Can Large Language Models be Effective Online Opinion Miners?

Ryang Heo   Yongsik Seo   Junseong Lee   Dongha Lee[†]
Yonsei University
{ryang1119, ysseo, brulee, donalee}@yonsei.ac.kr

Abstract

The surge of user-generated online content presents a wealth of insights into customer preferences and market trends. However, the highly diverse, complex, and context-rich nature of such content poses significant challenges to traditional opinion mining approaches. To address this, we introduce **O**nline **O**pinion **M**ining **B**enchmark (**OOMB**), a novel dataset and evaluation protocol designed to assess the ability of large language models (LLMs) to mine opinions effectively from diverse and intricate online environments. OOMB provides extensive *(entity, feature, opinion)* tuple annotations and a comprehensive opinion-centric *summary* that highlights key opinion topics within each content, thereby enabling the evaluation of both the extractive and abstractive capabilities of models. Through our proposed benchmark, we conduct a comprehensive analysis of which aspects remain challenging and where LLMs exhibit adaptability, to explore whether they can effectively serve as opinion miners in realistic online scenarios. This study lays the foundation for LLM-based opinion mining and discusses directions for future research in this field[1].

1 Introduction

The explosive growth of user-generated content has fundamentally transformed marketing strategies and business decision-making. Companies now analyze vast amounts of user opinions scattered across platforms such as social media, review sites, and online communities to understand how consumers truly perceive their products and services (Rahayu et al., 2021; Chen et al., 2022). As a result, *opinion mining*—the task of extracting and analyzing opinions from online text—has become a core capability in today's data-driven landscape.

[†] Corresponding author
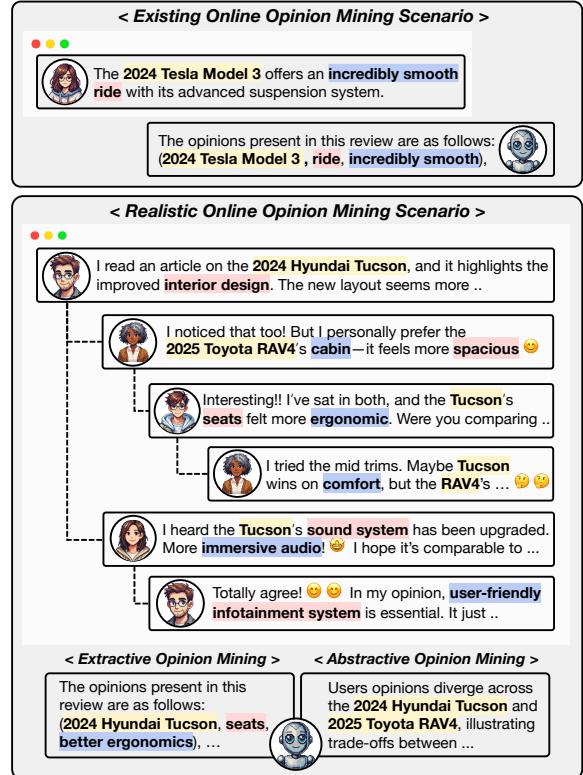[1] https://github.com/ryang1119/Online-Opinion-Mining

Figure 1: Existing opinion mining scenarios assume a simple input structure (**Upper**). In contrast, our study facilitates both extractive and abstractive opinion mining in complex, multi-threaded web discussions, enabling flexible and context-aware mining (**Lower**).

Existing opinion mining approaches have primarily focused on identifying and extracting opinion expressions or spans within text (İrsoy and Cardie, 2014; Xia et al., 2021; Li et al., 2022; Zhang et al., 2022a,b). Over time, these methods have evolved to incorporate sentiment analysis (Zhao et al., 2020; Zhang et al., 2021; Seo et al., 2024), allowing for a deeper understanding of user preferences.

Despite these advances, existing approaches still face two critical limitations. **(1) Underrepresentation of real-world input complexity**: Previous benchmarks predominantly focus on single-sentence reviews (Peng et al., 2019; Cai et al., 2021;

Benchmark	#Test Examples	Avg #Tokens	Avg #Tuples	Tuple Components	Content Types	Task Ext.	Task Abs.
ASTE (Peng et al., 2019)	1,468	15.7	1.7	(a, o, s)	Reviews	✓	✗
ACOS (Cai et al., 2021)	1,399	15.2	1.5	(a, c, o, s)	Reviews	✓	✗
ASQP (Zhang et al., 2021)	1,081	14.9	1.5	(a, c, o, s)	Reviews	✓	✗
DiaASQ-EN (Li et al., 2022)	100	179.7	8.5	(t, a, o, s)	Conversation	✓	✗
OOMB (Ours)	600	648.7	14.4	(e, f, o)	Reviews, Blogs, Conversation	✓	✓

Table 1: A comparison of our benchmark to existing opinion related benchmarks. Each tuple component represents the following: *a*: aspect, *c*: aspect category, *o*: opinion, *s*: sentiment, *t*: target, *e*: entity, and *f*: feature.

Zhang et al., 2021) or preprocessed dialogue scenarios (Li et al., 2022). However, in real-world online environments, user opinions appear in far more complex and structurally diverse formats. In practice, opinion streams span multi-party threaded discussions, long-form narratives with interleaved pros/cons, and domain-specific markers (e.g., emojis, slang, abbreviations) that introduce implicit sentiment signals (Figure 1 Lower). The absence of a setting that comprehensively captures these realistic and diverse forms of opinion expression makes **it difficult to assess under what conditions and to what extent large language models (LLMs) can effectively perform opinion mining.** This gap poses a significant challenge to evaluating the utility of LLMs and understanding their applicability to real-world applications.

(2) Confinement to extraction-centric tasks: As mentioned earlier, most prior tasks have focused on extracting opinion spans or structured tuples from input texts. However, this extraction-centric approach can excessively simplify or compress the nuanced contextual information and emotional nuances that are essential for strategic decision-making. For instance, the tuple (*"Tesla Model 3"*, *"interior"*, *"larger than the previous model"*) fails to capture critical contextual background—such as whether the user inspected the vehicle in person or harbored an implicit purchase intent. In real-world industry settings, marketers and product teams are more interested in cohesive, topic-level insights rather than isolated fragments of information (Yuan et al., 2015; Santos and Gonçalves, 2021; Han et al., 2023b). These observations highlight the need to explore opinion mining paradigms that **move beyond raw extraction and aim to preserve the emotions, contextual subtleties, and user intent embedded** in real-world discourse.

To address these challenges, we propose **O**nline **O**pinion **M**ining **B**enchmark, named **OOMB**, a novel benchmark specifically tailored to evaluate the opinion mining capabilities of LLMs across

realistic, complex, and diverse online scenarios. Unlike previous datasets, OOMB incorporates content from structurally distinct platforms—including blogs, review sites, Reddit threads, and YouTube comments—capturing long-form content, single & multi-user interactions representative of authentic online discussions. Each content instance is enriched with dual-layer annotations: (1) structured sets of (*entity, feature, opinion*) tuples reflecting explicit user perspectives, and (2) context-rich, opinion-centric *summaries* organized around key thematic insights from a marketer's viewpoint.

Building upon this benchmark, we introduce two complementary tasks: **(1) Feature-centric opinion extraction (FOE)** evaluates whether LLMs can accurately extract structured opinions from online content and **(2) Opinion-centric insight generation (OIG)** assesses whether LLMs can mine high-level topics and insights from user opinions expressed in online content. We conduct extensive experiments on ten proprietary and open-source LLMs to provide an in-depth analysis of their respective capabilities and limitations. The evaluation results demonstrate that while the models struggle with extracting structured opinions from online content, they exhibit relatively strong adaptability in synthesizing diverse user opinions into meaningful insights. Based on these findings, we discuss key takeaways and potential future directions to further advance the field of opinion mining. Specifically, our contributions are as follows:

- We present OOMB, a realistic and richly annotated benchmark that evaluates LLMs across structurally diverse online content using both structured tuples and insight-oriented summaries.

- We define two complementary tasks FOE and OIG—to jointly assess extraction and abstraction capabilities of LLMs from diverse online content.

- We extensively evaluate both proprietary and open-source LLMs, highlighting their strengths, limitations, and opportunities for further work.
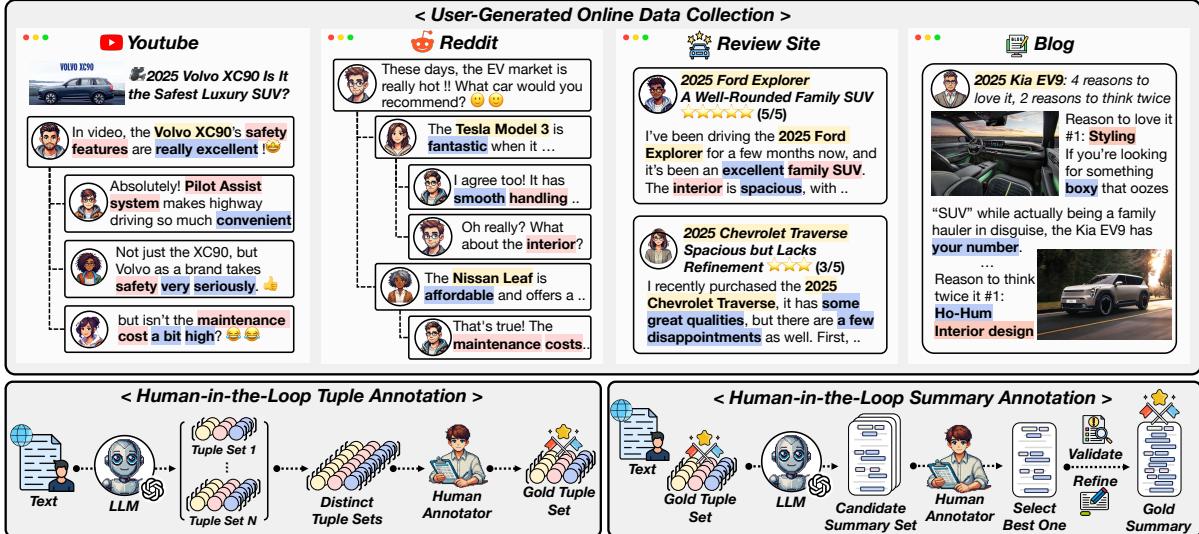
Figure 2: The overview of our OOMB benchmark construction pipeline.

2   OOMB Benchmark

In this section, we introduce the construction of **OOMB**, a benchmark designed to effectively represent real-world online content. Figure 2 illustrates the overall construction pipeline.

2.1   Data Collection

To reflect realistic user-generated content and a wide range of online structures, we collect textual data from four different sources: *Blog*, *Review Site*, *Reddit*, and *YouTube*. Blog and review site provide detailed long-form posts and specific car reviews, while reddit and youtube capture multi-threaded and single-threaded discussions, respectively. Specifically, we curate the sources of each website from Feedspot,[2] a platform that organizes and manages content across various topics. More details about our data collection process and sources are provided in Appendix A.1.

2.2   Data Annotation

For each collected user-generated content, we construct a dual-layer annotation for each content, consisting of both structured opinion tuples and free-form opinion-centric summaries. Following recent studies demonstrating that LLMs with advanced reasoning capabilities can serve as effective tools for data annotation (He et al., 2024; Tan et al., 2024), we adopt a human-in-the-loop process in which LLM is first used as the initial annotator, and human annotators verify and refine them to ensure high-quality, reliable labels. The detailed annotation process is described in Appendix A.2.

---

[2] https://www.feedspot.com/

Entity-feature-opinion tuple   We annotate each content with structured *(entity, feature, opinion)* tuples that capture user perspectives. In contrast to the commonly used (aspect, category, opinion, sentiment) schema, this design more closely reflects how real users express opinions—typically without explicit category or sentiment labels, but rather through direct mentions of entity features. Specifically, *entity* refers to the specific subject or object under discussion in the content (e.g., "Volvo XC90"). *feature* indicates a characteristic, attribute, or component of the entity that a user mentions or evaluates (e.g., "interior design"). *opinion* represents the subjective or objective judgment, reaction, experience, evaluation, or feedback regarding a feature (e.g., "luxurious"). If a feature is implicit and does not appear explicitly in the text, it is labeled as "NULL" following (Cai et al., 2021). In all other scenarios, each component of the tuple is assumed to be an explicitly mentioned span in the content.

Tuple annotation   To maximize the coverage and diversity of (entity, feature, opinion) tuple sets from each input content, we perform five rounds of zero-shot prompting using GPT-4o-mini. We then take the union of all generated tuples and remove duplicates to form a distinct preliminary tuple sets. Subsequently, five trained human annotators review every candidate tuple for correctness, eliminate hallucinated entries, and complement any missing tuples. To support consistent decision-making, we design detailed task-specific annotation guidelines and conduct a one-week training session for all annotators, including case-based instruction and edge-case discussions. This process was applied

to every content, thereby ensuring high coverage, consistency, and reliability in the final annotations.

Opinion-centric summary We annotate each content with an opinion-centric summary, a free-form text that organizes diverse opinions into high-level topics for meaningful insights. Specifically, from a marketing manager's perspective, opinions are grouped into broad categories, highlighting frequently mentioned or standout aspects to reveal key trends. This summary follows a *three-to-five-line form*, providing a cohesive structure for clear and concise representation of core discussions.

Summary annotation We generate five independent candidate summaries using the input content and the associated final set of tuples as input. Then, the same five human annotators review each summary from the perspective of a marketing manager and select the highest-quality one that best captures the opinions in the final tuples at the topic level. Similar to the tuple annotation process, we design detailed annotation guidelines to ensure consistent decision-making, and all five annotators undergo a one-week training session. If all candidate summaries are deemed insufficient in quality, the annotators collaboratively rewrite a new summary that more accurately reflects the key insights. For the selected summary, the annotators collaboratively refine and finalize it by checking for missing opinions, eliminating hallucinations, and ensuring conciseness in a three-to-five-line format.

2.3 Statistics and Analysis

As shown in Table 1, unlike previous benchmarks, OOMB features substantially longer average token lengths and a significantly higher number of tuples, making it considerably more challenging. Additionally, it covers a broader and more diverse range of content types while supporting two tasks: extraction and abstraction. This dual-task setup enables the evaluation of LLMs in more realistic settings by reflecting the complexity and variability of real-world opinion expressions. Detailed our benchmark statistics are presented in Appendix A.3.

3 Experiments

3.1 Feature-centric opinion extraction (FOE)

Problem formulation Similar to existing opinion mining approaches (Fan et al., 2019; Xia et al., 2021), this task aims to enable LLMs to accurately identify and extract structured set of opinion tuples from the given input content. Formally, given a content c, our goal is to identify and extract a set of tuples $\mathcal{T} = \{(e_i, f_i, o_i)\}_{i=1}^{N}$, where e_i represents the entity, f_i the feature, and o_i the opinion.

