A WHAT TO DO NEXT WITH PANOSENT?

In this paper, we introduce a novel benchmark for Multimodal Conversational Aspect-based Sentiment Analysis, which includes two innovative subordinate tasks: Panoptic Sentiment Sextuple Extraction and Sentiment Flipping Analysis. We have proposed the Chain-of-Sentiment reasoning method based on our MLLM, which has demonstrated strong benchmark performance on our dataset, PanoSent. We firmly believe that this pioneering work will inaugurate a new era for the sentiment analysis community. Several important directions for future research can emerge from our work.

► **Exploring Multimodality in PanoSent** In this paper, we encode multimodal information in a straightforward manner using common techniques. Given the critical role of multimodal information for this task, future efforts should focus on developing more powerful methods for multimodal feature extraction and integration. Additionally, investigating the impact of different modalities on sentiment recognition across various scenarios promises to be a fruitful area of research.

► **Identifying Implicit Sentiment Elements** Compared to explicit sentiment elements, the identification of implicit elements poses a greater challenge. Our approach, based on MLLM, autonomously determines the recognition of implicit sentiment elements through an understanding of the input data's content. We believe there are more accurate methods to be discovered for identifying implicit elements.

► **Sentiment Cognition and Reasoning Mechanisms** Our new task involves complex sentiment cognition, for which we propose a reasoning framework. Future research should delve deeper into the mechanisms of interaction and triggering among sentiment elements, as well as the mechanisms behind Sentiment flipping, in order to develop more robust sentiment reasoning solutions.

► **Modeling Dialogue Context** Dialogue scenarios closely resemble the natural ways in which people express emotions. This work processes the overall content of dialogues through the model, allowing it to understand conversations autonomously. Next steps in research could focus on how to more effectively enhance the model's ability to model dialogue context, thus better addressing cross-utterance issues. For example, further consideration could be given to modeling dialogue structure and speaker coreference resolution features.

► **Sentiment-aware Instruction Fine-tuning** Our work involves tasks based on a MLLM, which is fine-tuned on our training set. Research indicates that the setup of instruction fine-tuning significantly affects the LLM's performance on downstream tasks. We believe that developing superior methods for instruction fine-tuning, such as designing approaches that increase the LLM's sensitivity to sentiment, holds great promise.

► **Cross-lingual Transfer Learning** Our dataset includes three popular languages from different language families: English, Chinese, and Spanish, with non-parallel annotations across languages. Subsequent research could explore cross-lingual transfer learning in a multimodal scenario, investigating the supportive role of language-invariant features (multimodal information) for sentiment learning across languages.

► **Cross-domain Transfer Learning** Our dataset is extensive, covering hundreds of different domains and everyday scenarios. It would be interesting to study the variations of panoptic sentiment across different scenes and domains, making cross-domain transfer learning a meaningful direction for future work.

► **Weak/Unsupervised Sentiment Analysis** Our paper primarily focused on supervised learning using a large amount of annotated data. However, MLLMs already possess significant unsupervised generalization capabilities. It is crucial to leverage our benchmark for weak or even unsupervised sentiment recognition. In the subsequent part of the Appendix, we provide an analysis and exploration of few-shot sentiment recognition.

B ETHIC CONSIDERATIONS

In conducting this research and developing the PanoSent benchmark, several ethical considerations have been taken into account to ensure the responsible use and application of the technologies involved.

► **Privacy and Data Protection** Given that the raw dataset includes multimodal dialogues that may contain personal information, rigorous measures have been implemented to anonymize and protect any potentially sensitive data. This includes the removal of personally identifiable information (PII) from texts, images, audio, and video content. Additionally, the dataset has been reviewed to ensure compliance with relevant data protection regulations such as GDPR and CCPA, aiming to respect user privacy fully. Our data collection procedures have been carefully designed to focus on factual knowledge acquisition without infringing on privacy rights, thereby upholding our strong commitment to privacy and ethical research standards.

► **Data Collection** For the creation of the PanoSent dataset, all data was collected from publicly available sources or through contributions from individuals who were informed about the purposes of the research and provided their explicit consent. Efforts were made to ensure that contributors understood their rights, including the right to withdraw their data at any point.

► **Annotator and Compensation** Acknowledging the significant role of human annotators in the creation of the PanoSent dataset, we have engaged a diverse group of annotators including well-trained individuals from crowdsourcing platforms, native speakers, and senior postgraduate students with specialized training for the annotation tasks. The estimated time required for annotating each dialogue utterance is between 4 to 6 minutes, reflecting the complexity and detailed nature of the task. Annotators are compensated at a rate of $0.50 for each dialogue they complete, which is designed to fairly reflect the effort and skill involved. Additionally, the compensation for linguists and native speakers involved in the project is determined based on the average time commitment, ensuring fair and equitable remuneration for their expertise and contribution.

► **Intellectual Property Protection** The PanoSent dataset includes content collected from publicly available sources on a popular Chinese social media platform, utilizing its officially open API. This collection method ensures compliance with intellectual property laws and respects the terms of service of the platform. Permission for the use, distribution, and modification of this content is granted under the terms of the Weibo API distribution agreement. This approach safeguards the intellectual property rights of the content creators while facilitating academic research and development.

► **Bias and Fairness** Recognizing the potential for bias in AI systems, this research includes an analysis of the PanoSent dataset for biases related to gender, ethnicity, language, and other sociodemographic factors. Steps have been taken to mitigate these biases through diverse and representative data collection across multiple languages and scenarios. However, it is acknowledged that complete eradication of bias is challenging, and continuous efforts are required to identify and address biases as the benchmark evolves.

► **Misuse Potential** The research team is aware of the potential misuse of sentiment analysis technologies, such as applications in surveillance or the manipulation of public opinion. Therefore, alongside the release of the PanoSent benchmark and the associated models, guidelines have been developed to encourage ethical use. These guidelines emphasize the importance of consent, transparency, and accountability in any application or further development of the technologies presented in this paper.

► **Accessibility and Inclusivity** In line with our commitment to fostering an inclusive research community, all code and data related to the PanoSent benchmark will be made openly available. This ensures that researchers and practitioners from diverse backgrounds and with varying levels of resources have equal opportunities to contribute to, and benefit from, the advancements in multimodal conversational aspect-based sentiment analysis.

C MORE DETAILS OF DATASETS

C.1 Extended Details of Data Construction

C.1.1 Data Acquisition

► **Step1. Platform Selection and Data Collection.** Our initial step involves identifying a diverse range of social media and forum platforms as sources for our dataset, including but not limited to Twitter, Facebook, Reddit, Weibo, Xiaohongshu, BeReal. These platforms are chosen for their rich conversational content across multiple languages and the vast user engagement they facilitate. We target some influential bloggers within specific domains and the discussions surrounding trending topics related to our research themes. Conversations on these platforms typically originate from a root post, with users participating in multi-thread and multi-turn dialogues based on the initial post. In addition to text, these interactions often include multimodal content such as images, videos, and audios. While less common than text, this multimodal interaction is a crucial component of our dataset, and we make extra efforts to collect conversations incorporating these elements. Given that these platforms generally do not support audio replies as a standalone feature, we extract the audio tracks from video content to collect audio modal information. Data collection is automated through publicly available APIs provided by these platforms, with conversations being categorized based on their thematic relevance and the types of modal information they contain. The process of data acquisition and preprocessing is depicted in Figure 7. Following the collection process, a total of approximately 24,000 instances of data are gathered.

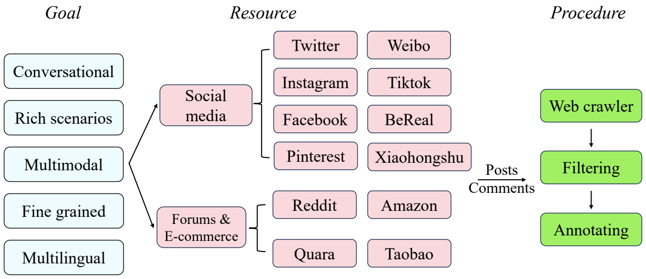


Figure 7: The workflow of data acquisition and preprocessing.

► **Step2. Data Cleaning and Re-organization.** To ensure the dataset is free from harmful content, privacy violations, irrelevant, or low-quality conversations, we employ a combination of manual inspection and automated tools. A keyword library is constructed based on previous related studies and the expertise of team members in social media analysis and specific thematic areas. This library includes keywords indicating potential harm, privacy infringement, and irrelevance to the research topic. Scripts are developed to automatically scan the collected data for these keywords, with flagged conversations undergoing manual review to determine their suitability for inclusion in the dataset. Additionally, we utilize the Toxic BERT model, capable of identifying various forms of harmful speech, including insults, discrimination, and harassment, by analyzing extensive online textual data. This model provides probability scores for detected categories and identifies the specific locations of toxic speech within the text. The output from the model is also subject to manual review, considering the context of the conversations to make final decisions on content inclusion. Multimodal content is manually reviewed due to its relatively lower volume, focusing not only on the potential harm but also on the relevance of the content to the conversation, with any mismatched multimodal content being removed. After the cleaning process, approximately 18,000 instances of data remained.

C.1.2 Human Annotation

We have recruited a team of annotators who possess relevant background knowledge, including well-trained individuals from crowdsourcing platforms, native speakers, and senior postgraduate students. Before commencing manual annotation, we developed detailed annotation guidelines based on the definitions from SemEval related to ABSA and the specific requirements of our task. All annotators have undergone uniform training to ensure consistency and objectivity in their work. Based on the task's complexity and the time needed for careful annotation, we estimate that annotators will require 4 to 6 minutes per data entry. Therefore, we have decided to pay annotators $0.50 per dialogue to acknowledge the effort required for accurate and detailed annotation. Each piece of data has been annotated by at least three independent annotators, and we have calculated the Cohen's Kappa Score to measure the consistency among them. Achieving a score of 0.88, which reflects the high quality of our annotated dataset, data with Kappa scores below a predefined standard undergo review and discussion. In cases of disagreement or ambiguity, linguists and native speakers collaborate to reach a consensus. Data that cannot reach consensus or remains ambiguous is discarded to maintain the quality of the dataset. Following manual inspection and annotation, the dataset has been further refined, resulting in a final dataset size of 9,280.

C.1.3 Automatic Synthesis

Our task mandates rigorous data requirements, necessitating dialogue context that is fine-grained enough to encompass all six defined elements, includes both implicit and explicit expressions, and incorporates multimodal information. Given that only a minuscule proportion of real-world data meets these criteria, and considering the proven success of LLMs in generating data, we have opted to utilize the capabilities of GPT-4 for automated data generation and corresponding element annotation. The process unfolds in several steps.

► **Step1. Creation of Dialogue Instances.** Drawing from high-quality real dialogues, we meticulously crafted a small batch of dialogue instances tailored to our task's needs. These instances display diversity in themes, participant count, length, turn-taking, reply structure, and types of included multimodal information. They undergo multiple rounds of modification and inspection by our team to ensure comprehensive coverage and quality.

