment Extraction

Task

You are an expert in Aspect-Based Sentiment Analysis (ABSA). Given a review, extract all sentiment triplets in the format: (*Aspect*, *Opinion*, *Sentiment*).

Aspect

A word or short phrase representing a property, feature, or characteristic of the product that the reviewer is evaluating (e.g., *battery life*, *price*, *screen*, *build quality*). Aspects can be either:

Explicit: the aspect is directly named in the review

(e.g., “*The screen is bright*” → aspect: *screen*).

Implicit: the aspect is not literally stated but clearly implied by the opinion

(e.g., “*It broke after a week*” → aspect: *durability*).

If the sentiment expresses overall satisfaction or dissatisfaction with the product rather than a specific attribute, use *#satisfaction* as the aspect.

Opinion

The opinion expression in the review that describes the aspect. Use the exact phrase or span from the review text. If multiple opinion phrases describe the same aspect, combine them with “ \ ” (e.g., “*great screen \ very bright*”).

Sentiment

The polarity of the opinion toward the aspect. Must be one of:

POS (positive), **NEG** (negative), **NEU** (neutral or mixed).

Output Format

Return a Python list of tuples: [(*‘aspect’*, *‘opinion’*, *‘SENTIMENT’*), ...]. If no triplets can be extracted, return an empty list. Do not include any explanation, only the list.

Example 1

Review: After a few years of heavy use, the earpads on my headset started splitting open. Since the headset still worked great, I decided to try these replacement earpads. Taking the old earpads off and putting the new ones on was really easy. Afterward it was almost like having a brand new headset! After about a year, one of the earpads has started to split. They are still usable, but I expected them to last longer than 12 months.

Triplets: [(*‘installation’*, *‘really easy’*, *‘POS’*), (*‘#satisfaction’*, *‘almost like having a brand new headset’*, *‘POS’*), (*‘durability’*, *‘After about a year, one of the earpads has started to split \ expected them to last longer than 12 months’*, *‘NEG’*)]

Explanation: *installation* is an implicit aspect (the reviewer describes the installation process without naming it). *#satisfaction* captures overall satisfaction not tied to a specific feature. *durability* is implicit, with two phrases describing the same aspect combined using \.

Example 2

Review: This nail polish remover doesn’t remove nail polish easily. I had to put a lot of pressure on and go over a nail multiple times. However, it hydrates my cuticles and makes my nails look shiny. It didn’t dry my nails – my nails look like I just applied cuticle oil. It smells really good too. I would still prefer this remover.

Triplets: [(*‘removal’*, *“doesn’t remove nail polish easily \ had to put a lot of pressure on and go over a nail multiple times”*, *‘NEG’*), (*‘functionality’*, *‘hydrates my cuticles’*, *‘POS’*), (*‘shine’*, *‘makes my nails look shiny’*, *‘POS’*), (*‘dryness’*, *“didn’t dry my nails”*, *‘POS’*), (*‘scent’*, *‘smells really good’*, *‘POS’*), (*‘#satisfaction’*, *‘I would still prefer this remover’*, *‘POS’*)]

Explanation: *removal* is explicit (core function directly criticized). *functionality*, *shine*, *dryness* are implicit aspects inferred from described effects. *scent* is explicit. *#satisfaction* captures overall preference.

Example 3

Review: towel is a little thin but what do you expect for \$8. it’ll work just fine

Triplets: [(*‘thickness’*, *‘a little thin’*, *‘NEG’*), (*‘price’*, *‘what do you expect for \$8’*, *‘NEU’*), (*‘functionality’*, *“it’ll work just fine”*, *‘POS’*)]

Explanation: *thickness* and *price* are implicit. *price* is annotated as NEU because the reviewer neither praises nor condemns it but uses it to set expectations. *functionality* is implicit (“work just fine” implies adequate basic function).

Input

Review: {review}

Extract all triplets:

Table 35: 3-shot prompt used for evaluation, including task definition and three demonstration examples with explanations.