Evaluation protocol To evaluate structured opinion extraction capabilities of LLMs, we utilize three types of tuple matching evaluation methods. **(1) Exact Match (EM)**: Consistent with existing opinion-related extraction tasks (Zhang et al., 2022a; Xia et al., 2021), a predicted tuple is considered correct only if all its elements exactly match the corresponding elements in the gold tuple. **(2) Relaxed Match (RM)**: To provide a more flexible evaluation beyond strict exact matching, we evaluate the similarity of each tuple component using both lexical and semantic matching. A tuple is considered a relaxed match if the similarity score of all its components exceeds a predefined threshold of 0.7, formally defined as:

$$\text{RM}(t_p, t_g) = \begin{cases} 1, & \text{if } \forall x \in \{e, f, o\}, \ \text{Sim}(x_p, x_g) \geq 0.7 \\ 0, & \text{otherwise} \end{cases}$$

where $t_p = (e_p, f_p, o_p)$ and $t_g = (e_g, f_g, o_g)$ are the predicted and gold tuples, respectively. Drawing on recent works (Han et al., 2023a; Li et al., 2024), we utilize the Python's difflib library[3] to compute token-level overlap scores for lexical similarity (L-RM), while employing a Sentence Transformer[4] for semantic similarity (S-RM). **(3) Contextual Match (CM)**: Inspired by (Fu et al., 2023; Fane et al., 2025), we design a method that leverages the reasoning capabilities of LLMs to match tuples in a manner similar to human judgment. Specifically, we utilize GPT-4o to evaluate both predicted and gold tuples, enabling the model to count how many tuples match. This metric allows recognition of semantically equivalent tuples even when surface forms differ significantly, used the prompt shown in Table 17. Note that for both RM and CM, we measure recall by counting each gold tuple at most once to avoid double counting, even if multiple predicted tuples match the same gold tuple. For all evaluation metrics, we primarily use the F1 score while also reporting precision and recall.

3.2 Opinion-centric insight generation (OIG)

Problem formulation This task aims to analyze whether LLMs can group scattered opinions from

---

[3] https://docs.python.org/3/library/difflib.html
[4] We use `all-MiniLM-L6-v2`, a lightweight model optimized for efficient sentence similarity computation.

Models	EM			L-RM			S-RM			CM		
	Pre	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Proprietary LLMs												
GPT-4o-mini	3.91	1.99	2.62	11.29	5.70	7.52	15.50	7.76	10.27	**65.07**	36.61	43.19
GPT-4o	7.27	5.18	6.02	15.86	11.39	13.20	21.23	15.31	17.71	59.34	**45.88**	**48.28**
Claude-3.5-Haiku	6.13	3.02	4.01	15.02	7.51	9.94	20.60	10.36	13.68	63.70	37.25	43.62
Claude-3.5-Sonnet	**11.12**	**6.32**	**7.97**	**22.97**	**13.02**	**16.46**	**29.30**	**16.52**	**20.93**	62.00	39.83	44.90
Open-source LLMs												
Llama3-8B-Instruct	**8.49**	**6.28**	**7.17**	**16.75**	**12.18**	**14.02**	**21.33**	15.43	**17.80**	51.92	42.51	43.18
Llama3-70B-Instruct	7.26	5.57	6.27	15.21	11.66	13.13	19.94	15.18	17.15	53.15	42.55	**43.67**
Gemma2-9B-it	6.37	4.51	5.25	14.17	10.17	11.78	17.73	12.59	14.64	**53.71**	41.93	43.61
Gemma2-27B-it	7.05	5.61	6.20	14.33	11.77	12.82	19.29	**15.83**	17.25	52.58	**42.98**	43.42
Qwen2.5-7B-Instruct	6.55	4.18	5.05	12.97	8.32	10.02	18.22	11.83	14.18	52.48	39.15	41.14
DeepSeek-7B-chat	3.00	1.63	2.07	5.86	3.13	4.02	8.20	4.34	5.61	49.25	30.33	33.12

Table 2: Performance comparison of various LLMs for the FOE task across diverse tuple matching metrics.

user-generated online content into high-level topics, providing context-aware and meaningful insights. Formally, given content c, our objective is to generate a free-form text summary S that cohesively encapsulates user opinions into high-level topics.

Evaluation protocol To broadly assess the quality of opinion insight summaries generated by the model across various aspects, we employ both lexical and semantic automated evaluation metrics. For lexical evaluation, we adopt **ROUGE-1,10, L** (Lin, 2004), which measure word overlap between the reference and generated summaries. For semantic evaluation, we leverage **BERTScore (BS)** (Zhang et al., 2019) and **A3CU** (Liu et al., 2023b). BS computes similarity between the reference and generated texts using contextual embeddings, while A3CU compares texts without extracting atomic content units, providing a human-aligned assessment of content similarity. For both ROUGE and A3CU, we report F1 scores. Moreover, to ensure a systematic and comprehensive evaluation, we also conduct reference-free assessments using an LLM as the judge. Inspired by (Siledar et al., 2024), we design the following six well-defined criteria: *Faithfulness, Coverage, Specificity, Insightfulness, Intent* and *Fluency*. This analysis extends beyond automated lexical and semantic metrics, providing a broader perspective on the abstractive opinion mining capabilities of LLMs. A detailed description is provided in Appendix B.3.

3.3 Experimental setup

Models We conduct extensive experiments on two types of LLMs: **(1) Proprietary LLMs** that are available via APIs, such as GPT-4o-mini, GPT-4o (OpenAI et al., 2024), and Claude 3.5 Haiku, Sonnet (Anthropic, 2024). **(2) Open-source LLMs**

Metric	Pearson r	Spearman ρ	Kendall τ
EM	0.4505	0.4722	0.4215
L-RM	0.4584	0.4754	0.4244
S-RM	0.5514	0.5531	0.4937
CM	**0.8337**	**0.8155**	**0.7279**

Table 3: Correlation coefficients between each metric and human judgment (*p-value* < 0.05) based on pairwise comparisons by five human evaluators. Detailed experimental settings are provided in Appendix B.2.

such as Llama-3-Instruct (8B, 70B, Grattafiori et al. 2024), Gemma 2-it (9B, 27B, Team et al. 2024b), Qwen2.5-7B-Instruct (Yang et al., 2024), and DeepSeek-7B-chat (Bi et al., 2024).

Implementation details Following recent studies demonstrating the reasoning capabilities of LLMs in zero-shot settings (Wang et al., 2024; Qin et al.; Liu et al., 2024; Chhabra et al., 2024), we perform both tasks using zero-shot prompting. This means the models rely solely on their pretrained knowledge without any task-specific finetuning. To ensure consistent and reliable performance across all experiments, we set the temperature to 0 for all generations. Our more detailed experimental setup presented in Appendix B.

4 Results and Discussion

In this section, we present the main findings of our study. Each subsection addresses the research question—*Can LLMs serve as effective online opinion miners?*—from various perspectives, supported by detailed experimental results and analyses.

4.1 RQ1: What makes it challenging for LLMs to extract structured opinions?

Performance on tuple extraction As shown in Table 2, LLMs consistently struggle to extract
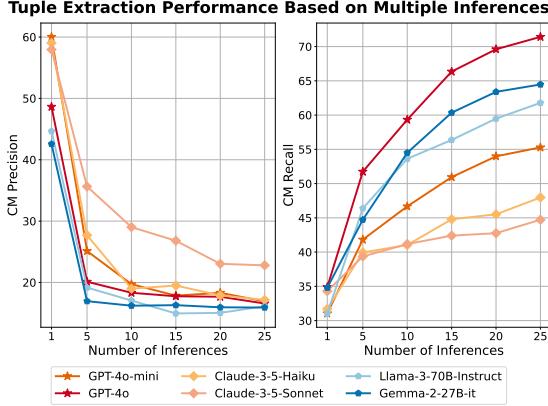
Figure 3: Performance comparison of various LLMs on the FOE task, increasing the numbers of inferences.

Content: Interior materials remain perfectly adequate for the price of the truck and Honda's reputation. **Gold**: ('2021 honda ridgeline', 'interior', 'adequate for the price') **Predicted**: ('2021 honda ridgeline', 'interior materials', 'perfectly adequate')			
EM: ✗	L-RM: ✗	S-RM: ✓	CM: ✗
Content: It's not a tower of power by any stretch but gets the job done, even with a payload of swingset. **Gold**: ('2021 honda ridgeline', 'power', 'gets the job done') **Predicted**: ('2021 honda ridgeline', 'engine', 'gets the job done')			
EM: ✗	L-RM: ✗	S-RM: ✗	CM: ✓
Content: The major update to the Ridgeline for the 2021 model year isn't in its powertrain (remains the same), interior (reclaims a physical volume knob) **Gold**: ('2021 honda ridgeline', 'interior', 'reclaims a physical volume knob') **Predicted**: ('2021 honda ridgeline', 'volume knob', 'excellent')			
EM: ✗	L-RM: ✗	S-RM: ✗	CM: ✗

Table 4: Examples of comparisons between gold and predicted tuples for structured opinion extraction.

structured opinions. Specifically, even the best-performing model fails to achieve an F1 of 30 on both the rigid EM metric and the more relaxed RM metric, demonstrating significantly low performance. In contrast, employing the CM leads to a notable and consistent improvement in both tuple matching accuracy and overall recall. This metric effectively leverages LLMs' reasoning capabilities to mirror human judgment and has been shown to align most closely with human evaluations (see Table 3). Nevertheless, even with CM, **most models fail to both accurately predict the correct tuples and comprehensively cover all tuples present in the input content**, revealing inherent limitations in LLMs' extraction capabilities. This highlights that structured opinion extraction remains a highly complex and challenging task for LLMs, particularly in the context of realistic online content.

Effect of multiple inference on tuple extraction performance To investigate how extensively an LLM can extract structured tuples from content, we perform multiple inference iterations per single input and measure the model's extraction performance. For evaluation, we take the union of all tuples generated across iterations, remove duplicates, and consider only the unique *(entity, feature, opinion)* tuple sets. To capture a broader range of tuple sets, we set the temperature to 1.0 during inference. As shown in Figure 3, most models generate a significantly larger number of predicted tuples as the number of inference iterations increases, but the number of correctly matched tuples does not keep pace. Notably, recall improves significantly across most models but eventually reaches a plateau, where the rate of increase diminishes. **This implies that LLMs recognize a fixed set of opinions within the content, making**

it challenging to cover every opinion merely by increasing the number of inference iterations. Therefore, improving the extraction capabilities of LLMs requires exploring alternative strategies beyond merely repeating the inference process.

Case study: LLMs' extraction capability We conduct a case study to identify key failure patterns that limit LLMs' ability to extract structured opinions. Table 4 illustrates the comparison between the gold tuples and GPT-4o's predicted tuples for actual input content across EM, L-RM, S-RM, and CM. Despite being explicitly instructed in the input prompt to extract spans as-is, LLM often produces semantically related but non-identical spans—"interior materials" instead of "interior"—substitutes related concepts such as "engine" for "power", and even hallucinates opinions like "excellent" in place of "reclaims a physical volume knob". These patterns indicate that LLMs tend to transform or reinterpret textual information rather than extracting it verbatim as structured tuples. Such behavior underscores a fundamental limitation of LLMs in this task and suggests that structured extraction may not be an effective approach for opinion mining with LLMs.

4.2 RQ2: How insightfully can LLMs generate abstractive opinion summaries?

Automated evaluation results Table 5 reports the performance for the OIG task, using both lexical and semantic evaluation metrics. While models show strong word-level overlap (R-1 and R-L) in their summaries, they exhibit significantly lower bigram recall (R-2), highlighting difficulty in sus-

6

Models	Lexical			Semantic	
	R-1	R-2	R-L	BS	A3CU
Proprietary LLMs					
GPT-4o-mini	<u>39.30</u>	<u>14.05</u>	<u>34.58</u>	**90.35**	**38.49**
GPT-4o	**39.36**	**14.77**	**34.85**	<u>89.86</u>	38.39
Claude-3.5-Haiku	33.47	10.06	29.00	88.50	31.91
Claude-3.5-Sonnet	33.60	9.47	29.53	88.79	31.67
Open-source LLMs					
Llama3-8B-Instruct	<u>37.48</u>	**13.15**	<u>33.43</u>	89.91	30.50
Llama3-70B-Instruct	**37.61**	<u>13.04</u>	**33.18**	**90.15**	<u>31.48</u>
Gemma2-9B-it	35.03	11.47	30.99	88.25	31.16
Gemma2-27B-it	35.40	11.69	31.02	<u>90.08</u>	**34.09**
Qwen2.5-7B-Instruct	33.84	10.87	27.94	89.56	25.34
DeepSeek-7B-chat	35.03	10.68	30.72	76.89	25.80

Table 5: Performance comparison of various LLMs for the OIG task across automated evaluation metrics.

taining coherent phrase structures. Additionally, they achieve relatively high BS; their performance on A3CU remains substantially lower, suggesting that LLMs often capture surface-level semantic similarity but struggle to reflect deeper, human-aligned content understanding. Thus, to thoroughly gauge LLMs' abstractive strengths—particularly their capture of intent, subtle sentiment shifts, and deeper insights beyond surface semantics—a multifaceted evaluation framework is needed.

LLM-Judge evaluation across multiple perspectives To comprehensively analyze how well models generate abstractive opinion summaries, we conduct a reference-free evaluation using an LLM as the judgemeter. From the results in Figure 4, we derive the following key conclusions: **(1) LLMs consistently provide natural and readable summaries while preserving the original content without distortion or unnecessary modification.** This demonstrates their strength in faithfulness and fluency, ensuring that the generated summaries remain accurate and coherent. **(2) However, LLMs struggle to capture implicit user intentions, nuanced expressions, and meaningful insights that are not explicitly stated in the input content.** This limitation is reflected in lower scores for insightfulness and intent, indicating that while LLMs can summarize well, they lack deeper abstraction and contextual understanding.

Impact of structured opinions on summary Figure 6 demonstrates that augmenting opinion tuples during opinion-centric summary generation not only leads to substantial improvements in automatic evaluation metrics, but also provides practical benefits from a user perspective. In particular,
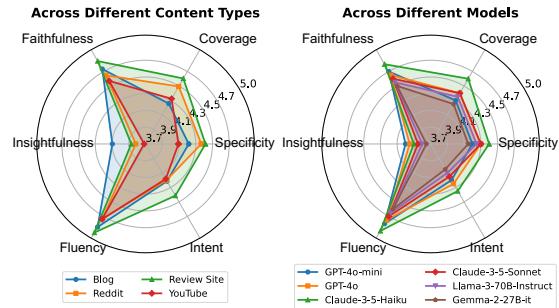

Figure 4: Radar charts for LLM-Judge evaluations of the OIG task. Comparison of the average model performance across different content types (**Left**). Comparison of performance across different models (**Right**).

the notable gains in *coverage* and *insightfulness* suggest that the model becomes more effective at capturing key opinions and delivering more informative summaries. Conversely, slight decreases in *intent* and *fluency* indicate that the added structure may sometimes interfere with natural expression and tone preservation. These results suggest that integrating structured opinion tuples into the insight generation pipeline is a key strategy for effective opinion mining, while also **highlighting the need for continued research into both the extractive and abstractive capabilities of LLMs.**

4.3 RQ3: Do LLMs effectively adapt to diverse online text environments?

To assess the adaptability of LLMs to the highly varied nature of online content, we analyze their performance across several dimensions. Figure 5 presents a comparative analysis of how different LLMs perform when these attributes vary.