► **Step2. Prompt Template Design and Data Generation.** We develop structured and coherent prompt templates to guide GPT-4[3] in understanding our requirements and generating dialogue data that aligns with them. After several iterations of adjustments and tests, we finalize a prompt template. This template instructs GPT-4 not only to generate dialogues but also to annotate them with the defined sextuples and identify instances of sentiment flips. Moreover, for certain dialogue utterances, GPT-4 is tasked with creating suitable captions as placeholders for images, audios, and videos, reflective of the context. Approximately 20,000 pieces of data are generated using this methodology.

> **An example of our prompt template**:
> You are a professional and creative playwright on dialogues related to 'Televisions'. Please comply with the following instructions. Do not comment, judge, or output other texts and only return the results.
> 1. Generate a nonlinear dialogue replying structure among 4 speakers, and the turns of the dialogue must be 3.
> 2. Each speaker in the dialogue should have a unique 'speaker_id' and a unique 'speaker_name', and each dialogue should have a unique 'doc_id'.
> 3. Dialogue should incorporate discussions around a specific 'aspect' of a 'target', attributed to a 'holder'. Such discussion, or 'opinion', evaluates or comments on the 'target', supported by 'rationale' explaining the reasoning behind these opinions.
> 4. Annotate and 'order' the occurrence of 'holder', 'target', 'aspect', 'opinion', and 'rationale' in HTML format in the 'annotation'.
> 5. Every utterance except the first utterance is a reply to dialogue sentence with index n, the reply property of this utterance should be n, the first utterance is -1.
> 6. The conversation needs to be granular enough to include all the following five elements: 'holder', 'target', 'aspect, 'opinion', and 'rationale'.
> 7. You need to store all five parts of the conversation content tag according to the format of the following example. And you should decide a sentiment for every combination based on the corresponding opinion, and the sentiment must be one of positive, negative, neutral.
> 8. Use your excellent imagination and strong content generation to add image and video modalities in the conversation. Please ensure the relevance of the image and video annotations and use the '<img> content </img>' format to annotate. If the utterance has an image or video annotation, the 'modality' should include 'type', 'caption', 'id'; the 'type' is always 'img' or 'video', the 'caption' is the corresponding annotation. If there is no image or video annotation, the 'modality' should be set to 'None' .
> 9. The dialogue should include at least one clear instance where a holder exhibits a change in sentiment towards an aspect of a target, triggered specifically by the 'participant feedback and interaction'.
> 10. Please ensure you fully comprehend the example provided and apply it to create a dialogue that fulfills all specified criteria. For instance, a sample json output would be: {sample_json_string}

► **Step3. Multimodal Information Retrieval.** With the annotated dialogues, we use the captions to retrieve the piece of information in the corresponding modality (image, audio, or video) from extensive databases such as COCO, Flickr30k for images, AudioSet and WaveText5K for audios, and WebVid for videos. These databases, rich in (image, audio, or video)-caption pairs, enable us to match dialogue captions with database captions using SentenceTransformer[4], focusing on the top-10 most similar candidates for each modality. For the associated multimodal content, three annotators score each of the ten candidates on a scale of 1-10. The content with the highest average score is selected as the definitive multimodal segment. Should none of the candidates meet the desired criteria—indicating a lack of suitable matches within the databases—we resort to direct retrieval from the Google search engine[5] to ensure exhaustive inclusivity.

► **Step4. Manual Review.** Each generated dialogue, along with annotations related to the two sub-tasks and multimodal content, undergoes a thorough review by at least two staff members. Any potentially problematic instances are discarded. We also calculate the Cohen's Kappa Score, achieving a score of 0.82, which attests to the consistency and validity of our annotation process. Following this rigorous review process, 10,720 data instances remained.

In Table 6, we illustrate a complete data instance (a conversation) with our annotation (English version is shown).

C.2 Detailed Summary of Dataset Insights

Here, we extend the content of Section §4.2 from the main article, to provide a more comprehensive introduction to all the highlights of our dataset.

► **Panoptic Fine-grained Sentiment Definition.** Compared to existing ABSA datasets, the PanoSent dataset stands out for its fine-grained and exhaustive annotation of sentiment elements, featuring six key items essential for ABSA: holder, target, aspect, opinion, sentiment, and rationale. The 'holder' represents the entity expressing the viewpoint, which, despite frequently being the speaker in conversational contexts, can also encompass instances where the holder is not the speaker. The 'target' pertains to the subject of discussion, such as a digital gadget, a service, or an activity. 'Aspect' refers to specific attributes or facets of the target, for example, the battery, screen, or camera quality of a smartphone. 'Opinion'

[3] *gpt-4-1106-preview* version API, https://openai.com/gpt-4

[4] https://huggingface.co/sentence-transformers

[5] https://www.google.com/

Table 6: A snippet of an annotated data instance in PanoSent dataset.

Key	Value
Dialogue-ID	00024
Dialogue	1. Ava: I recently purchased a new digital camera, and its image quality is absolutely stunning, capturing every detail with such clarity and vibrant colors that photos almost look lifelike. 2. Liam: That sounds amazing! What about its low-light performance? Does it capture sharp and clear images in low-light conditions? 3. Ava: The low-light performance is quite impressive. It captures sharp and clear images even in dimly lit environments. 4. Mia: That's great to hear! How about its video recording? 5. Ava: The camera excels in video recording. It captures high-quality videos with excellent stabilization, resulting in smooth and professional-looking footage. 6. Noah: What about its battery life? 7. Ava: The battery life is disappointing. It drains quickly, requiring frequent recharging. 8. Liam: It's worth noting that the camera's advanced features naturally demand more power, which is common for high-performance devices. Compared to similar models, our camera holds up well in terms of battery life, making it a fair trade-off for its quality. 9. Ava: That's a good point. Considering the advanced features and comparing it with other cameras, the battery life does seem acceptable. I hadn't looked at it that way before.
Replies	-1, 0, 1, 2, 3, 0, 5, 6, 7
Speakers	0, 1, 0, 2, 0, 3, 0, 1, 0
Holders	Ava, Liam
Targets	digital camera
Aspects	image quality, low-light performance, video recording, battery life
Opinions	absolutely stunning, quite impressive, excels in, disappointing, acceptable
Sextuples	(Ava, digital camera, image quality, absolutely stunning, positive, capturing every detail with such clarity and vibrant colors that photos almost look lifelike) (Ava, digital camera, low-light performance, quite impressive, positive, captures sharp and clear images even in dimly lit environments) (Ava, digital camera, video recording, excels in, positive, captures high-quality videos with excellent stabilization, resulting in smooth and professional-looking footage) (Ava, digital camera, video recording, excels in, positive, captures high-quality videos with excellent stabilization, resulting in smooth and professional-looking footage) (Ava, digital camera, battery life, disappointing, drains quickly, requiring frequent recharging) (Liam, digital camera, battery life, holds up well, positive, compared to similar models) (Ava, digital camera, battery life, acceptable, neutral, considering the advanced features and comparing it with other cameras)
Sentiment Flip	Holder-Target-Aspect: (Ava, digital camera, battery life) Initial Sentiment-Flipped Sentiment: (negative, neutral) Trigger Type: logical argumentation

denotes the expressed view or judgement, while 'sentiment' captures the emotional polarity associated with the opinion, classified as positive, neutral, or negative. Finally, 'rationale' elucidates the underlying reasons or justifications that give rise to a particular opinion. This meticulous approach to sentiment analysis not only enhances the depth of understanding around each conversational element but also significantly advances the precision and applicability of ABSA methodologies in dissecting and interpreting complex dialogues.

► **Cognitive Causal Rationale.** We not only prioritize the identification of sentiment states and the granularity of emotional details within dialogues but also emphasize the significance of understanding the underlying reasons behind expressed opinions. Building on this premise, we introduce the rationale element into ABSA for the first time, refining its definition to include a focus on the motivations behind sentiments. This approach aids in a more comprehensive analysis from a logical perspective, unveiling the catalysts behind viewpoints and attitudes, thereby enriching the extraction of deeper semantic insights.

► **Dynamic Sentiment Flipping.** In the complex scene of dialogues, analyzing dynamic sentiment changes is crucial. Participants in a conversation may alter their previous viewpoints and attitudes due to various triggers, a vital aspect for understanding the progression of events and emotional trends within dialogues, such as changes in characters' psychological states. This dynamic aspect of sentiment, however, has not been addressed in existing ABSA research. To comprehend the intricate dynamics of sentiment within multiparty dialogues, we categorize four distinct and clearly defined types of triggers that can lead to sentiment flips: **introduction of new information**, **logical argumentation**, **participant Feedback and interaction**, and **personal experiences and self-reflection**. Each of these triggers plays a critical role in the natural evolution of sentiment within conversations, providing a deeper insight into the fluid nature of human emotions and thoughts in dialogue contexts.

Table 7: Detailed categorization of domains in PanoSent dataset.

Principal Domains	Sub-Domains
Electronic Products	Smartphones, Personal Computers, Televisions, Wearable Technology, Cameras, Audio Systems, Gaming Hardware, Home Automation, Tablets, Drones, Smart Home Devices, E-Readers
Technology	Artificial Intelligence, Blockchain, Virtual Reality, Cybersecurity Measures, Cloud Solutions, Quantum Devices, Robotics, Network Innovations, Sustainable Energy, Advanced Biotech, Space Exploration Technologies
Fashion	High Fashion, Urban Streetwear, Designer Brands, Vintage Apparel, Accessories, Children's Wear, Sportswear, Sustainable and Ethical Fashion, Techwear, Seasonal Collections
Food and Cuisine	Plant-based Cuisine, Global Street Eats, Gourmet Dining, Mobile Food Services, Regional Delicacies, Sweets and Confectionery, Health-conscious Foods, International Fusion, Culinary Skills, Beverage Crafting
Movies and Entertainment	Major Studio Releases, Indie Films, Documentaries, Streaming Originals, Celebrity Culture, Awards Season, Reality Shows, Animation, Genre Cinema, Film Festival, Web Series, Fan Culture and Fandom
Health and Wellness	Mental Health Awareness, Fitness Regimens, Dietary Plans, Mindfulness and Meditation, Retreats for Wellbeing, Holistic Medicine, Beauty and Dermatology, Sleep Science, Nutritional Supplements, Wellness Gadgets
Finance and Economy	Equities Market, Savings and Budgeting, Property Market, Pensions and Retirement, Fiscal Policies, Insurance Schemes, Trading Strategies, Financial Tech, International Commerce, Crypto Assets
Sports and Athletics	Team Sports, Basketball, Racquet Sports, Olympic Disciplines, Adventure Sports, Digital Gaming Competitions, Gymnastics, Aquatic Activities, Motorsport, Outdoor Challenges, E-Sports Technology, Urban Sports and Street Games
Travel and Tourism	Offbeat Adventures, Cultural Expeditions, Green Travel, Opulent Journeys, Economical Excursions, Sea Cruises, Solo Explorations, Family Getaways, Heritage Sites, Gastronomic Tours
Art and Culture	Modern Art, Musical Variations, Performing Arts, Literary Works, Exhibition Spaces, Cultural Celebrations, Photographic Arts, Sculptural and Installations, Traditional Crafts, New Media Art

1) **Introduction of New Information** encapsulates instances where new data, research findings, news reports, or previously undiscussed information are introduced into the dialogue. Such information can alter or influence participants' understanding or emotional stance toward a topic.