Contextual Match (CM) Judgment Prompt
System Prompt You are an expert evaluator for aspect-based sentiment analysis. For each predicted-vs-ground-truth triplet pair, return two boolean judgments: aspect_match: true if the aspects refer to the same target, feature, or entity. aos_match: true if the aspect, opinion, and sentiment all match semantically. Rules For aspect_match, allow paraphrases, synonyms, and minor wording differences. For aos_match: the aspects must refer to the same target, feature, or entity; the opinions must express the same specific meaning, not just a related topic; and the sentiment labels must be the same. Do not require exact string match. Return false when the opinion is weaker, broader, narrower, or materially different in meaning. Do not return explanations. Output Format Return JSON only. Do not include any introduction, explanation, notes, or commentary. Do not describe the pairs. The response must be a JSON array. Each item must contain exactly: pair_id, aspect_match, aos_match. User Prompt Judge the following triplet pairs and return JSON only. {pair_items}

Table 36: Prompt used for the CM criterion. A judge LLM receives each predicted–gold triplet pair and returns semantic-equivalence judgments for the aspect alone (aspect_match) and for the full triplet (aos_match).

ASOR Prompt for Aspect-Sentiment-based Review-Opinion Extraction
Task You are an expert in Aspect-Based Sentiment Analysis (ABSA). Given an <i>Aspect</i> and a <i>Sentiment</i>, find the review(s) that are expected to contain that aspect-sentiment pair, and extract the <i>Opinion</i> expression for the pair from each matching review. Definitions Aspect: A word or short phrase representing a product property, feature, or characteristic being evaluated. Explicit: the aspect is directly named in the review. Implicit: the aspect is not literally stated but is clearly implied by the opinion expression. If the sentiment expresses overall satisfaction or dissatisfaction with the product rather than a specific attribute, the aspect may be <i>#satisfaction</i>. Sentiment: The polarity toward the aspect. Must be one of: POS (positive), NEG (negative), NEU (neutral or mixed). Opinion: The exact phrase or span in the review that expresses the given sentiment toward the given aspect. Task Steps 1. Read the target Aspect and Sentiment. 2. Read all candidate reviews. 3. Select only the candidate review(s) where the target Aspect-Sentiment pair is expected to exist. 4. For each selected review, extract the Opinion expression that supports the target Aspect-Sentiment pair. Important Rules Match the meaning of the aspect, not only the exact word. For example, “<i>broke after a week</i>” can imply the aspect <i>durability</i>. Do not select a review only because it contains the aspect word; the review must express the target sentiment toward that aspect. Do not select a review if the sentiment toward the aspect is different from the target sentiment. The opinion must be copied from the review text as exactly as possible. If multiple opinion phrases in the same review support the same target Aspect-Sentiment pair, combine them with “ \ \ ”. If no candidate review matches the target Aspect-Sentiment pair, return an empty list: <i>[]</i>. Output Format Return a Python list of tuples: <i>[(‘review_id’, ‘opinion’), ...]</i>. Use the exact <i>review_id</i> shown in the candidate reviews. Do not include the aspect or sentiment in the output. Do not include any explanation, markdown, or extra text. Example Aspect: <i>durability</i> Sentiment: NEG Candidate reviews: [r1] Taking the old earpads off and putting the new ones on was really easy. Afterward it was almost like having a brand new headset. [r2] After about a year, one of the earpads has started to split. They are still usable, but I expected them to last longer than 12 months. [r3] The color matches my headset perfectly and the material feels soft. Answer: <i>[(‘r2’, ‘After about a year, one of the earpads has started to split \ \ expected them to last longer than 12 months’)]</i> Input Aspect: {aspect} Sentiment: {sentiment} Candidate reviews: {reviews} Answer:

Table 37: Prompt used for the ASOR (Aspect–Sentiment-conditioned Opinion Retrieval) task, where the model selects reviews containing a given aspect-sentiment pair and extracts the corresponding opinion span.