LLMs struggle with dense and lengthy content Scenarios involving long content or a large number of tuples inherently present verbose and opinion-rich online content. The more densely packed the information within these texts, the more LLMs noticeably struggle to extract opinions and derive insights. This observation highlights the considerable challenges LLMs face in opinion mining when dealing with highly condensed and information-dense content. Therefore, it is crucial to explore more adaptive and effective mining approaches tailored to such complex scenarios (e.g., long-form user-generated texts, multi-thread discussions).

LLMs are robust in complex entity and multi-user environments In contrast, in environments where multiple users participate or the entity complexity increases, both extraction and opinion in-
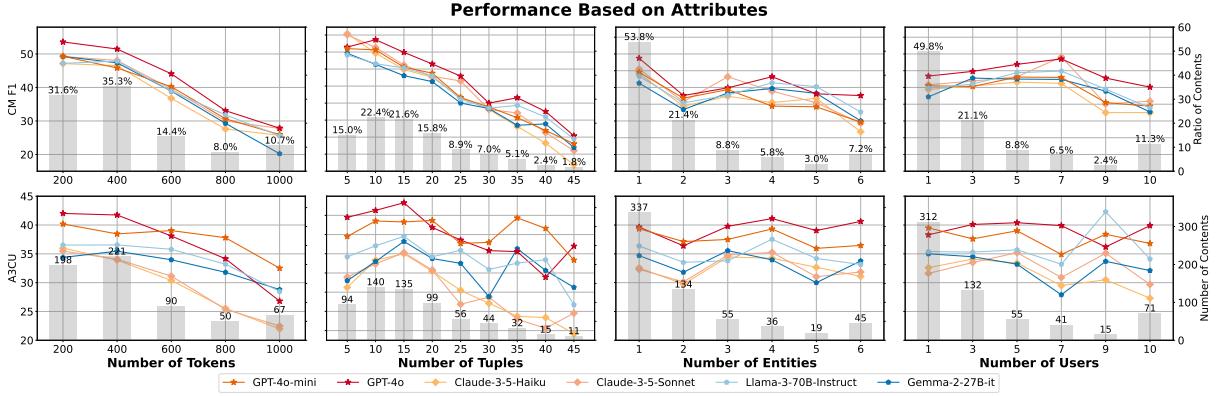
Figure 5: Performance comparison of various LLMs based on changes in different attributes within online content. CM F1 scores for the FOE task (**Upper**), and A3CU scores for the OIG task (**Lower**).
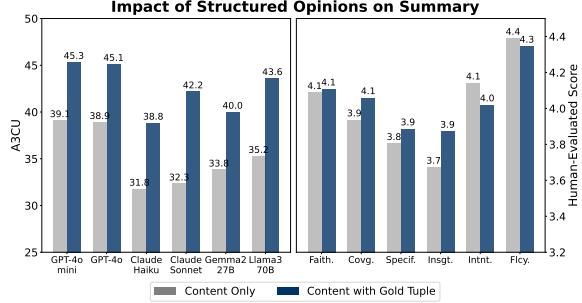

Figure 6: Comparison of the OIG task performance between content-only input and input with gold tuples. Automated metric performance by model (**Left**) and human evaluation of GPT-4o outputs (**Right**). Detailed human evaluation is presented in Appendix B.3.

sight generation show a relatively weaker downward trend. These results indicate that in settings such as forum discussions or multi-user comment threads, LLMs do not face significant challenges in extracting information and synthesizing opinions. This finding implies that forums and communities tend to consist of simple comments or relatively easily recognizable subtopics, allowing LLMs to identify and summarize key opinions with ease.

5 Related Work

Early studies on opinion mining (Pang et al., 2008) primarily focused on identifying and classifying opinion-related expressions or spans within text (Yang and Cardie, 2013; İrsoy and Cardie, 2014; Katiyar and Cardie, 2016; Xia et al., 2021; Liu et al., 2021; Zhang et al., 2022b). In particular, extracting opinions about specific aspects of products and services received significant attention (Fan et al., 2019; Wu et al., 2020; Zhao et al., 2020; Chen et al., 2020), and subsequent work extended this to jointly predict sentiment, enabling more complex and insightful analyses (Peng et al., 2019; Cai et al.,

2021; Zhang et al., 2021; Li et al., 2022; Kim et al., 2024a; Seo et al., 2024; Bai et al., 2024).

Recently, large language models (LLMs) (OpenAI et al., 2024; Grattafiori et al., 2024; Team et al., 2024a) have demonstrated remarkable zero-shot and in-context learning capabilities across a range of tasks, including information extraction (Kim et al., 2024b; Perot et al., 2024; Liu et al., 2024) and abstractive summarization (Chhabra et al., 2024; Tang et al., 2024; Siledar et al., 2024). While these advances suggest that LLMs have great potential in opinion mining, existing benchmarks fall short of capturing the complexity of real-world inputs and remain focused on simplified, structured extraction settings. As a result, they fail to fully assess the true potential of LLMs in this domain. To bridge this gap, we introduce the OOMB benchmark, which encompasses a wide spectrum of realistic online content and enables comprehensive investigation of both extraction and abstractive capacities of LLMs.

6 Conclusion

In this paper, we introduce OOMB, a novel benchmark designed to assess LLMs' capabilities in both structured opinion extraction and insight-oriented opinion generation across diverse and realistic online content scenarios. To the best of our knowledge, OOMB is the first comprehensive benchmark for evaluating LLMs in both structured and abstractive opinion mining tasks under real-world conditions. Our research reveals the dual challenge of precise opinion extraction and contextual insight generation, highlighting the need for future research to improve the effectiveness of both approaches. This work lays the foundation for LLM-based opinion mining and serves as a stepping stone for future research in this field.

Limitations

Despite its contributions, this study has several limitations, each of which also suggests promising directions for future research and practical extensions. First, although OOMB includes a diverse range of user-generated online content, it is currently confined to the vehicle domain, which may limit its generalizability to other areas such as electronics or healthcare. However, since the benchmark construction pipeline—including data collection, tuple annotation, and summary generation—is designed to be domain-agnostic, it can be easily extended to other fields with only minor adjustments to data sourcing and annotation guidelines.

Second, the current benchmark does not take into account user-specific information (user profiles). In real-world applications, factors such as user expertise, preferences, usage context, and prior sentiment trends play a critical role in shaping actionable insights. Integrating user metadata or interaction history would enable a natural extension of the framework toward user-aware opinion mining. While this direction is beyond the current scope, enriching OOMB with such annotations and modeling could open up new avenues for personalized opinion mining, allowing LLMs to produce more tailored and context-sensitive outputs.

Third, although we adopted a human–machine collaborative annotation pipeline (Sharif et al., 2024; Seo et al., 2025) to construct high-quality labels, opinion extraction and summarization inherently involve subjective judgment. To mitigate this, we established detailed annotation guidelines and a multi-stage validation process; nevertheless, some degree of annotation variance is unavoidable. Future work may explore more systematic approaches to subjectivity, such as crowdsourced consensus annotation, uncertainty-aware learning frameworks, or prompt ensemble methods.

Ethical Statement

This study strictly adhered to ethical guidelines throughout the process of data collection and usage. Data crawling was conducted solely for non-commercial research purposes and performed at a controlled rate to avoid overloading servers or causing potential DDoS attacks. When collecting user reviews, personal information such as reviewer IDs, names, and locations was intentionally excluded, focusing only on text and dates to ensure user privacy. However, we cannot entirely rule out the possibility that the review text may contain personal details, hate speech, or inappropriate content. All data samples were collected and annotated in compliance with the terms and conditions of their respective sources. By making our dataset and models accessible, we aim to foster academic progress in generative event extraction research.

References

Anthropic. 2024. Claude 3.5 haiku and sonnet.

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

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

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.

Shaowei Chen, Jie Liu, Yu Wang, Wenzheng Zhang, and Ziming Chi. 2020. Synchronous double-channel recurrent network for aspect-opinion pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6515–6524, Online. Association for Computational Linguistics.

Tao Chen, Premaratne Samaranayake, XiongYing Cen, Meng Qi, and Yi-Chen Lan. 2022. The impact of online reviews on consumers' purchasing decisions: Evidence from an eye-tracking study. *Frontiers in Psychology*, 13:865702.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 1–11, Mexico City, Mexico. 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.

Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif, Eduardo Blanco, and Steven Corman. 2025. BEMEAE: Moving beyond exact span match for event argument extraction. 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 5734–5749, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *Preprint*, arXiv:2302.04166.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanaz-

eri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan Mc-Phie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023a. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Yi Han, Gaurav Nanda, and Mohsen Moghaddam. 2023b. Attribute-sentiment-guided summarization of user opinions from online reviews. *Journal of Mechanical Design*, 145(4):041402.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Ozan İrsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728, Doha, Qatar. Association for Computational Linguistics.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

Jieyong Kim, Ryang Heo, Yongsik Seo, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2024a. Self-consistent reasoning-based aspect-sentiment quad prediction with extract-then-assign strategy. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7295–7303, Bangkok, Thailand. Association for Computational Linguistics.

Seoyeon Kim, Kwangwook Seo, Hyungjoo Chae, Jinyoung Yeo, and Dongha Lee. 2024b. Verifiner: Verification-augmented ner via knowledge-grounded reasoning with large language models. *Preprint*, arXiv:2402.18374.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, et al. 2022. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. *arXiv preprint arXiv:2211.05705*.

Yinghao Li, Rampi Ramprasad, and Chao Zhang. 2024. A simple but effective approach to improve structured language model output for information extraction. *arXiv preprint arXiv:2402.13364*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Siyi Liu, Yang Li, Jiang Li, Shan Yang, and Yunshi Lan. 2024. Unleashing the power of large language models in zero-shot relation extraction via self-prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13147–13161, Miami, Florida, USA. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Ziheng Liu, Rui Xia, and Jianfei Yu. 2021. Comparative opinion quintuple extraction from product reviews. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3955–3965.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2019. Knowing what, how and why: A

near complete solution for aspect-based sentiment analysis. In *AAAI Conference on Artificial Intelligence*.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, Chen-Yu Lee, and Nan Hua. 2024. LMDX: Language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15140–15168, Bangkok, Thailand. Association for Computational Linguistics.

Yanzhao Qin, Tao Zhang, Yanjun Shen, Wenjing Luo, Yan Zhang, Yujing Qiao, Zenan Zhou, Wentao Zhang, Bin CUI, et al. Sysbench: Can llms follow system message? In *The Thirteenth International Conference on Learning Representations*.

Agus Rahayu, Dian Herdiana Utama, and Ririe Novianty. 2021. The impact of online customer reviews on purchase intention in online marketplace. In *5th Global Conference on Business, Management and Entrepreneurship (GCBME 2020)*, pages 471–477. Atlantis Press.

Susana Santos and Helena Martins Gonçalves. 2021. The consumer decision journey: A literature review of the foundational models and theories and a future perspective. *Technological Forecasting and Social Change*, 173:121117.

Kwangwook Seo, Donguk Kwon, and Dongha Lee. 2025. Mt-raig: Novel benchmark and evaluation framework for retrieval-augmented insight generation over multiple tables. *arXiv preprint arXiv:2502.11735*.

Yongsik Seo, Sungwon Song, Ryang Heo, Jieyong Kim, and Dongha Lee. 2024. Make compound sentences simple to analyze: Learning to split sentences for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11171–11184, Miami, Florida, USA. Association for Computational Linguistics.

Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2024. Explicit, implicit, and scattered: Revisiting event extraction to capture complex arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12061–12081, Miami, Florida, USA. Association for Computational Linguistics.

Tejpalsingh Siledar, Swaroop Nath, Sankara Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, and Nikesh Garera. 2024. One prompt to rule them all: LLMs for opinion summary evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12119–12134, Bangkok, Thailand. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *Preprint*, arXiv:2402.13446.

An Tang, Xiuzhen Zhang, Minh Dinh, and Erik Cambria. 2024. Prompted aspect key point analysis for quantitative review summarization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024a. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Gemma Team, 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, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda

Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A user-centric multi-intent benchmark for evaluating large language models. *arXiv preprint arXiv:2404.13940*.

Meixi Wu, Wenya Wang, and Sinno Jialin Pan. 2020. Deep Weighted MaxSAT for Aspect-based Opinion Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5618–5628, Online. Association for Computational Linguistics.

Qingrong Xia, Bo Zhang, Rui Wang, Zhenghua Li, Yue Zhang, Fei Huang, Luo Si, and Min Zhang. 2021. A unified span-based approach for opinion mining with syntactic constituents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1795–1804.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria. Association for Computational Linguistics.

Xiaojun Yuan, Ning Sa, Grace Begany, and Huahai Yang. 2015. What users prefer and why: a user study on effective presentation styles of opinion summarization. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15*, pages 249–264. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. *arXiv preprint arXiv:2110.00796*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022a. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

Xin Zhang, Guangwei Xu, Yueheng Sun, Meishan Zhang, Xiaobin Wang, and Min Zhang. 2022b. Identifying chinese opinion expressions with extremely-noisy crowdsourcing annotations. *Preprint*, arXiv:2204.10714.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248, Online. Association for Computational Linguistics.

A Benchmark Construction Details

A.1 Data Source

To construct a diverse and representative dataset for online opinion mining, we collected user-generated content from four distinct web content types: *Blog*, *Reddit*, *Review Site*, and *YouTube*.

Blog   We collected data from well-established automotive blogs, including The Drive[5] , Autoblog[6], and CarExpert.[7] Blog content is primarily written by experts and car owners, often providing detailed and comprehensive insights on a single entity. Compared to other content types, blog posts tend to be longer and more structured, covering multiple aspects of a vehicle in depth.

Reddit   We collected data from the r/cars subreddit[8] , a community of car enthusiasts. Users freely share their opinions about various vehicles through a multi-threaded structure, where multiple participants engage in open discussions. This interactive nature generates diverse automotive perspectives through community discussions, making it a valuable source for opinion mining.

Review Site   We collected data from Edmunds[9] , an automotive review platform where users provide star ratings along with detailed reviews for specific vehicles. Review sites explicitly encourage opinion sharing, leading to more direct and detailed user feedback. These structured reviews combine ratings with detailed feedback, making them rich in straightforward user opinions.

YouTube   We gathered comments from automotive YouTube channels[10] [11] listed by Feedspot[12], focusing on channels with large subscriber bases providing car reviews and analysis. When YouTubers share their vehicle reviews, viewer opinions and reactions appear in the comments. YouTube uses a single-threaded structure where viewers can leave comments and engage in discussions through replies. This structure allows for community participation through viewer responses to both the video content and other comments, creating an interactive space for opinion sharing.

A.2 Human-in-the-loop Annotation Details

To ensure high-quality, consistent annotations, we adopted a human-in-the-loop process in which GPT-4o-mini[13] serves as the initial annotator and human annotators[14] subsequently verify and refine its outputs. Prompts used to solicit these initial annotations are provided in Table 24 for tuples and Table 26 for summaries. All annotators underwent one full week of training on our detailed guidelines (Table 25 for tuples; Table 27 for summaries) before beginning any annotation work. This entire annotation–refinement process was applied to every sample in the dataset, ensuring higher overall quality and consistency in the resulting annotations.