2) **Logical Argumentation** involves constructing arguments through logical reasoning and analysis using known information or consensus. This trigger uses structured and persuasive logic to convince participants to adopt a viewpoint through rational analysis.

3) **Participant Feedback and Interaction** focuses on the direct feedback and interactions among participants in the dialogue, including opposition, questioning, or other forms of direct response. This category emphasizes how direct interpersonal communication can influence shifts in emotional stances.

4) **Personal Experiences and Self-reflection** covers instances where individuals trigger a change in their emotional stance by describing their own experiences, reflecting on their perceptions or experiences. This trigger is internal, based on personal memories and their current evaluation.

► **Multi-scenario.** PanoSent positions dialogue as its contextual backbone, incorporating 10 primary real-life domains that span over 100 sub-domains, thereby ensuring a broad diversity to facilitate research into sentiment analysis from a variety of perspectives. The 10 main domains include electronic products, technology, fashion, food and cuisine, movies and entertainment, health and wellness, finance and economy, sports and athletics, travel and tourism, and art and culture. Data within each main domain vary in distribution, and each domain encompasses at least 10 sub-domains. The specific classifications and details of these sub-domains are illustrated in Table 7, while the distribution of categories within each domain is depicted in Figure 8.

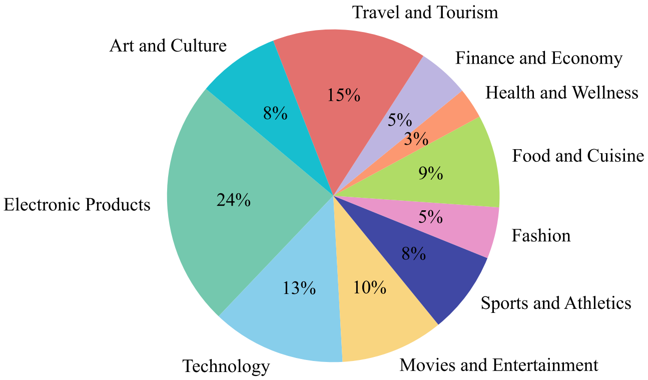


Figure 8: Distribution of categories within each domain.

► **Multimodality.** Our PanoSent dataset showcases a structured amalgamation of multimodal content within dialogues, reflecting the diverse interaction types prevalent in human communication. As elucidated in Figure 9, the majority of the dialogues remain text-based. Beyond text, certain dialogues are enriched with images, audios, or videos, thereby integrating visual and auditory dimensions into the textual conversations. The additional modalities include images (23%), audio (4.5%), video (4.5%), and mixed modalities (14%). The mixed modalities encompass combinations like image-audio (IA), image-video (IV), audio-video (AV), and image-audio-video (IAV). We ensure these non-textual modalities are abundant, relevant, and of high quality, aligning closely with the dialogue content.

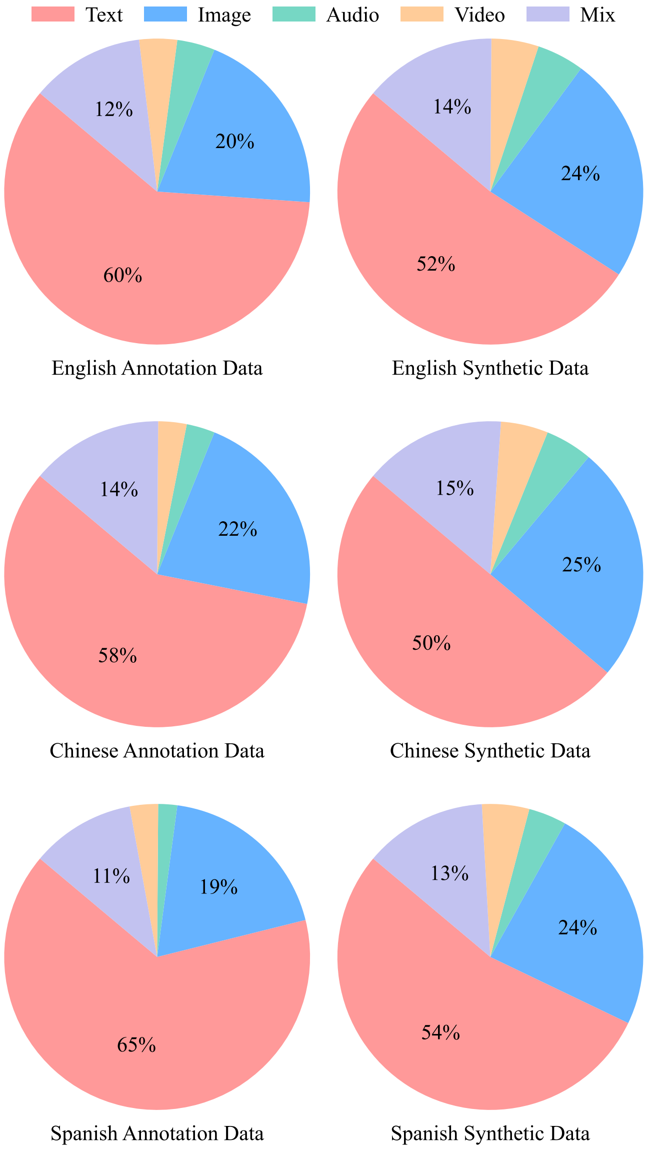


Figure 9: Distribution of multimodal composition in different languages and types.

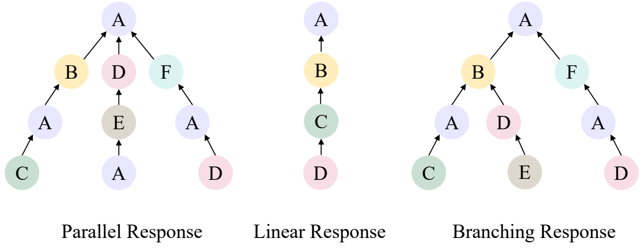


Figure 10: Different replying structure of dialogue.

► **Multilingualism.** PanoSent encompasses dialogues in three predominant languages: English (60%), Chinese (30%), and Spanish (10%), facilitating cross-lingual research in ABSA. To ensure the accuracy and standardization of annotations across each language, we employ online grammar checking tools for preliminary validation of the annotations. Additionally, we engage several native speakers for each language to conduct manual reviews and corrections, guaranteeing that the data annotations are not only standardized but also precise. This meticulous approach ensures the dataset's reliability for cross-lingual sentiment analysis studies.

► **Implicit ABSA.** Our dataset comprehensively accommodates implicit ABSA, thereby introducing heightened challenges into the field. Although most sextuple elements are explicitly mentioned in the utterance text, about 28% of sextuples include implicit elements that need to be inferred from the context of information presented across various modalities.

Contrastingly, most existing studies have predominantly focused on the extraction of explicit elements, largely overlooking the implicit dimensions. In reality, whether it's product reviews, daily conversations, or dialogues in other scenarios, a substantial portion comprises implicit elements. Hence, implicit elements are exceedingly common and should not be disregarded. This emphasis underscores the necessity of integrating both explicit and implicit element analysis to fully capture the nuances and complexities of sentiment in diverse communicative contexts.

► **Cross-utterance and inner utterance.** Given that elements of the same sextuple can originate from multiple distinct utterances, potentially spanning across two, three, or even more utterances, the extraction of information spanning multiple utterances poses greater demands on the model's capabilities. Our dialogue dataset includes such instances, laying a foundation for subsequent exploration and research. This consideration highlights the intricate dynamics of conversation analysis, emphasizing the necessity for models to adeptly navigate and interpret cross-utterance and inner-utterance relationships to fully understand the context and sentiments expressed.

► **Rich dialogue replying structure.** Commonly, every dialogue starts with a root post, with multiple users (speakers) participating by replying to previous utterances. Consequently, the diversity of a dialogue is manifested not only in superficial distinctions, such as the number of participants or the number of turns within the dialogue but also in the deeper variations of the reply structure. We have taken into account the diversity of reply structures and identified three distinct types of reply structures, as illustrated in Figure 10. These structures have been carefully considered during the automatic synthesis of dialogues to ensure a realistic and varied representation of conversational dynamics.

► **High-quality and Large-scale.** Through meticulous manual annotation and cross-validation, we ensure the high quality of PanoSent. By employing automated synthesis, we significantly expand the dataset's scale without compromising its quality. This results in a total of 20,000 dialogue instances and 77,303 sextuples. This high-quality large-volume dataset facilitates subsequent research.

Table 8: Data statistics of PanoSent dataset. 'Dia.', 'Utt.' and 'Spk.' refer to dialogue, utterance, and speaker, respectively. 'Hld.', 'Tgt.', 'Asp.', 'Opi.' and 'Rat.' refer to holder, target, aspect, opinion and rationale terms, respectively. 'Sext.' and 'Flip.' refer to sextuple and sentiment flip. 'Imp.' and 'Exp.' refer to implicit and explicit.

		Dialogue			Items					Sextuples		Manner	
		Dia.	Utt.	Spk.	Hld.	Tgt.	Asp.	Opi.	Rat.	Sext.	Flip.	Imp.	Exp.
EN	Total	12,000	83,668	45,753	40,341	14,351	34,848	40,524	38,124	44,348	16,132	11,904	30,388
	Train	9,600	67,060	36,970	32,180	11,590	27,724	32,336	30,524	35,532	12,856	9,931	25,601
	Valid	1,200	8,150	4,368	4,005	1366	3,506	4,065	3,747	4,287	1,616	976	1,255
	Test	1,200	8,458	4,415	4,156	1395	3,618	4,123	3,853	4,529	1,660	997	3,532
ZH	Total	6,000	43,462	25,440	22,662	7,263	17,096	23,039	21,241	23,901	8,051	6,623	17,278
	Train	4,800	34,608	20,478	18,193	5,788	13,753	18,480	17,033	19,273	6,466	5,645	13,628
	Valid	600	4,324	2,398	2,182	701	1617	2,247	2062	2,231	772	480	1,751
	Test	600	4,530	2,564	2,287	774	1726	2,312	2146	2,397	813	498	1,899
SP	Total	2,000	14,657	9,082	7,984	2,797	6,647	8,558	7,937	9,055	3,154	2,218	6,837
	Train	1,600	11,868	7,320	6,406	2,248	5,357	6,838	6,383	7,274	2,542	1,851	5,423
	Valid	200	1335	856	768	246	614	822	714	845	302	175	670
	Test	200	1454	906	810	303	676	898	840	936	310	192	744
	All	20,000	141,787	80,275	70,987	24,411	58,591	72,121	67,302	77,303	27,337	20,745	54,503

Table 9: Extended statistics of PanoSent train set (as extension of Table 2).

		Dialogue			Items					Sextuples		Mannaer	
		Dia.	Utt.	Spk.	Hld.	Tgt.	Asp.	Opi.	Rat.	Sext.	Flip	Imp.	Exp.
EN	Total	9,600	67,060	36,970	32,180	11,590	27,724	32,336	30,524	35,532	12,856	9,931	25,601
	Real	3,168	21,573	11,823	10,925	3,661	8,940	10,274	9,848	11,047	4,160	2,524	8,523
	Synth	6,432	45,487	25,147	21,255	7.929	18,784	22,052	20,676	24,485	8,696	7,407	17,078
ZH	Total	4,800	34,608	20,478	18,193	5,788	13,753	18,480	17,033	19,273	6,466	5,645	13,628
	Real	1,584	12,047	7,471	6,210	2,020	4,721	6,256	5,737	6,544	2,238	1,569	4,975
	Synth	3,216	22,561	13,007	11,983	3,768	9,032	12,224	11,296	12,729	4,228	4,076	8,653
SP	Total	1,600	11,868	7,320	6,406	2,248	5,357	6,838	6,383	7,274	2,542	1,851	5,423
	Real	528	3,812	2,290	2,040	716	1,793	2,234	2,149	2,408	822	494	1,914
	Synth	1,072	8,056	5,030	4,366	1,532	3,564	4,604	4,234	4,866	1,720	1,357	3,509
	All	16,000	113,536	64,768	56,779	19,626	46,834	57,654	53,940	62,079	21,864	17,427	44,652

C.3 Finally Dataset Statistics

We document the composition and distributions of the entire PanoSent dataset in detail. Table 8 provides an overview of the dataset, which comprises 20,000 dialogues in total. The dialogues are distributed across training, validation, and test sets in an 8:1:1 ratio. Table 9 focuses specifically on the training set within the PanoSent dataset, revealing that the ratio of synthetic to real dialogues is approximately 2:1. Each table categorizes the dataset by dialogues, utterances, and speakers, as well as semantic elements like holders, targets, aspects, opinions, and rationales. Additionally, the tables account for sextuples, flips, and the elements' implicit or explicit nature.

D MORE DETAILS OF METHODS

Here, we provide a more detailed introduction to the Chain-of-Sentiment (CoS) reasoning framework and the paraphrase-based verification (PpV) mechanism we proposed. We mainly present more details about the prompts we used.

D.1 Prompts for CoS Reasoning

To more clearly illustrate the workflow of our designed CoS mechanism, we provide a specific dialogue example to demonstrate the reasoning process. The content of the dialogue is as follows:

- *[0] Chris: I find the low-light performance is exceptional, capturing clear and vibrant photos even in dim settings. (reply = -1)* *[IMAGE$_1$](caption: Dusk light in the forest through a mobile phone lens.)*
- *[1] Emma: But the battery life to be quite disappointing. It tends to drain quickly even with minimal usage. (reply = 0)*
- *[2] Sophia: Yes, it is a significant issue, often needing recharging multiple times a day. (reply = 1)*
- *[3] Lucas: And the phone's design blends elegance with practicality. (reply = 0)*
- *[4] Chris: However, I don't see it that way; it seems to follow the same formula as its predecessors. (reply = 3)*
- *[5] Sophia: Have you guys noticed the new model's edge-to-edge display design? It's useful and maximizes screen size without increasing the phone's overall dimensions.(reply = 4)* *[VIDEO$_1$](caption: Showcasing the phone's special edge-to-edge display design.)*

- *[6] Chris: That's a good point. I hadn't really considered that aspect. The edge-to-edge display design is indeed impressive. I might have underestimated its design before. (reply = 5)*

Then, the reasoning process of our CoS goes as follows:

► Step 1: Target-Aspect Identification.

Input Data:
1. Chris: I find the low-light performance is exceptional, capturing clear and vibrant photos even in dim settings. (reply = -1)
2. Emma: But the battery life to be quite disappointing. It tends to drain quickly even with minimal usage. (reply = 0)
3. Sophia: Yes, it is a significant issue, often needing recharging multiple times a day. (reply = 1)
4. Lucas: And the phone's design blends elegance with practicality. (reply = 0)
5. Chris: However, I don't see it that way; it seems to follow the same formula as its predecessors. (reply = 3)
6. Sophia: Have you guys noticed the new model's edge-to-edge display design? It's useful and maximizes screen size without increasing the phone's overall dimensions. (reply = 4)
7. Chris: That's a good point. I hadn't really considered that aspect. The edge-to-edge display design is indeed impressive. I might have underestimated its design before. (reply = 5)

With encoded information of [IMAGE$_1$], [VIDEO$_1$]

Instruction: Based on the multi-party dialogue and its accompanying multimodal data, please identify all possible targets and their specific aspects mentioned in the dialogue. Extract each target and aspect explicitly from the utterance text spans, or infer them implicitly via your understanding of the input data. Ensure each identified target is paired with its aspect(s), forming target-aspect pairs.

Output: Target-aspect pairs: (phone, low-light performance), (phone, battery life), (phone, design)

► Step 2: Holder-Opinion Detection.

Input Data:
1. Chris: I find the low-light performance is exceptional, capturing clear and vibrant photos even in dim settings. (reply = -1)
2. Emma: But the battery life to be quite disappointing. It tends to drain quickly even with minimal usage. (reply = 0)
3. Sophia: Yes, it is a significant issue, often needing recharging multiple times a day. (reply = 1)
4. Lucas: And the phone's design blends elegance with practicality. (reply = 0)
5. Chris: However, I don't see it that way; it seems to follow the same formula as its predecessors. (reply = 3)
6. Sophia: Have you guys noticed the new model's edge-to-edge display design? It's useful and maximizes screen size without increasing the phone's overall dimensions. (reply = 4)
7. Chris: That's a good point. I hadn't really considered that aspect. The edge-to-edge display design is indeed impressive. I might have underestimated its design before. (reply = 5)

Target-aspect pairs: (phone, low-light performance), (phone, battery life), (phone, design)

Instruction: Based on the dialogue and each target-aspect pair identified previously, please identify the holder (the person who expresses an opinion, normally should be a speaker of certain dialogue utterance) and the opinion, both either directly extracted from the text or inferred from our understanding of the input data. Formulate your output into 'holder-target-aspect-opinion' quadruples, ensuring each element is clearly identified.

Output: Target-aspect pairs: (Chris, phone, low-light performance, exceptional), (Emma, phone, battery life, quite disappointing), (Sophia, phone, battery life, a significant issue), (Lucas, phone, design, good), (Chris, phone, design, ordinary), (Sophia, phone, design, useful), (Chris, phone, design, indeed impressive)

► Step 3: Sentiment-Rationale Mining.

Input Data:
1. Chris: I find the low-light performance is exceptional, capturing clear and vibrant photos even in dim settings. (reply = -1)
2. Emma: But the battery life to be quite disappointing. It tends to drain quickly even with minimal usage. (reply = 0)
3. Sophia: Yes, it is a significant issue, often needing recharging multiple times a day. (reply = 1)
4. Lucas: And the phone's design blends elegance with practicality. (reply = 0)
5. Chris: However, I don't see it that way; it seems to follow the same formula as its predecessors. (reply = 3)
6. Sophia: Have you guys noticed the new model's edge-to-edge display design? It's useful and maximizes screen size without increasing the phone's overall dimensions. (reply = 4)
7. Chris: That's a good point. I hadn't really considered that aspect. The edge-to-edge display design is indeed impressive. I might have underestimated its design before. (reply = 5)

Holder-target-aspect-opinion quadruples: (Chris, phone, low-light performance, exceptional), (Emma, phone, battery life, quite disappointing), (Sophia, phone, battery life, a significant issue), (Lucas, phone, design, good), (Chris, phone, design, ordinary), (Sophia, phone, design, useful), (Chris, phone, design, indeed impressive)

Instruction: Based on the dialogue and each holder-target-aspect-opinion quadruple identified previously, please identify the sentiment polarity associated with the opinion and analyze the causal rationale behind it. The sentiment polarity should be classified as 'positive', 'neutral', or 'negative'. The rationale should be extracted explicitly from the text, or inferred implicitly via your understanding of the input data. Formulate your output into 'holder-target-aspect-opinion-sentiment-rationale' sextuples, ensuring sentiment polarity is clearly analyzed and the other five elements is clearly identified.

Output: Holder-target-aspect-opinion-sentiment-rationale sextuples: (Chris, phone, low-light performance, exceptional, positive, capturing clear and vibrant photos even in dim settings), (Emma, phone, battery life, quite disappointing, negative, drain quickly even with minimal usage), (Sophia, phone, battery life, a significant issue, negative, often needing recharging multiple times a day), (Lucas, phone, design, good, positive, blends elegance with practicality), (Chris, phone, design, ordinary, neutral, follow the same formula as its predecessors), (Sophia, phone, design, useful, positive, maximizes screen size without increasing the phone's overall dimensions), (Chris, phone, design, indeed impressive, positive, reconsider the practical innovation of the edge-to-edge display)

► Step 4: Sentiment Flipping Trigger Classification.

Input Data:
1. Chris: I find the low-light performance is exceptional, capturing clear and vibrant photos even in dim settings. (reply = -1)
2. Emma: But the battery life to be quite disappointing. It tends to drain quickly even with minimal usage. (reply = 0)
3. Sophia: Yes, it is a significant issue, often needing recharging multiple times a day. (reply = 1)
4. Lucas: And the phone's design blends elegance with practicality. (reply = 0)

5. Chris: However, I don't see it that way; it seems to follow the same formula as its predecessors. (reply = 3)
6. Sophia: Have you guys noticed the new model's edge-to-edge display design? It's useful and maximizes screen size without increasing the phone's overall dimensions. (reply = 4)
7. Chris: That's a good point. I hadn't really considered that aspect. The edge-to-edge display design is indeed impressive. I might have underestimated its design before. (reply = 5)

Holder-target-aspect-opinion-sentiment-rationale sextuples: (Chris, phone, low-light performance, exceptional, positive, capturing clear and vibrant photos even in dim settings), (Emma, phone, battery life, quite disappointing, negative, drain quickly even with minimal usage), (Sophia, phone, battery life, a significant issue, negative, often needing recharging multiple times a day), (Lucas, phone, design, good, positive, blends elegance with practicality), (Chris, phone, design, ordinary, neutral, follow the same formula as its predecessors), (Sophia, phone, design, useful, positive, maximizes screen size without increasing the phone's overall dimensions), (Chris, phone, design, indeed impressive, positive, reconsider the practical innovation of the edge-to-edge display)

Instruction: Based on the dialogue and each holder-target-aspect-opinion-sentiment-rationale' sextuple, please identify instances where a sentiment flip occurs for the same holder regarding the specific target-aspect pair. Determine the trigger type for these flips from the predefined categories: *introduction of new information, logical argumentation, participant feedback and interaction, personal experience and self-reflection*. Formulate your output to include the holder, target, aspect, initial sentiment, flipped sentiment, and the trigger type, or state "None" if no flips are identified.