Tuple verification and refinement were performed via our annotation UI (see Figure 8 and 9), which displays the original content, each tuple's components with existence flags, and highlighted evidence sentences. Summary verification and refinement were conducted using a separate annotation UI (see Figure 10), which presents the content text, the associated gold tuple, and the working draft of the summary side by side for comparison and iterative improvement.

A.3 Dataset Statistics

We provide detailed statistics on key attributes—namely, the number of samples, average token count, number of users, and number of tuples (i.e., opinions)—for each of the four content types collected: blogs, Reddit, review sites, and YouTube. As shown in Table 6, we categorize the values of each attribute into predefined ranges to illustrate the distribution of samples across different levels. Token lengths were measured using the NLTK word_tokenize library[15].

A.4 Dataset Analysis

Additionally, Figure 7 provides the distribution of feature-opinion topics in our dataset.

Feature keywords   The t-SNE visualization shows that feature keywords form well-separated

---

[5] https://www.thedrive.com/category/car-reviews

[6] https://www.autoblog.com/reviews/

[7] https://www.carexpert.com.au/car-reviews

[8] https://www.reddit.com/r/cars/

[9] https://www.edmunds.com/car-reviews/

[10] https://www.youtube.com/@AutoTraderTV

[11] https://www.youtube.com/channel/UCsqjHFMB_JYTaEnf_vmTNqg

[12] https://videos.feedspot.com/car_youtube_channels/

[13] gpt-4o-mini-2024-07-18

[14] We recruit undergraduates and graduates who are proficient in English and knowledgeable in the automotive domain.

[15] https://www.nltk.org/api/nltk.tokenize.word_tokenize.html

clusters according to major product aspects. Categories like *Interior & Design*, *Driving Experience*, and *Performance & Powertrain* appear frequently and show high cohesion. *Engine & Driving Performance* and *Infotainment $ Digital Systems* are located near the core clusters, reflecting semantic proximity, while Overall Vehicle appears more scattered, indicating higher lexical variability. Outlier points on the periphery suggest rare or ambiguous feature expressions that may require finer handling.

Opinion keywords Opinion keywords also exhibit meaningful clustering patterns. *Price & Value* and *Specs & Performance* form tight clusters, while *Utilitarian and Emotional Evaluations* overlap in the center, suggesting a blend of practical and emotional judgments. *Tech & Functionality* Evaluations appear in a distinct region, separate from general *Positive & Negative* sentiment expressions, highlighting their specialized nature. Points between clusters reflect nuanced or polysemous opinions, suggesting the need for flexible sentiment understanding models.

B Experimental Details

B.1 Evaluation Models

Proprietary LLMs We used the most up-to-date versions of OpenAI APIs[16] and Anthropic AI[17]. Specifically, we used the following models:

- **GPT-4o-mini**: `gpt-4o-mini-2024-07-18`
- **GPT-4o**: `gpt-4o-2024-08-06`
- **Claude-3.5-Haiku**:
  `claude-3-5-haiku-20241022`
- **Claude-3.5-Sonnet**:
  `claude-3-5-sonnet-20241022`

Open-sourced LLMs We used Hugging Face model cards and ran them on two NVIDIA A100 GPUs. Specifically, we used the following models:

- **Llama3-8B-Instruct**:
  `meta-llama/meta-llama-3-8b-instruct`
- **Llama3-70B-Instruct**:
  `meta-llama/meta-llama-3-70b-instruct`
- **Gemma2-9B-it**: `google/gemma-2-9b-it`
- **Gemma2-27B-it**: `google/gemma-2-27b-it`
- **Qwen2.5-7B-Instruct**:
  `Qwen/Qwen2.5-7B-Instruct`

---

[16]https://openai.com/index/openai-api/
[17]https://www.anthropic.com/

- **DeepSeek-7B-chat**:
  `deepseek-ai/deepseek-llm-7b-chat`

B.2 Feature-centric opinion extraction

RM metric threshold selection We set the threshold of the Relaxed Match (RM) metric to 0.7, as it empirically provides the optimal balance—capturing meaningful semantic similarities without being overly permissive. Prior information extraction (IE) research has highlighted that exact span matching may underestimate model performance due to its overly strict nature. To alleviate this issue, previous studies have proposed overlap-based evaluations with thresholds set at 0.5 (Han et al., 2023a) or 0.75 (Sharif et al., 2024).

As shown in Table 7 and 8, lower thresholds tend to excessively acknowledge partial overlaps, inflating recall to an unrealistic degree. Conversely, higher thresholds often miss semantically valid matches due to minor textual variations, causing the RM metric performance to converge toward Exact Match (EM) scores and consequently lose its intended flexibility. Thus, our experiments confirm that a threshold of 0.7 achieves optimal RM performance, which we subsequently adopt for our main experiments.

Human Alignment in Tuple Matching This experiment aims to identify the most appropriate matching metric for reliably evaluating LLMs' tuple extraction performance. To this end, we investigate which of the three evaluation metrics used in the FOE task—Exact Match (EM), Relaxed Match (RM), and Contextual Match (CM)—best aligns with human judgment. First, we randomly select 100 tuples predicted by GPT-4o given the input content texts. Then, five human annotators evaluated the validity of each predicted-gold tuple pair using binary judgments: 1 if they considered the pair to be a match, and 0 otherwise. Based on these judgments, we computed the correlation coefficients between human agreement and each metric.

Table 3 reports the Pearson r, Spearman ρ, and Kendall τ correlation coefficients, averaged across the five annotators over the 100 samples. Across all correlation metrics, CM achieved the highest alignment with human judgment, indicating that Contextual Match best reflects how humans assess tuple matching. These findings suggest that CM serves as the most reliable and appropriate metric for evaluating the performance of LLMs in tuple extraction tasks.

B.3 Opinion-centric insight generation

LLM-Judge Evaluation Criteria To evaluate summary quality across diverse criteria, we use GPT-4o and randomly sample 100 pieces of content from each of the four content types. The scoring scale for each evaluation follows the previous NLG evaluation framework, G-EVAL (Liu et al., 2023a), and is measured on a scale from 1 to 5. We adopt the following six criteria:

- **Faithfulness**: Evaluate whether the summary faithfully reflects the original review without distortion and check for any hallucinations.

- **Coverage**: Evaluate whether the summary effectively captures and represents the key opinions expressed in the review.

- **Specificity**: Evaluate whether the summary presents meaningful and relevant details rather than being vague or overly generic.

- **Insightfulness** : Evaluate whether the summary provides meaningful insights that enhance understanding or decision-making for the reader.

- **Intent**: Evaluate whether the summary accurately preserves the author's original tone, intent, and nuances without altering the emotional or stylistic essence of the review.

- **Fluency**: Evaluate whether the summary is naturally written, grammatically correct, and easy to read.

Case study: annotated opinion-centric summary Table 28 presents representative examples of gold opinion-centric summaries for each content type.

Case study: LLMs' summarize capability We illustrate our LLM-Judge evaluation protocol with two case studies: one on a review site (Tables 29 and 31), and one on a Reddit (Tables 30 and 32). In each set, the first table shows the input content, the model-generated summary (GPT-4o), the gold opinion-centric summary, and the automatic metric A3CU score, while the second table provides a reference-free, six-dimension LLM-Judge assessment complete with per-dimension scores and detailed reasoning.

Although the A3CU metric assigns the review site example a high score (60.91) and the Reddit example a low score (10.12), our reference-free LLM-Judge evaluation reveals that the model's performance on the Reddit content is in fact stronger across several human-aligned dimen-

sions—particularly *Coverage* and *Specificity*. This divergence highlights the limitations of purely reference-based, automatic metrics in capturing the nuanced, insight-oriented qualities of opinion summaries. We therefore conclude that for OIG evaluation, combining automatic reference-based metrics with a reference-free, human-aligned judging protocol yields a more comprehensive and reliable assessment of LLMs' true insight-generation capabilities.

Human Evaluation We assess the quality of the generated summary through a human evaluation conducted on Amazon Mechanical Turk (AMT). Specifically, we randomly sample 200 examples from our benchmark and ask three human judges per example to evaluate summaries generated by GPT-4o under two settings: 1) using only the input content, and 2) using both the input content and gold tuples. Each judge rates the quality of the summaries on a 1 to 5 scale across six criteria. The AMT interface used for human evaluation is presented in Figure 11, 12, 13, 14, 15, 16 and 17.

B.4 Performance of LLMs by content type

We report the Feature-centric opinion extraction performance of various LLMs for each content type in Table 9, 10, 11 and 12 , and the Opinion-centric insight generation performance in Table 13, 14.

B.5 Prompts

We present prompts used in our experiments:

- **Data Annotation**: The prompt designed for Feature-centric opinion extraction is shown in Table 24 and the prompt for Opinion-centric insight generation is shown in Table 26

- **Feature-centric opinion extraction**: The prompt designed for Feature-centric opinion extraction is shown in Table 15.

- **Opinion-centric insight generation**: The prompt designed for Opinion-centric insight generation is shown in Table 16.

- **Contextual Match**: The prompt used for performing Contextual Match is provided in Table 17.

- **LLM-Judge Evaluation**: The prompt for evaluating the LLM-Judge is presented in Table 18, 19, 20, 21, 22 and 23.

Attribute	Range	Content Type				Total	Ratio(%)
		Blog	Reddit	Review Site	YouTube		
Tuples	≤ 10	8	79	14	125	226	37.7
	≤ 15	14	49	33	32	128	21.3
	≤ 20	19	21	33	21	94	15.7
	≤ 25	20	15	13	7	55	9.2
	≤ 30	20	8	8	8	44	7.3
	≤ 35	34	5	5	9	53	8.8
Entities	1	58	53	89	128	328	54.7
	2	44	34	12	39	129	21.5
	3	9	24	3	17	53	8.8
	4	2	24	0	7	33	5.5
	5	1	10	1	3	15	2.5
	≥ 6	1	32	1	8	42	7.0
Tokens	≤ 200	0	100	0	81	181	30.2
	≤ 400	3	58	85	68	214	35.7
	≤ 1000	9	14	21	44	88	14.7
	≤ 2000	39	3	0	8	50	8.3
	≤ 3000	64	2	0	1	67	11.2
Users	1	58	47	89	128	322	53.7
	2	44	35	12	39	130	21.7
	3	9	24	3	17	53	8.8
	4	2	26	0	7	35	5.8
	≥ 5	2	45	2	11	60	10.0
Total		115	177	106	202	600	100.0

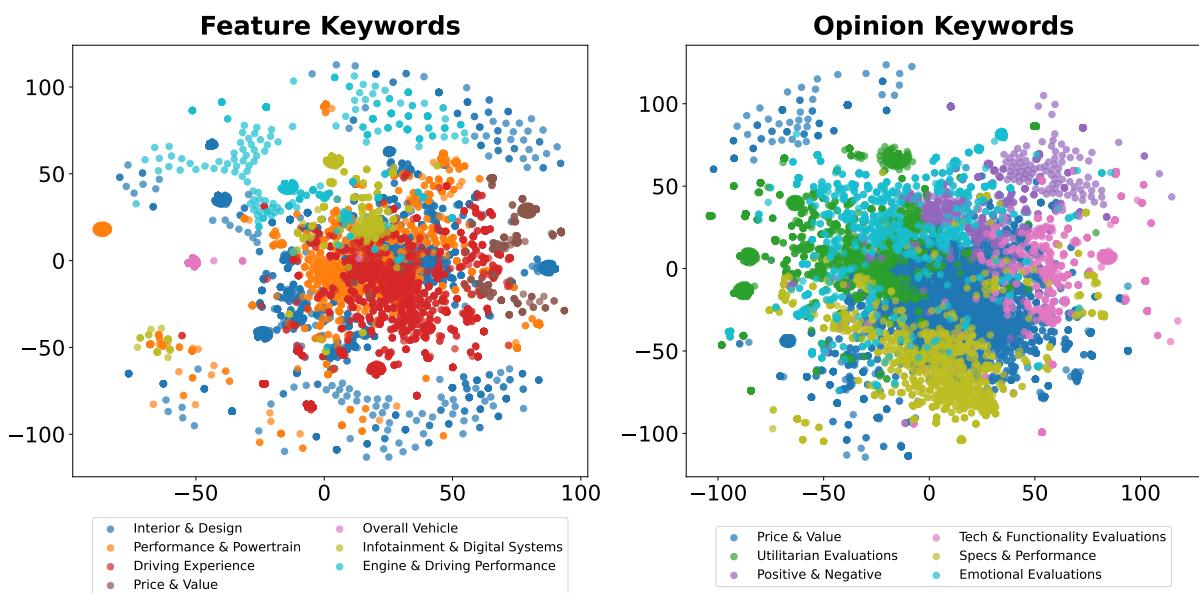Table 6: Statistics of the OOMB dataset across different attributes.

Figure 7: Visualization of feature keywords **(Left)** and opinion keywords **(Right)** extracted via K-means clustering using t-SNE. Each point represents a keyword, and colors indicate different topic clusters.

Models	EM			L-RM ≥ 0.7			L-RM ≥ 0.8			L-RM ≥ 0.9		
	Pre	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Proprietary LLMs												
GPT-4o-mini	3.91	1.99	2.62	11.29	5.70	7.52	7.72	3.68	4.98	4.80	2.29	3.10
GPT-4o	7.27	5.18	6.02	15.86	11.39	13.20	11.90	8.51	9.92	8.43	6.03	7.03
Claude-3.5-Haiku	6.13	3.02	4.01	15.02	7.51	9.94	10.61	4.97	6.77	7.17	3.36	4.57
Claude-3.5-Sonnet	**11.12**	**6.32**	**7.97**	**22.97**	**13.02**	**16.46**	**17.81**	**9.61**	**12.48**	**12.57**	**6.78**	**8.81**
Open-source LLMs												
Llama3-8B-Instruct	**8.49**	**6.28**	**7.17**	**16.75**	**12.18**	**14.02**	**13.13**	**9.17**	**10.80**	**9.85**	**6.88**	**8.10**
Llama3-70B-Instruct	7.26	5.57	6.27	15.21	11.66	13.13	12.07	9.07	10.36	8.67	6.52	7.44
Gemma2-9B-it	6.37	4.51	5.25	14.17	10.17	11.78	9.98	7.27	8.41	7.11	5.18	5.99
Gemma2-27B-it	7.05	5.61	6.20	14.33	11.77	12.82	10.80	9.08	9.87	7.55	6.35	6.90
Qwen2.5-7B-Instruct	6.55	4.18	5.05	12.97	8.32	10.02	9.99	6.14	7.60	7.35	4.51	5.59
DeepSeek-7B-chat	3.00	1.63	2.07	5.86	3.13	4.02	4.88	2.31	3.14	3.78	1.79	2.43

Table 7: Ablation study on the FOE task using three L-RM thresholds (≥ 0.7, ≥ 0.8, ≥ 0.9).