Output: (Chris, phone, design, neutral, positive, introduction of new information)

D.2 Prompts for Paraphrase-based Verification

In our paraphrase-based verification mechanism, the transformation of k-tuples into natural language expressions is carefully designed for each specific k-tuple. This ensures that the expressions accurately reflect the intended sentiment analysis's meaning and context. Each step in the verification process can yield multiple outcomes—such as pairs, quadruples, or sextuples—depending on the specific demands of the analysis task. For example, in the initial step, if k target-aspect pairs are identified, they are represented as $(t_1, a_1), (t_2, a_2), ..., (t_k, a_k)$. The verification templates that follow are structured to assess the consistency of these outcomes with the dialogue content, thereby validating the precision of our analysis.

► Step 1: Verification of Target-Aspect Identification

Input Data: D, $\{(t_i, a_i)\}$
Instruction: In this dialogue, participants discussed various targets and their corresponding aspects, including a_1 of t_1, a_2 of t_2, etc. Please based on the dialogue, verify whether these descriptions are consistent with the dialogue content and provide '1' for 'yes' or '0' for 'no' judgment.

Expected Output: 1 (if yes) or 0 (if no)

► Step 2: Verification of Holder-Opinion

Input Data: D, $\{(h_j, t_i, a_i, o_j)\}$
Instruction:In this dialogue, different participants expressed their opinions towards various aspects of targets, including the opinion of h_1 on a_1 of t_1 is o_1, and the opinion of h_2 on a_2 of t_2 is o_2, etc. Please based on the dialogue, verify whether these descriptions are consistent with the dialogue content and provide '1' for 'yes' or '0' for 'no' judgment.

Expected Output: 1 (if yes) or 0 (if no)

► Step 3: Verification of Sentiment-Rationale Mining

Input Data: D, $\{(h_j, t_i, a_i, o_j, s_k, r_l)\}$
Instruction: In this dialogue, the analysis has identified sentiments and rationales behind opinions, including h_1's opinion o_1 on a_1 of t_1 carries a sentiment s_1 with rationale r_1, etc. Please based on the dialogue, verify whether these descriptions are consistent with the dialogue content and provide '1' for 'yes' or '0' for 'no' judgment.

Expected Output: 1 (if yes) or 0 (if no)

► Step 4: Verification of Sentiment Flipping Trigger Classification

Input Data: D, $\{(h_j, t_i, a_i, o_j, s_k, r_l)\}$
Instruction: In this dialogue, instances of sentiment flipping and their triggers have been identified, including h_1's sentiment towards a_1 of t_1 initially was ζ_1 and later flipped to ϕ_1 due to trigger τ_1, etc. Please based on the dialogue and your commonsense knowledge, verify whether these descriptions accurately capture the emotional dynamics and their triggers in the dialogue and provide '1' for 'yes' or '0' for 'no' judgment.

Expected Output: 1 (if yes) or 0 (if no)

Upon receiving outcomes from the verification prompted by the MLLM, the next steps are as follows:

In case of inconsistency: If verification results show the expression is inconsistent with the dialogue content, we will instruct the LLM to regenerate and reverify the k-tuples.

In case of consistency: If the LLM confirms the expression is consistent with the dialogue content, it indicates that the current step's reasoning and transformation results are trustworthy. We then proceed with the next steps of analysis and verification based on this confirmed information.

This procedure ensures the analysis moves forward in an orderly manner. If inconsistencies arise, they are addressed by revisiting the analysis steps; once results are confirmed to be consistent, the analysis proceeds, leveraging these verified outcomes for subsequent steps.

E EXTENSIONS OF SETTINGS AND IMPLEMENTATIONS

In this section, we continue to provide more descriptions about the implementation details of our system and experiments.

E.1 System Training Details

E.1.1 Training Step 1: Multimodal Understanding Stage

► **Training Data**: The training data comprises 'text+X' pairs, where 'X' represents various forms of multimodal inputs including images, audios or videos. This diverse dataset structure is crucial for enabling LLM to learn from and interpret a wide range of multimodal information, thereby enhancing its ability to process and understand complex multimodal scenarios. Specifically, we employ well-established datasets such as LLaVA[39], miniGPT-4[83], and VideoChat[34], which have been designed for multimodal language model instruction tuning. These datasets not only provide a rich source of 'Text+X' pairs but also align with our objective to improve LLM's proficiency in generating textual responses from multimodal inputs, covering a broad spectrum of real-world scenarios and enhancing the model's understanding of multimodal content.

► **Training Objective:** The primary objective is to train the LLM to accurately interpret and generate textual descriptions for multimodal inputs, fostering a comprehensive understanding of both textual and non-textual content.

► **Loss Function**: We employ the Negative Log-Likelihood (NLL) Loss.

$$L_{NLL} = -\sum_{t=1}^{T} \log(p_{t,c_t}) \tag{5}$$

where T is the length of the text sequence, c_t represents the correct class (word) at time step t, and p_{t,c_t} is the probability assigned by the model to the correct word at time step t. This loss function aims to maximize the probability of the correct word sequence, thereby improving the model's ability to generate accurate and coherent textual descriptions from multimodal inputs.

E.1.2 Training Step 2: CoS Reasoning Process

► **Training Data**: Utilizes the PanoSent training set, segmented into instructions according to the Chain-of-Sentiment (CoS) reasoning framework for each of the task's four progressive steps.

► **Training Objective**: To enable the model to sequentially execute the CoS reasoning steps, facilitating an incremental understanding and processing of the given tasks' complexities.

► **Loss Function:** A composite loss function is applied, catering to the multi-task nature of the problem.

$$L = \sum_{i=1}^{n} \lambda_i L_{taski} \tag{6}$$

where L_{taski} denotes the loss for the i^{th} task, and λ_i represents the weight assigned to each task, signifying its importance. This function allows for simultaneous optimization across multiple reasoning steps, aligning with the objective of facilitating an incremental understanding and processing of the tasks' complexities. In our experiment, we treat each step equally important, we set $\lambda_i = 1$ for all tasks.

E.1.3 Training Step 3: Paraphrase-based Verification

► **Training Data:** Comprises paraphrase pairs that exhibit either an entailment or contradiction relation to the given context, aimed at verifying the accuracy of results from previous reasoning steps.

Table 10: Detail of the hyper-parameter setting.

Param	Value
Model size	11.3B
Tensor type	F32
Learning rate	1e-5
LoRA rank	8
LoRA alpha	32
Weight decay	1e-6
Batch size	8*8 (dialogues)
Epoch size	250
GPU	8*A100

► **Training Objective:** To train the model to distinguish between entailment and contradiction in the context of the provided paraphrases, ensuring the integrity and reliability of each reasoning step.

► **Loss Function:** For the task of classifying paraphrase pairs as entailment or contradiction, we use the Binary Cross-Entropy Loss function.

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{7}$$

where N is the number of samples, y_i indicates the true label (1 for entailment, 0 for contradiction), and p_i is the predicted probability of the i^{th} sample being an entailment. This loss function aims to optimize the model's ability to accurately classify the paraphrase pairs into the correct categories, enhancing the accuracy of the reasoning process.

E.2 More Detail of Model Configurations

In our experimentation, we use Flan-T5-XXL, an encoder-decoder language model with a size of 11.3 billion parameters, publicly available through Google on the HuggingFace platform. Hyperparameters are listed in Table 10. Our experimental settings include a learning rate of 1e-5, LoRA rank of 8, and LoRA alpha of 32. The tensor type used is F32, and we introduce a weight decay of 1e-6 to regularize the training. We processe the data in batches of 64 dialogues (8*8) over 250 epochs, using adafactor as the optimizer.

For the computational resources, we utilize an array of eight NVIDIA A100 GPUs. To account for variability and ensure robustness of our findings, we include stochastic elements by employing five different random seeds in our training. We fine-tune the model by employing the LoRA technique to modify a small subset of model parameters efficiently. The trained parameters specific to LoRA within Flan-T5-XXL is around 71 million, accounting for roughly 0.6% of the total parameter count.

E.3 Baseline Specification

Given the novel nature of our task and the lack of directly comparable prior research and methods, we thus establish several baselines through our own implementations of existing methods. First, we retrofit non-LLM-based systems, including UGF and DiaASQ, which are initially built for related ABSA tasks, e.g., sentiment triplet and quadruple extraction. The process involves adapting smaller-sized language models, specifically Multilingual BERT (Base) and mT5

(XXL), to perform our designated task. Since these models do not inherently support multimodal input processing, we employ a method similar to our model, using imagebind to encode multimodal information. Subsequently, we extend our comparisons to include LLM-based systems, which support text, audio, image, and video, such as Unified-IO 2 and NExT-GPT. We only use NExT-GPT for comparison of Sentiment Flipping Analysis task. For an equitable comparison, all systems are fine-tuned on the PanoSent training set. A comprehensive description of these baseline systems is presented to facilitate understanding and reproducibility.

E.3.1 Non-LLM-based Baseline Implementations

► **DiaASQ**: The DiaASQ[6] framework, which focuses on extracting quadruples from dialogues, utilizes a neural model that leverages dialogue-specific and discourse feature representations for effective end-to-end quadruple prediction. To adapt this framework for our sextuple extraction tasks (including ‘holder’ and ‘rationale’), we need to make some modifications to the architecture. First, the labeling scheme must be expanded to encompass ‘holder’ and ‘rationale’ elements, requiring adjustments to entity boundary labels and entity pair labels. Secondly, the dialogue-specific multi-view interaction layer should be adjusted by adding attention masks or features that can effectively distinguish and relate these new elements within the conversational context. Lastly, the decoding process, originally designed for quadruples, requires enhancements to recognize and extract sextuples.

► **UGF**: The Unified Generative Framework[7] transforms all ABSA subtasks into a unified generative task by treating each subtask target as a sequence composed of pointer indexes and sentiment class indexes. This approach leverages the BART sequence-to-sequence model to solve ABSA subtasks in an end-to-end manner. To adapt this framework for our sextuple extraction task, we propose architectural modifications to include new elements within the generative formulation. Specifically, this would involve extending the sequence representation to incorporate new indexes or tokens. Additionally, adjustments would be made to the model to enable the generation of these expanded sequences. These modifications would allow the framework to capture the extended relationships inherent in the sextuple extraction task, maintaining the unified and end-to-end nature of the original architecture.

E.3.2 LLM-based Baseline Implementations

► **Unified-IO 2**: Unified-IO 2[8] is an autoregressive multimodal model capable of understanding and generating content across images, text, audio, and action by encoding these inputs and outputs into a shared semantic space using a unified encoder-decoder transformer model. To adapt Unified-IO 2 for our sextuple extraction task, we leverage its inherent multimodal encoding capabilities. Specifically, we use conversational data and associated multimodal information as inputs. We use the same prompt of our method to guide the model to focus on extracting the sextuple elements.

► **NExT-GPT**: NExT-GPT[9] introduces an any-to-any MLLM system that seamlessly handles inputs and generates outputs across a variety of modalities including text, images, videos, and audio. The architecture is structured around connecting a LLM with multimodal adaptors and diffusion decoders. This design enables NExT-GPT to perceive and generate content in arbitrary combinations of modalities. To adapt NExT-GPT for our sextuple extraction task, we utilize its architecture to encode conversational data and associated multimodal information. By feeding the model conversational data along with relevant multimodal inputs, we can prompt NExT-GPT to perform inference and extract sextuples. For sentiment flipping analysis task, since the Non-LLM-based method mentioned above is specifically modeled for extraction tasks, we only use NExT-GPT in the LLM-based method for comparison. The test method is the same as sextuple extraction task, performing inference through prompts.

F EVALUATION SPECIFICATIONS

Here, we provide a detailed introduction on how we conduct the evaluation for the two subtasks.

F.1 Subtask-I Evaluation

For Subtask I, focusing on the extraction of fine-grained sentiment sextuples, our evaluation methodology is designed to rigorously assess the performance across various aspects of the task. We provide detailed specifications for element-wise, pair-wise, and overall sextuple evaluations.

F.1.1 Element-wise Evaluations

► **Explicit Elements.** For elements explicitly mentioned in the text, we apply the exact match metric for evaluation. Under this metric, a correct prediction must precisely match the term as annotated in the gold standard. Exact Precision (EP) is calculated as the proportion of correctly predicted terms among all predicted terms, while Exact Recall (ER) is the proportion of correctly predicted terms among all gold terms.

$$EP = \frac{\#\text{correct terms}}{\#\text{predicted terms}} \tag{8}$$

$$ER = \frac{\#\text{correct terms}}{\#\text{gold terms}} \tag{9}$$

$$\text{Exact Match F1} = 2 \cdot \frac{EP \cdot ER}{EP + ER} \tag{10}$$

Here, ‘#’ denotes the amount, and ‘correct terms’ refer to the predicted terms that exactly match the gold terms.

► **Implicit Elements.** For implicit elements not explicitly mentioned in the text, we utilize the binary match metric, which is a relaxation of the above exact one. For implicit elements not explicitly mentioned in the text, we utilize the binary match metric, which is a relaxation of the exact match metric. We evaluate if the predicted element is semantically identical to the gold term, as assessed by GPT-4, assigning a binary outcome (1 if yes, otherwise 0). When constructing such queries, it is crucial to include sufficient contextual information from the dialogue. This is because the meaning of terms can vary with context, and relying solely on the terms themselves may not accurately reflect their significance

[6] https://github.com/unikcc/DiaASQ
[7] https://github.com/yhcc/BARTABSA
[8] https://github.com/allenai/unified-io-2
[9] https://next-gpt.github.io/

in a specific dialogue. Therefore, it is essential that the prompts provided to GPT-4 contain complete dialogue content to enable accurate semantic evaluation. Our standard instruction template for GPT-4 is: "Given the context of the dialogue, do '[predicted term]' and '[gold standard term]' have similar meanings?"

$$BP = \frac{\#\text{semantically identical terms}}{\#\text{predicted terms}} \tag{11}$$

$$BR = \frac{\#\text{semantically identical terms}}{\#\text{gold terms}} \tag{12}$$

$$\text{Binary Match F1} = 2 \cdot \frac{BP \cdot BR}{BP + BR} \tag{13}$$

► **Element of Explicit Rationale.** For evaluating the explicit rationale element, we use the proportional match metric, which measures the proportional overlap between the predicted and gold standard terms. Proportional overlap assigns a score to represent the proportion of the overlapped region, rather than a binary value, 0 or 1. Proportional precision (PP) measures the proportion of the overlap between a predicted term and an overlapping gold term. Proportional recall (PR) measures the proportion of the overlap between a gold term and an overlapping predicted term.

$$PP = \frac{\#\text{correct terms}|\text{proportional overlap}}{\#\text{predicted terms}} \tag{14}$$

$$PR = \frac{\#\text{correct terms}|\text{proportional overlap}}{\#\text{gold terms}} \tag{15}$$

$$\text{Proportional Match F1} = 2 \cdot \frac{PP \cdot PR}{PP + PR} \tag{16}$$

► **F1 Score for Each Element.** The F1 score for each element is the average of the Exact F1 and the Relevant F1 score under that category, which could be either Binary F1 for implicit elements or Proportional F1 for explicit rationale, depending on the nature of the element.

► **Sentiment Classification.** The macro F1 Score is calculated as the average of F1 Scores for all sentiment classes, offering a balanced measure of model performance across different sentiment orientations. For each sentiment class c, we define:

$$CP_c = \frac{\#\text{correct predictions for class } c}{\#\text{predictions of class } c} \tag{17}$$

$$CR_c = \frac{\#\text{correct predictions for class } c}{\#\text{gold instances of class } c} \tag{18}$$

$$\text{Class F1}_c = 2 \cdot \frac{CP_c \times CR_c}{CP_c + CR_c} \tag{19}$$

$$\text{Macro F1} = \frac{\text{F1}_{\text{positive}} + \text{F1}_{\text{negative}} + \text{F1}_{\text{neutral}}}{3} \tag{20}$$

F.1.2 Pair-wise Evaluations

For a pair, the prediction must correctly identify both spans, and adhere to the evaluation standards for implicit elements and rationale.

► **Pair-wise F1 Score.** This metric evaluates the precision and recall of correctly identified pairs within the sextuples.

$$PP = \frac{\#\text{correct pairs}}{\#\text{predicted pairs}} \tag{21}$$

$$PR = \frac{\#\text{correct pairs}}{\#\text{gold pairs}} \tag{22}$$

$$\text{Pair-wise F1} = 2 \cdot \frac{PP \cdot PR}{PP + PR} \tag{23}$$

F.1.3 Sextuple Evaluations

For sextuple extraction, the prediction must accurately match all six elements, samely with consideration for the accuracy of implicit elements and rationale.

► **Micro F1 Score.** This metric evaluates the overall precision(OP) and overall recall(OR) for sextuple extraction.

$$OP = \frac{\#\text{correct sextuples}}{\#\text{predicted sextuples}} \tag{24}$$

$$OR = \frac{\#\text{correct sextuples}}{\#\text{gold sextuples}} \tag{25}$$

$$\text{Micro F1} = 2 \cdot \frac{OP \cdot OR}{OP + OR} \tag{26}$$

► **Identification F1 Score.** This metric focuses on the identification precision(IP) and identification recall(IR) of sextuples, excluding sentiment polarity.

$$IP = \frac{\#\text{correctly identified sextuples without sentiment}}{\#\text{predicted sextuples}} \tag{27}$$

$$IR = \frac{\#\text{correctly identified sextuples without sentiment}}{\#\text{gold sextuples}} \tag{28}$$

$$\text{Identification F1} = 2 \cdot \frac{IP \cdot IR}{IP + IR} \tag{29}$$

F.2 Subtask-II Evaluation

In Subtask-II, the evaluation of model performance in identifying sentiment flips and their triggers adopts specific measures tailored to the complexity of each task component. For assessing the identification of initial and flipped sentiments as well as their combined evaluation with triggers, the exact match F1 score is employed to account for the precision in capturing the interconnected aspects of sentiment transitions. Conversely, for the classification task of identifying triggers alone, the Macro F1 score is utilized to ensure a balanced evaluation across all trigger categories, reflecting equal importance to the accurate identification of each trigger type.

F.2.1 Flip Evaluations

To assess the model's ability to correctly identify both the initial sentiment and the flipped sentiment, we use the exact match F1 score. This measure accurately reflects the model's capability in detecting precise changes in sentiment:

$$\text{Exact Match F1} = 2 \cdot \frac{\text{Precision}_{\text{Flip}} \times \text{Recall}_{\text{Flip}}}{\text{Precision}_{\text{Flip}} + \text{Recall}_{\text{Flip}}} \tag{30}$$

F.2.2 Trigger Evaluations

We evaluate the identification of flipping triggers using the Macro F1 score, which accommodates the diversity of trigger categories within the dataset. This metric ensures that all categories are assessed with equal importance, providing a balanced measure of performance across varied types of triggers.

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} 2 \cdot \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{31}$$

Table 11: Rationale extraction evaluation results on 200 EN test samples.

	Rationale Extraction
Human Evaluation	68.31
Proportional Match F1	46.49
Exact Match F1	21.38

where N is the number of trigger categories, and Precision_i and Recall_i are the precision and recall for the i-th trigger category, respectively.

F.2.3 Overall Flip-Trig Evaluations

Finally, the model's overall performance in simultaneously identifying both the correct flipped sentiment and the correct trigger is assessed using the exact match f1 score, providing a comprehensive evaluation of the model's nuanced understanding of sentiment dynamics and their triggers:

$$\text{Exact Match F1} = 2 \cdot \frac{\text{Precision}_{\text{Flip-Trig}} \times \text{Recall}_{\text{Flip-Trig}}}{\text{Precision}_{\text{Flip-Trig}} + \text{Recall}_{\text{Flip-Trig}}} \tag{32}$$

G MORE EXPERIMENTS AND ANALYSES

We further present additional experimental results and analyses.

G.1 Evaluation on Rationale

This experiment aims to compare the applicability of the proportional match F1 versus exact match F1 evaluation metrics in the task of rationale extraction. We focus on empirically validating the performance of these two evaluation methods across 200 data entries, using human judgment as a benchmark to assess their effectiveness.

First, we calculate the exact match F1 and proportional match F1 scores for rationale extraction on the selected dataset. Next, we conduct a manual review of these 200 data entries, providing a binary match F1 score to assess whether the predicted rationale is semantically identical to the gold rationale. Lastly, these automatically computed scores are directly compared with the results of the manual review.

As shown in Table 11, the results demonstrate that the proportional match F1 scores are significantly more consistent with manual evaluations than the exact match F1 scores. This finding supports the effectiveness of the proportional match F1 evaluation metric in situations of partial text match for rationale. It indicates that proportional match F1 better captures and evaluates text segments that support specific sentiment judgments, compared to exact match F1. This discrepancy highlights the superior flexibility and alignment of proportional match F1 with human assessment practices in sentiment analysis tasks, especially those involving rationale extraction.

G.2 Extended Explorations of Impact of Using Different Backbone LLMs

In order to compare the performance of different LLM backbones on our two subtasks, we conduct a controlled experiment where we maintain consistent methodologies and architectures across two settings—Sentica and Sentica (+Cos+PpV)—while varying only the LLM backbone used for task reasoning. For a fair comparison, each model is evaluated using the same set of parameters and input data (only English dataset), ensuring that any performance differences could be attributed to the backbone itself, rather than external variables.