Models	EM			S-RM ≥ 0.7			S-RM ≥ 0.8			S-RM ≥ 0.9		
	Pre	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Proprietary LLMs												
GPT-4o-mini	3.91	1.99	2.62	15.50	7.76	10.27	9.86	4.70	6.37	5.48	2.61	3.54
GPT-4o	7.27	5.18	6.02	21.23	15.31	17.71	14.76	10.55	12.31	9.39	6.72	7.83
Claude-3.5-Haiku	6.13	3.02	4.01	20.60	10.36	13.68	13.41	6.28	8.56	8.55	4.00	5.45
Claude-3.5-Sonnet	**11.12**	**6.32**	**7.97**	**29.30**	**16.52**	**20.93**	**21.08**	**11.37**	**14.78**	**14.08**	**7.59**	**9.87**
Open-source LLMs												
Llama3-8B-Instruct	**8.49**	**6.28**	**7.17**	**21.33**	15.43	**17.80**	**15.52**	10.84	**12.77**	**10.59**	7.39	**8.71**
Llama3-70B-Instruct	7.26	5.57	6.27	19.94	15.18	17.15	13.80	10.37	11.85	9.44	7.09	8.10
Gemma2-9B-it	6.37	4.51	5.25	17.73	12.59	14.64	11.73	8.55	9.89	7.73	5.64	6.52
Gemma2-27B-it	7.05	5.61	6.20	19.29	**15.83**	17.25	13.09	**11.01**	11.96	8.18	6.88	7.48
Qwen2.5-7B-Instruct	6.55	4.18	5.05	18.22	11.83	14.18	12.36	7.59	9.41	8.38	5.15	6.38
DeepSeek-7B-chat	3.00	1.63	2.07	8.20	4.34	5.61	5.68	2.69	3.65	3.94	1.87	2.53

Table 8: Ablation study on the FOE task using three S-RM thresholds (≥ 0.7, ≥ 0.8, ≥ 0.9).

Models	EM			L-RM			S-RM			CM		
	Pre	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**	**Pre**	**Rec**	**F1**
Proprietary LLMs												
GPT-4o-mini	2.40	1.08	1.49	10.17	4.58	6.31	13.44	6.05	8.34	**49.14**	22.55	29.49
GPT-4o	4.67	**3.53**	**4.02**	14.27	10.78	12.29	17.91	13.53	15.41	38.21	**28.34**	**30.63**
Claude-3.5-Haiku	3.31	1.44	2.00	10.15	4.41	6.15	14.96	6.50	9.07	46.95	21.98	28.75
Claude-3.5-Sonnet	**5.68**	2.91	3.85	**18.75**	9.61	12.71	**25.51**	13.07	17.29	45.30	23.97	30.08
Open-source LLMs												
Llama3-8B-Instruct	4.36	2.78	3.39	11.74	7.48	9.14	14.82	9.44	11.54	**39.53**	25.06	**29.20**
Llama3-70B-Instruct	**5.86**	**4.41**	**5.04**	12.34	9.28	10.59	15.86	11.93	13.61	35.42	27.00	28.96
Gemma2-9B-it	3.38	2.80	3.06	10.01	8.36	9.15	11.98	9.92	10.86	33.45	26.90	27.64
Gemma2-27B-it	2.54	2.54	2.54	8.19	8.22	8.20	10.95	10.98	10.96	30.76	**27.42**	26.27
Qwen2.5-7B-Instruct	3.81	1.96	2.59	8.83	4.54	6.00	12.95	6.67	8.80	28.15	18.26	20.24
DeepSeek-7B-chat	1.57	0.49	0.75	2.72	0.85	1.29	3.24	1.01	1.54	34.89	12.07	16.25

Table 9: Performance comparison of different models for FOE task across various evaluation metrics on Blog type.

Models	EM			L-RM			S-RM			CM		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Proprietary LLMs												
GPT-4o-mini	3.65	1.43	2.06	9.23	3.62	5.20	14.30	5.61	8.05	68.73	33.07	41.11
GPT-4o	9.74	**5.65**	7.15	18.18	**10.54**	13.34	23.87	**13.84**	17.52	63.59	41.83	**47.70**
Claude-3.5-Haiku	6.25	2.43	3.49	14.34	5.57	8.02	18.34	7.12	10.25	68.80	34.67	42.43
Claude-3.5-Sonnet	**13.14**	5.41	**7.66**	22.32	9.18	13.01	28.60	11.77	16.68	68.68	34.84	42.90
Open-source LLMs												
Llama3-8B-Instruct	9.28	5.61	6.99	**17.31**	10.46	**13.04**	22.45	13.56	16.91	56.18	40.66	43.77
Llama3-70B-Instruct	8.39	4.97	6.24	15.38	9.11	11.44	21.29	12.60	15.83	58.97	37.66	42.89
Gemma2-9B-it	8.10	4.53	5.81	15.78	8.83	11.32	21.39	11.97	15.35	**61.53**	40.76	**45.74**
Gemma2-27B-it	**9.84**	5.92	**7.40**	15.92	9.58	11.96	21.60	13.00	16.23	60.40	37.04	42.54
Qwen2.5-7B-Instruct	8.67	4.41	5.85	15.85	8.07	10.70	21.00	10.70	14.17	38.56	19.64	26.03
DeepSeek-7B-chat	3.71	1.83	2.45	7.66	3.78	5.06	9.67	4.77	6.39	53.42	32.32	36.14

Table 10: Performance comparison of different models for FOE task across various evaluation metrics on Reddit type.

Models	EM			L-RM			S-RM			CM		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Proprietary LLMs												
GPT-4o-mini	4.87	2.93	3.66	13.24	7.97	9.95	17.72	10.67	13.32	**77.83**	48.15	57.65
GPT-4o	3.92	3.17	3.50	10.81	8.73	9.66	19.23	15.53	17.19	70.44	**57.93**	**61.80**
Claude-3.5-Haiku	5.54	3.40	4.21	19.10	11.72	14.53	27.70	17.00	21.07	73.41	46.84	56.01
Claude-3.5-Sonnet	**14.01**	9.44	**11.28**	29.16	19.64	23.47	36.64	24.68	29.49	70.94	49.52	57.14
Open-source LLMs												
Llama3-8B-Instruct	**9.82**	7.56	**8.54**	21.31	16.41	18.54	28.46	21.92	24.77	64.15	50.72	54.89
Llama3-70B-Instruct	5.90	4.63	5.19	17.16	13.48	15.10	24.33	19.11	21.41	**66.23**	52.56	**57.13**
Gemma2-9B-it	7.24	5.22	6.06	16.99	12.25	14.24	21.87	15.77	18.32	63.11	48.66	53.49
Gemma2-27B-it	6.39	5.63	5.99	16.98	14.95	15.90	23.97	21.10	22.44	62.52	53.93	56.37
Qwen2.5-7B-Instruct	3.55	2.99	3.24	9.74	8.21	8.91	17.66	14.89	16.16	62.82	53.48	55.82
DeepSeek-7B-chat	2.22	1.11	1.48	5.25	2.64	3.51	10.50	5.28	7.02	63.69	34.16	42.16

Table 11: Performance comparison of different models for FOE task across various evaluation metrics on Review Site type.

Models	EM			L-RM			S-RM			CM		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Proprietary LLMs												
GPT-4o-mini	4.81	2.62	3.39	12.87	7.02	9.09	16.50	9.00	11.65	63.77	42.11	45.49
GPT-4o	10.73	**8.51**	9.49	19.78	**15.68**	17.49	23.33	**18.50**	20.63	61.28	53.61	51.81
Claude-3.5-Haiku	9.13	4.75	6.25	16.27	8.46	11.13	21.41	11.13	14.64	63.03	43.5	46.79
Claude-3.5-Sonnet	**11.61**	7.62	9.20	20.95	13.75	16.60	25.47	16.72	20.19	60.1	48.79	48.93
Open-source LLMs												
Llama3-8B-Instruct	**10.65**	8.90	**9.70**	17.75	14.84	16.16	21.07	17.61	19.18	48.28	49.99	44.42
Llama3-70B-Instruct	9.93	8.95	9.42	**17.78**	16.02	**16.86**	21.19	19.09	20.08	50.54	51.07	45.76
Gemma2-9B-it	6.76	5.49	6.06	13.83	11.23	12.39	15.66	12.71	14.03	52.44	48.14	45.38
Gemma2-27B-it	9.45	8.36	8.87	16.22	14.34	15.22	20.64	18.25	19.37	50.8	50.02	45.46
Qwen2.5-7B-Instruct	10.45	7.62	8.81	17.38	12.66	14.65	21.18	15.43	17.85	**52.97**	45.63	43.92
DeepSeek-7B-chat	4.50	3.12	3.68	7.64	5.29	6.25	9.21	6.38	7.54	45.64	36.73	34.94

Table 12: Performance comparison of different models for FOE task across various evaluation metrics on Youtube type.

Models	Blog					Reddit				
	Lexical			Semantic		Lexical			Semantic	
	R-1	R-2	R-L	BS	A3CU	R-1	R-2	R-L	BS	A3CU
Proprietary LLMs										
GPT-4o-mini	**40.14**	**16.32**	**36.06**	**90.32**	**36.03**	38.41	12.60	33.15	90.17	37.71
GPT-4o	37.06	14.68	33.03	89.18	31.34	**41.13**	**15.33**	**36.15**	**90.22**	**40.65**
Claude-3.5-Haiku	31.15	9.91	27.01	87.45	25.75	35.20	10.49	30.42	89.10	34.22
Claude-3.5-Sonnet	31.87	9.04	28.44	87.96	25.69	34.77	9.88	30.49	89.28	34.62
Open-source LLMs										
Llama3-8B-Instruct	35.77	14.10	32.63	89.52	30.27	33.73	9.26	29.58	85.82	28.52
Llama3-70B-Instruct	**38.82**	**15.65**	**34.70**	**90.09**	**32.22**	**37.60**	**12.13**	**32.76**	89.70	**35.24**
Gemma2-9B-it	35.86	13.55	33.45	89.77	28.95	35.00	10.31	30.54	88.36	30.72
Gemma2-27B-it	34.73	12.42	31.07	89.74	31.00	36.27	11.04	31.38	**90.01**	34.36
Qwen2.5-7B-Instruct	35.54	12.45	32.29	89.88	26.30	31.27	8.74	26.92	89.37	25.24
DeepSeek-7B-chat	34.79	11.75	30.95	73.04	21.09	35.21	10.22	30.89	75.50	25.98

Table 13: Performance comparison of different models for the OIG task using lexical and semantic metrics on the Blog and Reddit Types.

Models	Review Site					YouTube				
	Lexical			Semantic		Lexical			Semantic	
	R-1	R-2	R-L	BS	A3CU	R-1	R-2	R-L	BS	A3CU
Proprietary LLMs										
GPT-4o-mini	**39.85**	14.35	**35.24**	**90.42**	**41.60**	38.90	12.82	33.90	**90.47**	38.96
GPT-4o	39.44	14.50	34.98	89.83	41.07	**40.15**	**14.67**	**35.65**	90.28	**41.45**
Claude-3.5-Haiku	34.57	10.20	29.76	88.70	36.87	33.36	9.76	29.14	88.76	31.57
Claude-3.5-Sonnet	33.80	9.44	29.22	88.84	34.34	34.12	9.57	30.03	89.11	32.47
Open-source LLMs										
Llama3-8B-Instruct	**40.97**	**15.85**	**36.60**	90.05	38.24	34.77	9.88	30.49	89.66	33.89
Llama3-70B-Instruct	36.00	12.00	31.64	90.34	37.30	**44.88**	**17.38**	29.49	**90.64**	**36.70**
Gemma2-9B-it	34.10	11.17	30.06	88.45	35.43	34.84	10.54	30.70	86.36	29.41
Gemma2-27B-it	35.48	12.43	30.84	**90.46**	**38.58**	35.12	10.88	30.87	90.10	32.90
Qwen2.5-7B-Instruct	33.10	10.71	29.61	89.73	25.29	35.37	11.27	22.89	89.27	24.47
DeepSeek-7B-chat	33.74	9.81	29.01	83.65	29.12	36.47	10.84	**32.14**	77.44	27.21

Table 14: Performance comparison of different models for the OIG task using lexical and semantic metrics on the Review Site and YouTube Types.

Feature-centric opinion extraction (FOE) task prompt

[Task Description]

You are a car opinion miner for the user. Your task is to extract tuples in the form of (entity, feature, opinion) by identifying attributes, specific features, or components mentioned in the text and associating opinions with each feature.

1. A "entity" is the name of the car model with brand which opinions are expressed (e.g., "volvo xc90", "toyota camry", "Nissan Sentra").
2. A "feature" as a specific characteristic, attribute, or component of an entity that users mention or evaluate (e.g., "interior design," "fuel efficiency," "safety features"). - The performance, design, or experience (e.g., "handling", "ride comfort"). - Distinct functions or technologies in a vehicle (e.g., "infotainment system"). - Physical parts or systems that make up the vehicle (e.g., "brake", "transmission").
3. An "opinion" is a subjective or objective judgment, reaction, experience, evaluation, or feedback about the entity's feature, including assessments of quality, performance, or value, as well as direct responses or reactions from users based on their experience.

IMPORTANT:

- Features and opinions MUST be extracted in the input text. Never generate words or terms that do not exist in the text.

- However, if the feature corresponding to an opinion does not exist in the text and is implicit, the feature is treated as "NULL".

- The opinion MUST be no more than 5 words.

- The output must be in valid JSON format, but **DO NOT** include "json" code block delimiters (e.g., "'json ... "').

- Return **only** the JSON object, without any extra text, explanations, or comments.

- Provide only the tuples. Do not mention your process or how you arrived at it.

- Note: Return your results in JSON format only, with the following structure: {'opinion_tuple': [{'entity': <str>, 'feature': <str>, 'opinion': <str>}, {'entity': <str>, 'feature': <str>, 'opinion': <str>}, ..., {'entity': <str>, 'feature': <str>, 'opinion': <str>}]}

[Content Text]

...

Table 15: The prompt for Feature-centric opinion extraction.

Opinion-centric insight generation (OIG) task prompt

[Task Description]
You are a product and marketing manager at a global automotive company. Your task is to produce a free-form summary that categorizes and organizes a user text into higher-level insights, such that the report alone provides a clear understanding of the key opinions expressed. This summary should be written in natural, human-like language and structured around the core topics (features).

Step to Follow:
1. Read and Understand

- Examine the online text to identify its main points.

2. Organize Top-Level Topics

- Group similar or related tuples into clear categories (e.g., "Engine Issues," "Warranty Feedback," etc.).
- Reflect on the frequency or intensity of opinions if it helps convey importance.

3. Create a Three-Line Report

- Line 1: Highlight the most frequently mentioned or emphasized features, grouping related opinions into high-level categories.
- Line 2: Focus on the features that a user strongly praised or criticized, incorporating the intensity or frequency of opinions where applicable.
- Line 3: Provide a cohesive conclusion summarizing the overarching sentiment or key takeaway from a user's text.

4. Write the Summary

- Make it short (three to five lines).
- Use clear, direct language.
- Ensure that reading only this summary sufficiently conveys a user's main viewpoints.

What to Avoid:

- Provide only the summary. Do not mention your process or how you arrived at it.
- Do not include introductory phrases such as "Here is a summary of the review" or "Based on the review text."
- Do not directly copy sentences from the online text; rephrase and synthesize information.
- Carefully analyze the given text to determine the number of users and decide whether to use "user" or "users" accordingly.

Note: Return your results in JSON format only, with the following structure: {'summary': <str>}

[Content Text]
...

Table 16: The prompt for Opinion-centric insight generation.

Contextual Match (CM) prompt

[Task Description]
You are given two lists of tuples, each in the form [(entity, feature, opinion), (entity, feature, opinion), ...]. One list represents the Gold (correct) tuples, and the other list represents the Model's Predicted tuples. Your goal is to calculate four values: - matched_pred_tuple: The matched Predicted tuple. - matched_gold_tuple: The matched Gold tuple.

Match Criteria:
- Examine the two lists of tuples to identify their main points.
- Convert each element (entity, feature, opinion) to lowercase before comparing (e.g., "kia soul" vs. "Kia Soul" are equivalent).
- Allow flexibility when matching tuples by considering semantic equivalence, synonyms, rephrased expressions, or other valid variations that convey the same context or meaning. For instance, the following cases should be considered as valid matches:

- Pred: ("toyota corolla", "brakes", "getting hot"), Gold: ("toyota corolla le", "brakes", "getting hot")

- Pred: ("toyota camry", "null", "looks better"), Gold: ("camry", "looks", "better")

- Pred: ("porsche 911", "performance", "can go effortlessly fast"), Gold: ("porsche 911", "drive", "effortlessly fast")

- Pred: ("aston martin vanquish", "rear badge", "would look better"), Gold: ("aston martin vanquish", "badge", "better")

- A tuple is considered a relaxed match if all three elements are semantically equivalent after applying these transformations.
- Do not count a match more than once if there are duplicates.
- Note: Return your results in JSON format only, with the following structure:
```
{'matched_tuple_pair': [ {'matched_pred_tuple': ('entity', 'feature',
'opinion'), 'matched_gold_tuple': ('entity', 'feature', 'opinion') ...
},
```

[Pred Tuples]
...

[Gold Tuples]
...

Table 17: The prompt for Contextual Match (CM).

LLM-Judge Evaluation prompt - (Faithfulness)

[Task Description]
You are provided with an online user's automobile review text along with an opinion-centric insight summary that groups user opinions at the topic level to offer insights. Your task is to evaluate the summary based on its Faithfulness. Make sure you understand the following evaluation metric very clearly.

Evaluation Criteria:
Faithfulness (1-5) – Evaluate whether the summary accurately represents the original review without distorting its meaning, omitting key details, or introducing hallucinated information that was not present in the original text.
• Score 1: The summary completely distorts the original online content
and contains many incorrect information. It cannot be trusted at all.
• Score 2: The summary significantly misrepresents the original online content with several incorrect information.
• Score 3: The summary partially reflects the original online content but has some minor incorrect information.
• Score 4: The summary significantly misrepresents the original online content with several incorrect information.
• Score 5: The summary completely reflects the original online content without any distortions and incorrect information.

[Review Text:] ...

[Summary:] ...

Evaluation Steps:
1. Read through the review text provided.
2. Read the corresponding opinion-centric insight summary.
3. Evaluate whether the summary faithfully reflects the original review without any distortion.
4. Check if the summary contains any hallucinated information not present in the review.
5. Assign a score from 1 to 5 for Faithfulness, where 1 indicates very poor faithfulness and 5 indicates excellent faithfulness.

Table 18: The prompt for LLM-Judge Evaluation (Faithfulness).

LLM-Judge Evaluation prompt - (Coverage)

[Task Description]
You are provided with an online user's automobile review text along with an opinion-centric insight summary that groups user opinions at the topic level to offer insights. Your task is to evaluate the summary based on its Coverage. Make sure you understand the following evaluation metric very clearly.

Evaluation Criteria: Coverage (1-5) – Evaluate whether the summary effectively captures and represents the key opinions expressed in the review without omitting significant details or key points. Ensure that all essential opinions are included and accurately represented.
•Score 1: The summary fails to capture any key opinions from the online content. The content is either completely missing or irrelevant to the original opinions.
•Score 2: The summary captures only a small portion of key opinions. Many important opinions from the online content are missing.
•Score 3: The summary captures some key opinions but misses others. The coverage is partial and could be more comprehensive.
•Score 4: The summary effectively captures most key opinions from the online content. The coverage is good but may miss minor details.
•Score 5: The summary comprehensively captures all key opinions from the online content. Nothing important is missing, and the coverage is complete.

[Review Text:] ...

[Summary:] ...

Evaluation Steps:
1. Read through the review text provided.
2. Read the corresponding opinion-centric insight summary.
3. Evaluate whether the summary effectively captures and represents the key opinions expressed in the review.
4. Consider if any significant details or opinions are missing from the summary.
5. Assign a score from 1 to 5 for Coverage, where 1 indicates very poor coverage and 5 indicates excellent coverage.

Table 19: The prompt for LLM-Judge Evaluation (Coverage).

LLM-Judge Evaluation prompt - (Specificity)

[Task Description]
You are provided with an online user's automobile review text along with an opinion-centric insight summary that groups user opinions at the topic level to offer insights. Your task is to evaluate the summary based on its Specificity. Make sure you understand the following evaluation metric very clearly.

Evaluation Criteria: Specificity (1-5) – Evaluate whether the summary presents meaningful and relevant details by including concrete information rather than being vague or overly generic. Ensure that the summary provides clear, detailed points that directly relate to the review content.
•Score 1: The summary is extremely vague and generic, lacking any meaningful details. It uses broad generalizations without specific examples or descriptions.
•Score 2: The summary includes very few specific details. Most information is presented in a general way without concrete examples.
•Score 3: The summary includes some specific details but could be more precise. There is a mix of specific and generic information.
•Score 4: The summary provides good specific details in most areas. The information is concrete and meaningful, though some minor points could be more detailed.
•Score 5: The summary is highly specific throughout, providing precise and meaningful details. All information is concrete with relevant examples and descriptions.

[Review Text:] ...

[Summary:] ...

Evaluation Steps:
1. Read through the review text provided.
2. Read the corresponding opinion-centric insight summary.
3. Evaluate whether the summary provides specific, concrete details and avoids overly general or ambiguous statements.
4. Assign a score from 1 to 5 for Specificity, where 1 indicates very poor specificity and 5 indicates excellent specificity.

Table 20: The prompt for LLM-Judge Evaluation (Specificity).

LLM-Judge Evaluation prompt - (Insightfulness)

[Task Description]
You are provided with an online user's automobile review text along with an opinion-centric insight summary that groups user opinions at the topic level to offer insights. Your task is to evaluate the summary based on its insightfulness. Make sure you understand the following evaluation metric very clearly.

Evaluation Criteria: Insightfulness (1-5) – Evaluate whether the summary provides meaningful insights that enhance understanding or decision-making for the reader. Ensure that the summary goes beyond a simple restatement of the review to offer unique interpretations or conclusions that add value.
•Score 1: The summary provides no meaningful insights. It simply restates basic facts without adding any value for understanding or decision-making.
•Score 2: The summary offers very limited insights. Most information is superficial and does not help readers gain deeper understanding.
•Score 3: The summary provides some useful insights but could go deeper. It offers moderate value for understanding and decision-making.
•Score 4: The summary provides good insights in most areas. The information is valuable for understanding and decision-making, though some points could be more insightful.
•Score 5: The summary provides excellent insights throughout. All information meaningfully enhances understanding and is highly valuable for decision-making.

[Review Text:] ...

[Summary:] ...

Evaluation Steps:
1. Read through the review text provided.
2. Read the corresponding opinion-centric insight summary.
3. Evaluate whether the summary offers insightful, value-added interpretations that deepen understanding or guide decisions.
4. Assign a score from 1 to 5 for Insightfulness., where 1 indicates very poor Insightfulness. and 5 indicates excellent insightfulness.

Table 21: The prompt for LLM-Judge Evaluation (Insightfulness).

LLM-Judge Evaluation prompt - (Intent)

[Task Description]
You are provided with an online user's automobile review text along with an opinion-centric insight summary that groups user opinions at the topic level to offer insights. Your task is to evaluate the summary based on its Intent. Make sure you understand the following evaluation metric very clearly.

Evaluation Criteria: Intent (1-5) – Evaluate whether the summary accurately preserves the author's original tone, intent, and nuances without altering the emotional or stylistic essence of the review. Consider if the summary maintains the original sentiment and communication style throughout.
•Score 1: The summary completely fails to preserve the original tone and intent. The emotional essence and nuances are lost or significantly distorted.
•Score 2: The summary largely misrepresents the original tone and intent. Many nuances are missed or altered, though some basic sentiments remain intact.
•Score 3: The summary somewhat preserves the original tone and intent. Some nuances are captured while others are missed or altered.
•Score 4: The summary generally preserves the original tone and intent well. Most nuances and emotional elements are accurately captured, with only minor alterations.
•Score 5: The summary perfectly preserves the original tone, intent, and nuances. The emotional and stylistic essence is captured with complete accuracy.

[Review Text:] ...

[Summary:] ...

Evaluation Steps:
1. Read through the review text provided.
2. Read the corresponding opinion-centric insight summary.
3. Evaluate whether the summary maintains the original tone, intent, and nuanced details of the review without altering its emotional or stylistic qualities.
4. Assign a score from 1 to 5 for Intent, where 1 indicates very poor preservation and 5 indicates excellent preservation.

Table 22: The prompt for LLM-Judge Evaluation (Intent).

LLM-Judge Evaluation prompt - (Fluency)

[Task Description]
You are provided with an online user's automobile review text along with an opinion-centric insight summary that groups user opinions at the topic level to offer insights. Your task is to evaluate the summary based on its Fluency. Make sure you understand the following evaluation metric very clearly.

Evaluation Criteria: Fluency (1-5) – Evaluate whether the summary is naturally written, grammatically correct, and easy to read. Consider whether the summary flows smoothly, uses proper grammar, and avoids awkward phrasing, ensuring it is accessible to the reader.
•Score 1: The summary is extremely difficult to read. It is filled with errors, awkward phrasing, and lacks proper grammar.
•Score 2: The summary is difficult to read. It contains many errors, awkward phrasing, and lacks proper grammar.
•Score 3: The summary is somewhat difficult to read. It has some errors, awkward phrasing, and lacks proper grammar.
•Score 4: The summary is generally easy to read. It has few errors, awkward phrasing, and lacks proper grammar.
•Score 5: The summary is extremely easy to read. It is filled with errors, awkward phrasing, and lacks proper grammar.

[Review Text:] ...

[Summary:] ...

Evaluation Steps:
1. Read through the review text provided.
2. Read the corresponding opinion-centric insight summary.
3. Evaluate whether the summary is written in a natural and grammatically correct manner with clear and smooth flow.
4. Assign a score from 1 to 5 for Fluency, where 1 indicates very poor fluency and 5 indicates excellent fluency.

Table 23: The prompt for LLM-Judge Evaluation (Fluency).

Entity-feature-opinion tuple Annotation Prompt

[Task Definition]

You are a car opinion miner for the user. I want to perform feature-centric opinion extraction which is identifying attributes, specific features, or components mentioned in the text and associating opinions with each feature. To maximize coverage and diversity, you will generate a comprehensive set of tuples using your reasoning and understanding of the text. You will receive a user-generated car-related text. Your task is to extract all possible tuples in the form of (entity, feature, opinion) that exist within the text.

- **Entity**: the brand and model of a vehicle for which an opinion is expressed (e.g., "volvo xc90", "toyota camry", "nissan sentra").

- **Feature**: Feature: a specific characteristic, attribute, or component of the entity that users mention or evaluate. Includes aspects of performance, design, driving experience, features, systems, or parts (e.g., "ride comfort", "handling", "infotainment system", "interior", "brakes").

- **Opinion**: Opinion: a subjective or objective evaluation, reaction, or judgment about a feature. The opinion span must be taken exactly from the text and contain no more than 5 words. If the feature is implicit (not explicitly stated in the text), label it as "NULL".

- **Evidence Sentence**: the exact sentence from the input text that contains both the feature and the opinion (or where the opinion is clearly expressed about the implicit feature).

[Example]

- **Text**: The EyeSight system of Toyota Camry SE is amazing compared to Toyotas whatever it is.

- **Entity**: Toyota Camry SE

- **Feature**: EyeSight system

- **Opinion**: amazing compared to Toyotas whatever it is

[Annotation Requirements]

- All tuples must come only from the input text. Never invent or infer content that does not exist in the text.

- Do not hallucinate any information or generate tuples not explicitly supported by the input.

- Do not include any explanations, reasoning, or formatting outside the JSON object.

- The output must be in valid JSON format, and contain only the JSON object as follows:
```
{'opinion_tuple': [{'entity': <str>, 'feature': <str>, 'opinion':
<str>, 'evidence': <str>}, {'entity': <str>, 'feature': <str>,
'opinion': <str>, 'evidence': <str>}, ..., {'entity': <str>,
'feature': <str>, 'opinion': <str>, 'evidence': <str>}]}
```

[Content Text]

...

Table 24: OOMB Entity-feature-opinion tuple annotation prompt.

Entity-feature-opinion tuple Annotation & Verification Guideline

[Task Definition]
Your task is to analyze what kind of opinions users express about cars in online user-generated content.

- **Entity**: Car brand & model (e.g., volvo xc90, toyota camry, Nissan Sentra)

- **Feature**: A characteristic, attribute, or component of the mentioned entity (e.g., handling, ride comfort, infotainment system, ...)

- **Opinion**: The user's subjective/objective judgment, reaction, experience, evaluation, or feedback about the feature (e.g., terrible, incredibly stable)

[Example]

- **Text**: The EyeSight system of Toyota Camry SE is amazing compared to Toyotas whatever it is.

- **Entity**: Toyota Camry SE

- **Feature**: EyeSight system

- **Opinion**: amazing compared to Toyotas whatever it is

[Description]

- Each web page displays a online content along with a single (entity, feature, opinion) tuple at a time.

- A content can contain multiple **(entity, feature, opinion)** tuples.

- Sentence refers to the evidence sentence for the feature and opinion.

- **feature_exist, opinion_exist**: Whether the feature or opinion exists in the sentence.

- **sentence_exist**: Whether the sentence exists in the document on the right.

- If a feature appears in the content, it is highlighted in **bold red** in both the Document Info and Document Text.

- If an opinion appears in the document, it is highlighted in **bold blue**.

- The evidence sentence is highlighted in **bold black**.

- Each tuple is shown in the *Data Information* section, while the content text appears in the *Document Text* section. Annotators follow the process described below to perform accurate verification and refinement.

[Annotation & Verification Process]
(1) Entity Verification Check if the car brand and name (displayed in the entity field) is correctly identified from the document. The entity typically appears in titles, subtitles, or once in the main text. Confirm the context is indeed about this vehicle and the name is recorded exactly as it appears in the document.