Table 12: Comparason of LLM Backbones on EN Dataset.

	PLM	**Method**	**Result**	
			Sextuple	Flip-Trig
M1	mT5-XXL	Sentica	13.29	37.66
M2	mT5-XXL	Sentica(+Cos+PpV)	16.09	40.72
M3	Vicuna 7B	Sentica	23.26	63.49
M4	Vicuna 7B	Sentica(+Cos+PpV)	28.70	68.33
M5	Llama2	Sentica	24.16	65.09
M6	Llama2	Sentica(+Cos+PpV)	29.97	68.83
M7	Flan-T5-XXL	Sentica	26.06	66.71
M8	Flan-T5-XXL	Sentica(+Cos+PpV)	**32.18**	**71.39**

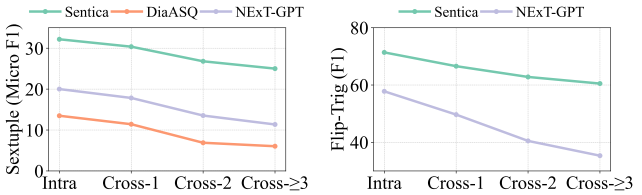


Figure 11: Performance of two subtasks on different cross-utterance levels.

As presented in Table 12, the results indicate that the Flan-T5-XXL backbone outperforms others in both subtasks. This superior performance is evident in the consistently higher scores achieved in the subtasks, confirming the efficacy of Flan-T5-XXL as a backbone for the Sentica framework.

G.3 Cross-utterance Sextuple Extraction and Sentiment Flip Trigger Identification.

In assessing the impact of cross-utterance dialogue dynamics on emotion analysis tasks, our experimental results demonstrate a consistent trend across both subtasks evaluated, shown in Figure 11. Cross-utterance interaction presents a discernible challenge that invariably leads to a degradation in performance. However, our Sentica mitigates this effect more robustly than comparative methodologies. This is evidenced by a relatively smaller decline in F1 scores, particularly in scenarios with increased cross-utterance complexity. Subtask I, which entails the extraction of sentiment sextuples, inherently requires a deeper contextual comprehension, making it more vulnerable to cross-utterance disturbances than Subtask II's focus on sentiment trigger identification and classification. When comparing LLM-based methods (Sentica and NExT-GPT) with non-LLM-based methods (DiaASQ), the former exhibits superior capability in contending with cross-utterance intricacies. Specifically, our model outstrips DiaASQ significantly under cross-utterance conditions, maintaining a higher performance baseline. For Subtask II, a similar pattern prevails with our method outperforming NExT-GPT. This underlines our model's robustness, not only in intra-utterance contexts but also when navigating the complexities introduced by cross-utterance dialogue sequences.

Table 13: Performance comparison of LLM backbones on EN dataset.

	PLM	Method	Result	
			Sextuple	Flip-Trig
M1	mT5-XXL	/	15.96	39.21
M2	mT5-XXL	+Cos+PpV	18.47	43.04
M3	Vicuna 7B	/	28.59	68.38
M4	Vicuna 7B	+Cos+PpV	28.70	68.33
M5	Llama2	/	28.11	67.64
M6	Llama2	+Cos+PpV	34.28	72.07
M7	Flan-T5 XXL	/	29.67	69.06
M8	Flan-T5 XXL	+Cos+PpV	**36.82**	**73.44**

Table 14: Ablation study results of instruction tuning on Flan-T5-XXL.

Model Configuration	Sextuple	Flip-Trig Trip
Complete Model (All Stages)	**32.18**	**71.39**
w/o Stage 1 (Multimodal Learning)	18.41$_{(\downarrow 13.77)}$	68.05$_{(\downarrow 3.34)}$
w/o Stage 2 (CoS Reasoning)	27.37$_{(\downarrow 4.81)}$	66.81$_{(\downarrow 4.58)}$
w/o Stage 3 (PpV Verification)	29.73$_{(\downarrow 2.45)}$	69.27$_{(\downarrow 2.12)}$

G.4 Comparisons with LLMs under Text-only Setting

To assess the performance of different LLM backbones when processing text-only data, we encode dialogue texts under two subtasks, selecting instances that contain only text modality for training and testing. This experiment is designed to exclude the influence of multimodal information and purely compare the text processing capabilities of each backbone. As shown in Table 13, in a text-only environment, the Flan-T5-XXL backbone demonstrated the best performance on two subtasks, showing Flan-T5-XXL's exceptional ability in pure text understanding and reasoning.

G.5 Influence of Different Instruction Tuning Strategies

In this experiment, we conduct an ablation study focusing on the three stages of Instruction Tuning to assess their individual contributions to the performance of our Sentica model under the Flan-T5-XXL backbone. The experiment is designed to isolate and evaluate the impact of each training phase—understanding multimodal representations, executing the CoS reasoning process, and mastering the PpV verification—by sequentially removing training stages and observing the resultant effect on model performance. By individually removing these training phases, we clearly demonstrate the specific contribution of each phase to model performance. The results as shown in Table 14 indicate that each independent training phase significantly enhances the model's understanding and reasoning abilities, particularly when these phases are utilized in conjunction. Sentica's performance in our tasks is notably improved when integrating all three stages.

Table 15: Comparison of joint and separate execution for subtask-II on EN data.

	Sextuple	Flip-Trig Trip
Joint	53.81	**71.39**
Separate	53.81	64.06

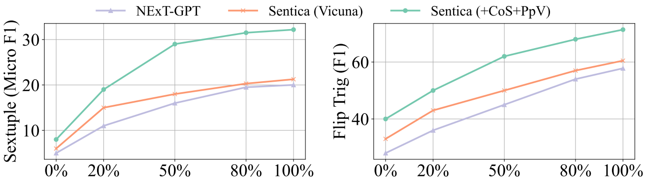


Figure 12: Performance on different training data volume.

G.6 Impact of Joint VS. Separate Subtask Execution.

The experiment aims to determine the effects of jointly performing Panoptic Sentiment Sextuple Extraction (subtask-I) and Sentiment Flipping Analysis (subtask-II) as opposed to processing them separately. In our CoS framework, we adopt a joint (cascade) approach. Comparative analysis of the results reveals that Subtask-II, when informed by the sentiment sextuples inferred from Subtask-I, demonstrates increased accuracy in identifying the Flip-Tri pair within dialogues. This improvement is significantly reflected in the increase of the Flip-Tri pair metric from 64.06 to 71.39, as shown in Table 15. The findings confirm that the sentiment sextuples from Subtask-I serve as critical reference information for Subtask-II, significantly enhancing the precision of sentiment flip identification and analysis, thereby highlighting the necessity and efficacy of an integrated approach to complex sentiment analysis tasks.

G.7 Influence of Training with Different Data Amount

In this study, we explore the effects of varying the volume of supervised training data on a LLM across five different data levels: 0%, 20%, 50%, 80%, and 100% of the training set. This investigation aims to pinpoint how different quantities of training data influence the model's performance in a supervised setting, with a particular focus on understanding the incremental benefits of additional data. We systematically increase the proportion of the dataset used for training, allowing for a direct comparison of the model's performance across these varying levels of data availability. The result, as shown in Figure 12, shows a consistent improvement in the model's effectiveness as the amount of supervised training data increases. Notably, the increase from 0% to 20% of the training data yields the most significant performance boost, demonstrating that early additions of supervised data substantially enhance the model's capabilities.

G.8 Few-shot Learning Experiments

The experiment is designed to compare the efficacy of our model against GPT-4 in few-shot learning scenarios without prior task-specific training. In conducting this comparison, few-shot instances

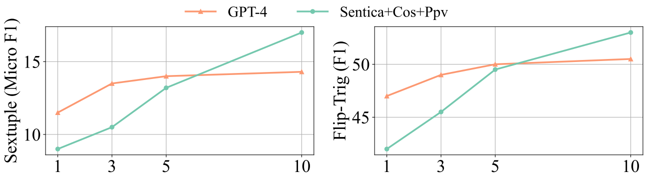
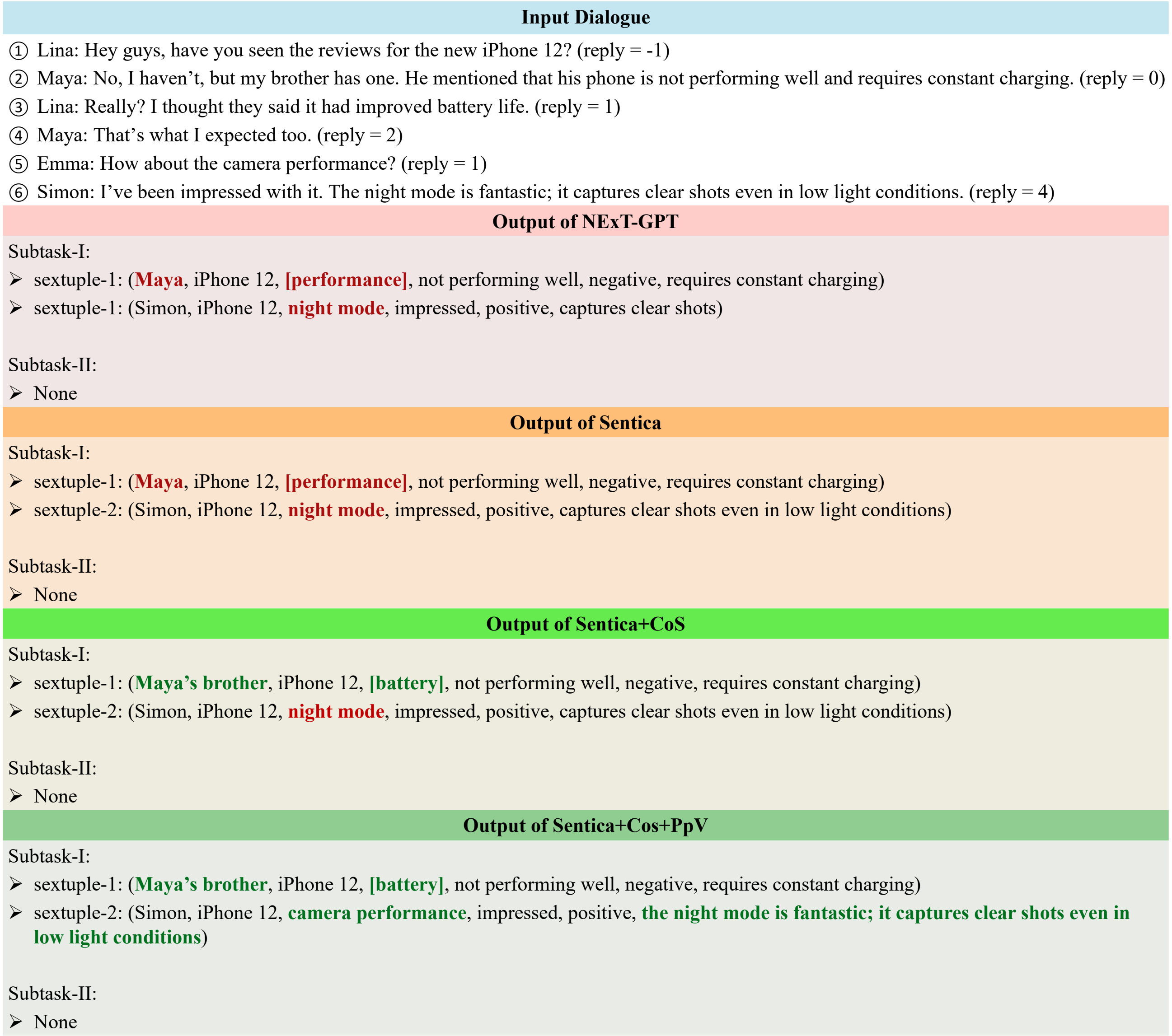


Figure 13: Performance comparison of our model and GPT-4 across different few-shot learning.

of 1, 3, 5, and 10 are chosen to observe how both models adapt and learn from an increasing number of examples. The results in Figure 13 shows that both models performing modestly with just 1 and 3 shots, due to the limited amount of information available. GPT-4 performs significantly better in scenarios with minimal examples. However, as the shot count is elevated to 5 and then to 10, our model demonstrate a notable uptick in performance, indicative of its enhanced capability to assimilate and apply the task's salient features and patterns effectively.