(2) Opinion Existence Check Verify that opinions exist explicitly in the document as words or phrases. When an opinion is found in the text (highlighted in blue), confirm opinion_exist is marked 'O'. If no explicit opinion is found in any sentence, verify opinion_exist is marked 'X'.

(3) Feature-Opinion Relationship Verification Check if the feature (highlighted in red) being discussed in relation to the opinion exists explicitly in the document. For explicit features, verify feature_exist is marked 'O' and the exact text from the document is used. For implicit features (not explicitly mentioned), verify they are marked as 'NULL' but feature_exist is still 'O'.

(4) Sentence Documentation Check Verify that the evidence sentence (in bold black) properly contains either both feature and opinion (when they appear in the same sentence) or spans from feature to opinion (when they appear in different sentences). Confirm sentence_exist is marked 'O' when this evidence appears in the document.

(5) Duplicate Tuple Check When multiple identical or similar feature-opinion pairs exist in the same document, verify that only one correct sample is kept (with proper feature_exist and opinion_exist marks) while others are marked 'X' to avoid duplication.

Table 25: OOMB Entity-feature-opinion tuple annotation and verification guideline.

Opinion-centric-summary Annotation Prmopt

[Task Definition]
You are a product and marketing manager at a global automotive company. You will be given a piece of user-generated automotive content, along with its final gold-standard set of (entity, feature, opinion) tuples. Your goal is to write a free-form opinion-centric summary that organizes and synthesizes the content into high-level, insightful categories.
Here is definition of (entity, feature, opinion) tuple and example:

- **Entity**: the brand and model of a vehicle for which an opinion is expressed (e.g., "volvo xc90", "toyota camry", "nissan sentra").

- **Feature**: Feature: a specific characteristic, attribute, or component of the entity that users mention or evaluate. Includes aspects of performance, design, driving experience, features, systems, or parts (e.g., "ride comfort", "handling", "infotainment system", "interior", "brakes").

- **Opinion**: Opinion: a subjective or objective evaluation, reaction, or judgment about a feature. The opinion span must be taken exactly from the text and contain no more than 5 words. If the feature is implicit (not explicitly stated in the text), label it as "NULL".

This summary should be written in natural language, structured around the core features discussed, and clearly convey the most salient and recurring opinions.

Step 1. Read and Understand

- Thoroughly examine both the input content and the associated tuples.

- Identify the main opinion clusters and the intensity or frequency of key topics.

Step 2. Organize into High-Level Topics

- Group similar tuples into broad categories (e.g., "Interior Design", "Performance & Handling").

- Reflect on user sentiment and frequency to prioritize key points.

Step 3. Generate a Three-to-Five-Line Summary

- Line 1: Highlight the most frequently mentioned or emphasized features.

- Line 2: Emphasize features that were praised or criticized with notable intensity or consensus.

- Line 3+: Provide a concluding sentence summarizing the overall sentiment or takeaway.

Important Instructions

- The summary should be concise, natural, and informative—suitable for a marketing manager's report

- Ensure it covers the major opinions expressed in the tuple set

- The tone should be neutral and professional, without exaggeration

- Do not directly copy sentences from the content

- Do not mention the annotation process or describe how the summary was generated

- Do not include phrases like "Based on the review..." or "Here is a summary..."

- Carefully analyze whether the input reflects a single user or multiple users, and adjust nouns/pronouns accordingly ("user" vs. "users").

- Note: Return your results in JSON format only, with the following structure: {'summary': <str>}

[Content Text]
...

Table 26: OOMB: Opinion-centric summary annotation prompt.

Opinion-centric-summary Annotation & Verification Guideline

[Task Definition]
Your task is to analyze a text that summarizes the main topics and insights derived from online user-generated content. The provided summary is not a simple condensation, but an opinion-centric summary that synthesizes multiple users' perspectives to deliver high-level insights useful for marketing and strategic decision-making.

[Description]

- Each page displays a online content, a single gold tuple from the content, and its corresponding summary.

- Each summary is paired with a single gold tuple from the content and displayed on the page. When the page is turned, the next gold tuple appears.

- Each summary and a single gold tuple is shown in the *Data Information* section, while the content text appears in the *Document Text* section. Annotators follow the process described below to perform accurate verification and refinement.

[Annotation & Verification Process]
(1) Read & Understand Original Text and Summary
Carefully read the original user-generated content and Summary.

(2) Check Factual Accuracy
Thoroughly examine the summary for any hallucinations or factual inaccuracies that contradict information in the original content. Check whether all claims in the summary are directly supported by explicit statements in the source texts. Verify that no fabricated or assumed information is included, particularly for technical specifications, features, or entity attributes mentioned in the summary.

(3) Check Subject Coherence
Verify the correct attribution of opinions to appropriate subjects, considering the multi-user discussion context. Ensure opinions are not misattributed across different entities or users, especially in complex, multi-threaded discussions. Check that the summary properly distinguishes between individual opinions and collective sentiments when aggregating views from multiple users. Confirm that referenced features are associated with their correct corresponding entities.

(4) Check Sentiment Consistency
Check for sentiment polarity errors, particularly in cases involving irony, sarcasm, or nuanced expressions. Ensure that positive opinions are not mistakenly presented as negative and vice versa. Verify that the summary accurately captures the tone and emotional valence of the original opinions, including subtle sentiment expressions that may be context-dependent. Confirm that intensity modifiers (e.g., "very," "somewhat," "extremely") are appropriately preserved when they significantly impact the expressed opinion.

(5) Verify Three-Line Structure
Line 1 (Frequent or Emphasized Features) Verification

- Confirm that the summary accurately identifies and highlights features most frequently mentioned or emphasized in the original text

- Check frequency counts to verify that truly common themes are included in the summary

- Check if minority opinions aren't overrepresented or majority opinions underrepresented

Line 2 (Strongly Praised or Criticized Features) Verification

- Confirm the summary clearly identifies features that received particularly strong praise or criticism

- Verify that the intensity of opinions is accurately conveyed (using appropriate intensity indicators)

- Check that the distinction between mild opinions and strong sentiments is preserved

Line 3 (Conclusion/Key Takeaway) Verification

- Confirm the presence of a concise conclusion that synthesizes the overall sentiment

- Verify this conclusion accurately reflects the predominant message across all original comments

- Check that the conclusion doesn't introduce new information not supported by the original text

Table 27: OOMB: Opinion-centric summary annotation and verification process.

Input Content	Opinion-centric Summary
[YouTube] **Title:** Here's Why Everyone Hates the Mercedes-AMG GLE63 Coupe **Post #1 (Person 1):** I actually like the ways these look. They look big and aggressive **Comment #1 (Person 2):** Said no one ever **Comment #2 (Person 3):** not everyone has good taste. I dont like it because it looks good but the coupe roof line, and that's the point of the car. So it's like it could be perfect but it's not, to me at least. Looks wise. They should've changed the front too. **Comment #3 (Person 4):** no they look obnoxious and awkwardly proportioned. you'd have to be insane to spend over \$10K more for a less efficient, less practical version of a much better car in every aspect. **Comment #4 (Person 5):** If u like dome shaped overpriced. . . **Comment #5 (Person 6):** It's a fat sedan	The Mercedes-AMG GLE63 Coupe receives mixed reviews, with users criticizing its design as awkwardly proportioned, less efficient, and less practical compared to other models. Some find the coupe's roofline and overall look obnoxious and overpriced, while a minority appreciate its big and aggressive appearance. Overall, the sentiment leans negative, with the car's aesthetics and practicality being the main points of contention.
[Reddit] **Title**: What are some versions/generations of cars that have been mostly under the radar? **Post #11518 (Author: Person 1):** Y34 Infiniti M45.340hp V8 luxury sedan with strong, square styling. Roughly 9,000 imported over the two years on sale.Parts availability is a bitch. **Comment #11519 (Author: Person 2, Reply to reply comment #11518):** Good choice. An actual car that people dont talk about. There was also the M56 with a VK56 V8 that made 420 horsepower. Also rare. Heard one with an exhaust a while ago and it sounded pretty good. **Comment #11520 (Author: Person 3, Reply to reply comment #11519):** Yeah, the M56 is really cool and almost never gets talked about. Infiniti also continued the V8 when they renamed the M37M56 to the Q70, but a V8 Q70 is even rarer than an M56. **Comment #11521 (Author: Person 4, Reply to reply comment #11518):** And they look so rad. Quirky Japanese styling that wasnt super well received by the western market at the time, but few things look as good as a murdered out M45 **Comment #11522 (Author: Person 5, Reply to reply comment #11518):** Always loved these and thought of them as cars that should never have made it here stateside because of how JDM the styling is. Could be such a timeless car with some right modifications **Comment #11523 (Author: Person 6, Reply to reply comment #11518):** Most sinister looking car since the w109 Mercedes	The Y34 Infiniti M45 is celebrated for its strong, square styling and potential for timeless appeal with the right modifications, though parts availability is a challenge. Its quirky Japanese design was not initially well-received in the Western market, yet it is now appreciated for its unique and sinister look, comparable to the w109 Mercedes. The M56 variant, with its 420 horsepower V8, is noted for its rarity and impressive sound, while the V8 Q70 is even rarer, adding to the allure of these distinctive models.
[Reveiw Site] **Title**: Turn OFF Automatic Software Updates Took awhile but have grown from tolerating to appreciating this SUV. Replacement of a defective 12v battery and the dealer updating my car from the 2.0 to the 2.1 software resolved the initial issues that initially were bricking my XC60 B5. Second, I turned OFF the automatic software updates which you can do in check for software update submenu. Sticking with version 2.1 for now since it has made my car stable (at least on my B5 trim). 2.2 is available which includes Apple CarPlay but that version was JUST released and the initial bugs have not been worked out. So buggy that 2.2 is not available for an OTA update specifically for B5 trims since it causes known problems. While CarPlay is in high demand the native Google OS has been great and the Google map appears on the driver display which CarPlay currently wont. That said, recommend finding the software version that enables stability for your car/trim and then turn off automatic updates in the settings. For my XC60 B5 I am going to stick with 2.1 until forced to update or until the blogs indicate Volvo and Google are aligning better. Aside from the software gremlins this SUV is comfortable, super handsome in my r-design trim, and feels a touch unique.	The Volvo XC60 B5 receives mixed feedback, with users appreciating its comfort and unique design, particularly in the R-design trim. However, software issues are a significant concern, as the latest 2.2 version, which includes Apple CarPlay, is buggy and unavailable for OTA updates on B5 trims. Users recommend sticking with the stable 2.1 version and turning off automatic updates to maintain vehicle stability, while the native Google OS is praised for its functionality.
[Blog] **Title**: The Little Things: 2024 Hyundai Kona's Heat Buttons Are Oddly Satisfying The Hyundai Kona has a very open and room-like interior, with some well-placed buttons. I've been on an ergonomics kick lately, closely scrutinizing the button placement and cockpit layouts of every car I get into. The 2024 Hyundai Kona has a uniquely open and homey-feeling interior, but one aspect of it that struck me was how perfect the heated seat and steering wheel buttons are. Before your eyes roll out of your head at the thought of praising something as mundane as a button, consider this: Heated seat and steering wheel controls are features you might use every day, multiple times a day, for months and months while you own a car. It's fun to talk about max skidpad g's and 0-60 times, but realistically, how often do you take your vehicle to its extremes? When we're talking about what it's like to live with a car day in and day out, ergonomic touches can make or break the enjoyability of ownership. ...	The 2024 Hyundai Kona is praised for its interior design, featuring a uniquely open and homey feel with well-placed, physical buttons that enhance user engagement and ease of use. The heated seat and steering wheel controls receive high marks for their perfect placement and functionality, including a rare two-level heat setting for the steering wheel. Despite some criticism of its front-end design and limited acceleration, the overall sentiment is positive, highlighting the Kona's thoughtful ergonomic touches and practical features.

Table 28: Examples of OOMB opinion-centric-summaries by content type

Case Study: Review Site (1/2)	
Input Content	**Title**: Turn OFF Automatic Software Updates Took awhile but have grown from tolerating to appreciating this SUV. Replacement of a defective 12v battery and the dealer updating my car from the 2.0 to the 2.1 software resolved the initial issues that initially were bricking my XC60 B5. Second, I turned OFF the automatic software updates which you can do in check for software update submenu. Sticking with version 2.1 for now since it has made my car stable (at least on my B5 trim). 2.2 is available which includes Apple CarPlay but that version was JUST released and the initial bugs have not been worked out. So buggy that 2.2 is not available for an OTA update specifically for B5 trims since it causes known problems. While CarPlay is in high demand the native Google OS has been great and the Google map appears on the driver display which CarPlay currently wont. That said, recommend finding the software version that enables stability for your car/trim and then turn off automatic updates in the settings. For my XC60 B5 I am going to stick with 2.1 until forced to update or until the blogs indicate Volvo and Google are aligning better. Aside from the software gremlins this SUV is comfortable, super handsome in my r-design trim, and feels a touch unique.
Model Generated Summary	The reviewer initially faced issues with their Volvo XC60 B5 due to a defective 12v battery and problematic software updates, but these were resolved by upgrading to software version 2.1. They advise against automatic software updates, as version 2.2, though offering Apple CarPlay, is buggy and not recommended for B5 trims. The reviewer appreciates the native Google OS, especially its integration with the driver display, and plans to stick with version 2.1 for stability. Despite software challenges, they find the SUV comfortable, stylish in the R-design trim, and unique.
Gold Summary	The Volvo XC60 B5 receives mixed feedback, with users appreciating its comfort and unique design, particularly in the R-design trim. However, software issues are a significant concern, as the latest 2.2 version, which includes Apple CarPlay, is buggy and unavailable for OTA updates on B5 trims. Users recommend sticking with the stable 2.1 version and turning off automatic updates to maintain vehicle stability, while the native Google OS is praised for its functionality.
A3CU	60.91

Table 29: Case Study (1) of LLM-Judge Evaluation for the OIG Task. Summary generated by GPT-4o.

Case Study: Reddit (1/2)	
Input Content	**Title**: Ford starts deliveries of enhancedrange etransit **Comment 35276 (Author: Person 1)**: hate the trend of just using an ICE hull for an EV, so uninteresting and ugly. EV allows you to change the form factor of cars. I like new startups that make interesting and practical looking cars. **Comment 35277 (Author: Person 2, Reply to reply comment #35276)**: Its a work van. Fleets like having an already existing supply of commonly used and broken parts much more than they like having pretty looking vans. **Comment 35278 (Author: Person 3, Reply to reply comment #35276)**: Yeah but cars are the way they are for a lot of other reasons. And it actually makes a ton of financial sense. Some of the hypothetical benefits are not really as beneficial when insurance, repair, maintenance, learning curve for mechanics are taken into consideration. **Comment 35279 (Author: Person 4, Reply to reply comment #35276)**: Vans are gonna be fuckin boxes no matter what, its the point
Model Generated Summary	The review discussion centers around the design and practicality of electric vehicles (EVs) using traditional internal combustion engine (ICE) designs. Person 1 expresses dissatisfaction with the trend of using ICE hulls for EVs, finding them uninteresting and aesthetically displeasing, and prefers innovative designs from new startups. In contrast, Person 2 argues that for work vans, practicality and the availability of parts are more important than aesthetics. Person 3 supports this view, highlighting the financial and practical benefits of using existing designs, considering factors like insurance, repair, and maintenance. Person 4 bluntly states that vans are inherently boxy due to their functional purpose, implying that design changes are limited by their utility.
Gold Summary	The Ford e-Transit receives mixed feedback, with some users criticizing its design as uninteresting and ugly due to its use of an ICE hull, while others appreciate the practicality and financial sense of maintaining a familiar form. The vehicle's reliance on commonly used parts is seen as beneficial for fleet maintenance, despite the lack of aesthetic appeal. Overall, the sentiment reflects a divide between those valuing innovative design and those prioritizing practicality and cost-effectiveness.
A3CU	10.12

Table 30: Case Study (2) of LLM-Judge Evaluation for the OIG Task. Summary generated by GPT-4o.

Case Study: Review Site (2/2)

Faithfulness: 5	**Accuracy of Software and Battery Issues**: The summary correctly notes the initial issues the user faced due to a defective 12v battery and the problematic software updates, which closely aligns with what the review text describes. It also correctly mentions how these issues were largely resolved by upgrading to version 2.1.
	Software Updates and Recommendations: The summary accurately captures the user's advice against automatic software updates, specifically highlighting that version 2.2 is buggy, particularly for the B5 trim, which is consistent with the original review's observations. It also correctly represents the reviewer's intent to stick with version 2.1 for stability.
Coverage: 4	**Comprehensive Coverage of Main Issues**: The summary effectively highlights the initial issues faced by the reviewer, namely the defective 12v battery and problematic software updates, and how these were resolved by upgrading to software version 2.1. These are the critical components of the review text.
	Inclusion of Recommendation and Software Concerns: The insight summary correctly encapsulates the reviewer's advice against automatic updates due to the issues with version 2.2, emphasizing its buggy nature and incompatibility with B5 trims. It also mentions the preference to stay with version 2.1, another key opinion in the review.
	Minor Missing Details: Although the summary is generally accurate, it doesn't mention the specific aspect of finding a stable software version and then turning off automatic updates, which adds a nuanced detail to the reviewer's overall advice. This omission prevents the summary from achieving a perfect score.
Specificity: 4	**Resolution Details**: The summary covers the resolution process involving the upgrade to software version 2.1 to address initial vehicle issues. It explicitly mentions the defects in the 12v battery and problematic updates, demonstrating concrete detail awareness about the user's experience.
	Specific Features and Versions: The summary specifies the difference between software versions 2.1 and 2.2, noting the inclusion of Apple CarPlay in version 2.2. It captures the user's preference for version 2.1 for stability reasons and references the native Google OS's features, offering precise insights.
	General Impression and Recommendations: While largely specific in detail, certain generalizations are made, such as mentioning the comfort and uniqueness of the SUV without referencing specific features that contribute to these impressions. This is where the summary could benefit from slightly more precise detail, warranting a score of 4 rather than a 5.
Insightfulness: 4	**Comprehensive Coverage**: The summary effectively covers the key points of the review, including the issue with the battery, the software problems, and the resolution through the 2.1 software update. It emphasizes the importance of stability and cautions against automatic updates. This communicates the core guidance intended by the reviewer.
	Value-Added Advice: By recommending a strategic approach to software updates and emphasizing the stability offered by version 2.1, the summary provides actionable insights that can directly inform decision-making for potential users of the XC60 B5, particularly those concerned about software reliability.
	Balanced Perspective: The summary maintains a balance by acknowledging the non-software-related positives of the vehicle, such as comfort and aesthetic appeal, which enriches understanding for readers interested in a broader evaluation of the SUV's attributes. However, it could delve deeper into why the Google OS is particularly advantageous or why the SUV's design stands out, which slightly limits its insightfulness further.
Intent: 5	**Tone and Intent Preservation**: The summary effectively captures the reviewer's tone of cautious appreciation towards the vehicle, highlighting the resolve to prioritize stability over newer, but potentially problematic, software updates. The advice against automatic updates and emphasis on waiting for software stability clearly mirrors the reviewer's thoughtful, pragmatic approach.
	Emotional Nuances: The emotional essence of the original review, which oscillates between initial frustration and eventual appreciation, is well-preserved. The summary accurately reflects the transition from dissatisfaction due to battery and software issues to satisfaction with the SUV's comfort and aesthetic appeal.
	Detail and Specificity: Key details concerning the software update issues, particularly the problems associated with version 2.2 and the decision to stick with version 2.1, are thoroughly covered in the summary. The review's specific mentions of the native Google OS and Apple's CarPlay availability are adequately noted, maintaining the original's nuanced reflections.
Fluency: 5	**Natural Flow and Language Usage**: The summary is written in a manner that mimics a natural and conversational tone. It concisely recounts the original review's main points without introducing excessive technical jargon or unnecessary detail.
	Grammar and Sentence Structure: The summary is void of grammatical errors or awkward phrasing, making it easy for the reader to follow. Sentences are constructed well, contributing to a smooth reading experience.
	Clarity and Cohesion: Each point in the summary builds logically upon the previous, maintaining clear cohesion throughout. This makes it accessible even to readers unfamiliar with the context of the review text.

Table 31: Case Study of LLM-Judge Evaluation for the OIG Task. Summary generated by GPT-4o.

Faithfulness: 5	**Accurate Representation**: The summary accurately captures the main points and perspectives shared by the commenters without distorting their meaning. Each viewpoint, from aesthetics to practicality, is clearly represented as expressed in the original comments. **No Omitted Key Details**: The core opinions about the use of ICE hulls for EVs, the practical considerations for work vans, and the limited design flexibility due to utility needs are all present in the summary. This shows no significant omissions. **Absence of Hallucinated Information**: The summary does not introduce any information that was not present in the original text, ensuring that it stays true to the opinions and insights provided by the commenters. This maintains the integrity of the review discussion.
Coverage: 5	**Comprehensive Representation**: The summary thoroughly captures the primary opinions expressed by each participant in the review text. It details Person 1's dislike for using ICE hulls in EVs, Person 2's focus on practicality and cost-effectiveness discussed by Person 2, Person 3's considerations of financial benefits, and Person 4's functional view of van design. **Inclusion of Key Points**: No significant opinions are omitted. The contrasting views regarding aesthetic values versus practical functionality and financial considerations are clearly highlighted, showcasing the diverse perspectives discussed in the review. **Conciseness and Clarity**: The summary provides a concise yet clear representation of the review's key discussions, effectively relaying the main points without unnecessary details, ensuring all significant opinions are well-covered.
Specificity: 5	**Concrete Alignment with Comments**: The summary effectively captures the conversation between different users by directly referencing the points they make. It notes Person 1's dissatisfaction with ICE-based EV designs and preference for innovative startups, Person 2's emphasis on practicality and part availability for work vans, Person 3's mention of financial and practical factors, and Person 4's blunt assertion about vans' inherent boxiness. These details perfectly align with the provided comments, making the summary highly specific. **No Overgeneralization**: At no point does the summary veer into vague or generic statements. Each opinion is not only linked to the respective commenter but also articulated with clear explanation of their rationale, demonstrating a detailed and precise understanding of the user opinions. **Point-wise Clarity**: The summary maintains a structurally coherent format by presenting each perspective distinctly. The inclusion of each contributor's viewpoint provides a comprehensive view of the entire discussion, acknowledging the nuances in opinions rather than making an overarching generalization. This results in a precise and informative overview.
Insightfulness: 4	**Comprehensive Overview**: The summary effectively synthesizes various perspectives from the commentary, covering not only the dissatisfaction with traditional designs from Person 1 but also contrasting practical views from others who value utility over aesthetics. This balanced presentation allows readers to see different angles of the ongoing design debate, providing a fuller context. **Emphasis on Practical Factors**: By highlighting the points about practicality—such as access to parts, financial logic, and the vans' purpose—the summary goes beyond mere restatement of comments and offers a deeper understanding of the practical considerations influencing design decisions in the EV market. **Concludes with Functional Insight**: The summary concludes with Person 4's point about the inherent boxy design of vans due to their function, driving home the idea that practicality often trumps aesthetic innovation in vehicle design. This insight allows readers to grasp why some design choices remain seemingly unchanged.
Intent: 5	**Tone and Intent Preservation**: The summary accurately captures the tone and intent of each comment within the discussion. Person 1's dissatisfaction with the trend is clearly articulated, as is the practical-focused tone of Persons 2, 3, and 4. This indicates that the emotional elements of approval or disapproval were preserved effectively. **Nuance and Specificity**: The summary encapsulates the nuances in each participant's perspective. It highlights Person 1's preference for innovative design, contrasts this with the practical considerations highlighted by Person 2 and Person 3, and points out Person 4's straightforward viewpoint regarding the inherent functionality of vans. **Consistent Style and Balance**: The summary maintains a balanced view and communicates both sides of the argument without straying from the original intent of the discourse. This illustrates a good level of detail while maintaining the integrity and style of the original review, indicating an effective preservation of stylistic and emotional essence.
Fluency: 5	**Grammar and Syntax**: The summary is grammatically sound, with proper sentence structures, clear subject-verb agreement, and correct use of punctuation, enhancing readability. **Flow and Coherence**: The transition between points expressed by different persons is seamless, logically structured, allowing for a coherent understanding of conflicting viewpoints on EV and ICE designs. **Clarity and Readability**: The language used is straightforward and easy to follow, with precise vocabulary choices that appropriately convey complex opinions in an accessible manner.

Table 32: Case Study of LLM-Judge Evaluation for the OIG Task. Summary generated by GPT-4o.

Figure 8: OOMB – Annotation UI used for Entity-feature-opinion tuple set annotation. This example shows a case where both the feature and opinion are present in the content.

Figure 9: OOMB – Annotation UI used for Entity-feature-opinion tuple set annotation. This example shows a case where the feature is present in the content, but the opinion is not.

Figure 10: OOMB – Annotation UI used for opinion-centric summary annotation.

We are investigating the quality of summaries about specific cars from online contents.

We are **investigating the quality of summaries** that aggregate people's opinions about specific cars from online content.

You will be given a online content about a car and a summary summarizing people's opinions about cars as expressed in online content.
The writers have been instructed to group and synthesize various opinions about cars from online content to provide meaningful information to general users.
Your task is to **evaluate the summary based on 6 different criteria**.

[[Evaluation Criteria]]

[Criteria]
You need to evaluate summary based on below criteria, using a scale of 1 to 5, where 5 indicates the highest quality and 1 indicates the lowest quality.

- **Faithfulness**:
  Evaluate whether the summary faithfully reflects the original online content without distortion and check for any incorrect information.
- **Coverage**:
  Evaluate whether the summary effectively captures and represents the key opinions expressed in the online content.
- **Specificity**:
  Evaluate whether the summary presents meaningful and relevant details rather than being vague or overly generic.
- **Insightfulness**:
  Evaluate whether the summary provides meaningful insights that enhance understanding or decision-making for the reader.
- **Intent**:
  Evaluate whether the summary accurately preserves the author's original tone, intent, and nuances without altering the emotional or stylistic essence of the online content.
- **Fluency**:
  Evaluate whether the summary is naturally written, grammatically correct, and easy to read.

Online content about the car
${text}

Summary
${summary}

Figure 11: The interface for human evaluation (Instruction part).

Faithfulness: Evaluate whether the summary faithfully reflects the original online content without distortion and check for any incorrect information.

○ 1   ○ 2   ○ 3   ○ 4   ○ 5

1. The summary completely distorts the original online content and contains many incorrect information. It cannot be trusted at all.

2. The summary significantly misrepresents the original online content with several incorrect information.

3. The summary partially reflects the original online content but has some minor incorrect information.

4. The summary significantly misrepresents the original online content with several incorrect information.

5. The summary completely reflects the original online content without any distortions and incorrect information.

Figure 12: The interface for human evaluation (Faithfulness).

Coverage: Evaluate whether the summary effectively captures and represents the key opinions expressed in the online content.

○ 1   ○ 2   ○ 3   ○ 4   ○ 5

1. The summary fails to capture any key opinions from the online content. The content is either completely missing or irrelevant to the original opinions.

2. The summary captures only a small portion of key opinions. Many important opinions from the online content are missing.

3. The summary captures some key opinions but misses others. The coverage is partial and could be more comprehensive.

4. The summary effectively captures most key opinions from the online content. The coverage is good but may miss minor details.

5. The summary comprehensively captures all key opinions from the online content. Nothing important is missing, and the coverage is complete.

Figure 13: The interface for human evaluation (Coverage).

Specificity: Evaluate whether the summary presents meaningful and relevant details rather than being vague or overly generic.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

1. The summary is extremely vague and generic, lacking any meaningful details. It uses broad generalizations without specific examples or descriptions.

2. The summary includes very few specific details. Most information is presented in a general way without concrete examples.

3. The summary includes some specific details but could be more precise. There is a mix of specific and generic information.

4. The summary provides good specific details in most areas. The information is concrete and meaningful, though some minor points could be more detailed.

5. The summary is highly specific throughout, providing precise and meaningful details. All information is concrete with relevant examples and descriptions.

Figure 14: The interface for human evaluation (Specificity).

Insightfulness: Evaluate whether the summary provides meaningful insights that enhance understanding or decision-making for the reader.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

1. The summary provides no meaningful insights. It simply restates basic facts without adding any value for understanding or decision-making.

2. The summary offers very limited insights. Most information is superficial and does not help readers gain deeper understanding.

3. The summary provides some useful insights but could go deeper. It offers moderate value for understanding and decision-making.

4. The summary provides good insights in most areas. The information is valuable for understanding and decision-making, though some points could be more insightful.

5. The summary provides excellent insights throughout. All information meaningfully enhances understanding and is highly valuable for decision-making.

Figure 15: The interface for human evaluation (Insightfulness).

Intent: Evaluate whether the summary accurately preserves the author's original tone, intent, and nuances without altering the emotional or stylistic essence of the online content.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

1. The summary completely fails to preserve the original tone and intent. The emotional essence and nuances are lost or significantly distorted.

2. The summary largely misrepresents the original tone and intent. Many nuances are missed or altered, though some basic sentiments remain intact.

3. The summary somewhat preserves the original tone and intent. Some nuances are captured while others are missed or altered.

4. The summary generally preserves the original tone and intent well. Most nuances and emotional elements are accurately captured, with only minor alterations.

5. The summary perfectly preserves the original tone, intent, and nuances. The emotional and stylistic essence is captured with complete accuracy.

Figure 16: The interface for human evaluation (Intent).

Fluency: Evaluate whether the summary is naturally written, grammatically correct, and easy to read.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

1. The summary is extremely difficult to read. It is filled with errors, awkward phrasing, and lacks proper grammar.

2. The summary is difficult to read. It contains many errors, awkward phrasing, and lacks proper grammar.

3. The summary is somewhat difficult to read. It has some errors, awkward phrasing, and lacks proper grammar.

4. The summary is generally easy to read. It has few errors, awkward phrasing, and lacks proper grammar.

5. The summary is extremely easy to read. It is filled with errors, awkward phrasing, and lacks proper grammar.

Figure 17: The interface for human evaluation (Fluency).