G.9 Case Study

We present several examples to highlight the performance differences between our model and others. As shown in Figures 14, 15, and 16, our model exhibits a deeper understanding of complex dialogue contexts, skillfully capturing subtle nuances and inferring implicit intentions. Its superior ability to handle multimodal information results in a more accurate interpretation across various modalities. Additionally, our model excels at uncovering implicit elements within dialogues. These strengths collectively allow for more comprehensive extraction of sextuple information and also aid in a more accurate analysis of sentiment flips within conversations.



Input Dialogue

① Lina: Hey guys, have you seen the reviews for the new iPhone 12? (reply = -1)
② Maya: No, I haven’t, but my brother has one. He mentioned that his phone is not performing well and requires constant charging. (reply = 0)
③ Lina: Really? I thought they said it had improved battery life. (reply = 1)
④ Maya: That’s what I expected too. (reply = 2)
⑤ Emma: How about the camera performance? (reply = 1)
⑥ Simon: I’ve been impressed with it. The night mode is fantastic; it captures clear shots even in low light conditions. (reply = 4)

Output of NExT-GPT

Subtask-I:
➢ sextuple-1: (**Maya**, iPhone 12, **[performance]**, not performing well, negative, requires constant charging)
➢ sextuple-1: (Simon, iPhone 12, **night mode**, impressed, positive, captures clear shots)

Subtask-II:
➢ None

Output of Sentica

Subtask-I:
➢ sextuple-1: (**Maya**, iPhone 12, **[performance]**, not performing well, negative, requires constant charging)
➢ sextuple-2: (Simon, iPhone 12, **night mode**, impressed, positive, captures clear shots even in low light conditions)

Subtask-II:
➢ None

Output of Sentica+CoS

Subtask-I:
➢ sextuple-1: (**Maya’s brother**, iPhone 12, **[battery]**, not performing well, negative, requires constant charging)
➢ sextuple-2: (Simon, iPhone 12, **night mode**, impressed, positive, captures clear shots even in low light conditions)

Subtask-II:
➢ None

Output of Sentica+Cos+PpV

Subtask-I:
➢ sextuple-1: (**Maya’s brother**, iPhone 12, **[battery]**, not performing well, negative, requires constant charging)
➢ sextuple-2: (Simon, iPhone 12, **camera performance**, impressed, positive, **the night mode is fantastic; it captures clear shots even in low light conditions**)

Subtask-II:
➢ None

Figure 14: A conversation on domain of electronic products. Different colors represent two types of answers. The first type in red indicates the wrong one, yet the green is the correct answer. [*] indicates the implicit information in the text.

Input Dialogue

① Samantha23: I think the screenplay of the new action series is incredibly engaging. The unexpected plot twists consistently deliver high-stakes drama. (reply = -1)

🖼 [Dramatic poster showcasing main characters and a tagline about its thrilling plot twists.]

② TechSavvy: However, I’m unsatisfied with the character development; some are richly portrayed while others barely scratch the surface. (reply = 0)

③ StoryWeaver: Yet, the storyline itself is exceptionally intricate, marked by its layered narratives and compelling plot twists that keep the narrative dynamic. (reply = 1)

④ SuspenseSee: I also find the pacing of the series refreshingly slow, allowing for detailed character development and a more profound build-up of suspense. (reply = 2)

⑤ SuspenseSee: However, upon further viewing, I've come to see the slow pacing as a drawback. (reply = 3)

🎬 [Some slow scenes in series that impact the tension and pace of the action.]

Output of NExT-GPT

Subtask-I:

➢ sextuple-1: (Samantha23, action series, screenplay, incredibly engaging, positive, the unexpected plot twists consistently deliver high-stakes drama)

➢ sextuple-2: (TechSavvy, action series, character development, unsatisfied, negative, some are richly portrayed while others barely scratch the surface)

➢ sextuple-3: (StoryWeaver, action series, storyline, exceptionally intricate, positive, marked by its layered narratives and compelling plot twists)

➢ sextuple-4: (SuspenseSee, action series, pacing, refreshingly slow, positive, allowing for detailed character development)

➢ sextuple-5: (SuspenseSee, action series, **slowing pacing**, a drawback, negative, **upon further viewing**)

Subtask-II:

➢ sentiment-flip: **None**

Output of Sentica

Subtask-I:

➢ sextuple-1: (Samantha23, action series, screenplay, incredibly engaging, positive, the unexpected plot twists consistently deliver high-stakes drama)

➢ sextuple-2: (TechSavvy, action series, character development, unsatisfied, negative, some are richly portrayed while others barely scratch the surface)

➢ sextuple-3: (StoryWeaver, action series, storyline, exceptionally intricate, positive, marked by its layered narratives and compelling plot twists that keep the narrative dynamic)

➢ sextuple-4: (SuspenseSee, action series, pacing, refreshingly slow, positive, allowing for detailed character development and a more profound build-up of suspense)

➢ sextuple-5: (SuspenseSee, action series, **slowing pacing**, a drawback, negative, **upon further viewing**)

Subtask-II:

➢ sentiment-flip: **None**

Output of Sentica+CoS

Subtask-I:

➢ sextuple-1: (Samantha23, action series, screenplay, incredibly engaging, positive, the unexpected plot twists consistently deliver high-stakes drama)

➢ sextuple-2: (TechSavvy, action series, character development, unsatisfied, negative, some are richly portrayed while others barely scratch the surface)

➢ sextuple-3: (StoryWeaver, action series, storyline, exceptionally intricate, positive, marked by its layered narratives and compelling plot twists that keep the narrative dynamic)

➢ sextuple-4: (SuspenseSee, action series, pacing, refreshingly slow, positive, allowing for detailed character development and a more profound build-up of suspense)

➢ sextuple-5: (SuspenseSee, action series, **pacing**, a drawback, negative, **[impact the tension and pace of the action]**)

Subtask-II:

➢ sentiment-flip: (SuspenseSee, action series, pacing, positive, negative, **participant feedback and interaction**)

Output of Sentica+Cos+PpV

Subtask-I:

➢ sextuple-1: (Samantha23, action series, screenplay, incredibly engaging, positive, the unexpected plot twists consistently deliver high-stakes drama)

➢ sextuple-2: (TechSavvy, action series, character development, unsatisfied, negative, some are richly portrayed while others barely scratch the surface)

➢ sextuple-3: (StoryWeaver, action series, storyline, exceptionally intricate, positive, marked by its layered narratives and compelling plot twists that keep the narrative dynamic)

➢ sextuple-4: (SuspenseSee, action series, pacing, refreshingly slow, positive, allowing for detailed character development and a more profound build-up of suspense)

➢ sextuple-5: (SuspenseSee, action series, **pacing**, a drawback, negative, **[impact the tension and pace of the action])**

Subtask-II:

➢ sentiment-flip: (SuspenseSee, action series, pacing, positive, negative, **personal experiences and self-reflection**)

Figure 15: A conversation on domain of movies and entertainment.

Input Dialogue

① James: I've tried it from that food truck, and it's so delicious! The fresh ingredients make it stand out. (reply = -1)

🖼 [An image of a person happily enjoying a plate of tacos from a street vendor.]

② Sophia: I agree, the flavors are amazing. They blend traditional and unexpected ingredients perfectly, which really elevates the overall taste. (reply = 0)

③ Oliver: I think the portion sizes are too small, and they need to improve that. (reply = 1)

🔊 [An audio comparing portion sizes with another food truck.]

④ James: I have to disagree. I feel the portion sizes are just right, and I wouldn't want them any larger. (reply = 2)

⑤ Lucas: And the food truck's customer service is exceptional, they always go the extra mile! (reply = 1)

⑥ Isabella: I don't think so, I've had a few bad experiences with their customer service. Their customer service often makes me frustrated. (reply = 4)

⑦ Lucas: That's really disappointing to hear. Their inconsistent customer service is unacceptable, and it's really changed my view about them. (reply = 5)

Output of NExT-GPT

Subtask-I:

- sextuple-1: (James, **it**, [flavors], so delicious, positive, the fresh ingredients make it stand out)
- sextuple-2: (Sophia, **it**, [flavors], amazing, positive, they blend traditional and unexpected ingredients perfectly)
- sextuple-3: (Oliver, tacos, portion sizes, too small, negative, **they need to improve that**)
- sextuple-4: (James, tacos, portion sizes, just right, **positive**, I wouldn't want them any larger)
- sextuple-5: (Lucas, food truck, customer service, exceptional, positive, they always go the extra mile)
- sextuple-6: (Isabella, food truck, customer service, frustrated, negative, a few bad experiences)
- sextuple-7: (Lucas, food truck, customer service, unacceptable, negative, **that's really disappointing to hear**)

Subtask-II:

- sentiment-flip: (Lucas, food truck, customer service, positive, negative, participant feedback and interaction)

Output of Sentica

Subtask-I:

- sextuple-1: (James, **it**, [flavors], so delicious, positive, the fresh ingredients make it stand out)
- sextuple-2: (Sophia, **it**, [flavors], amazing, positive, they blend traditional and unexpected ingredients perfectly)
- sextuple-3: (Oliver, tacos, portion sizes, too small, negative, **they need to improve that**)
- sextuple-4: (James, tacos, portion sizes, just right, **positive**, I wouldn't want them any larger)
- sextuple-5: (Lucas, food truck, customer service, exceptional, positive, they always go the extra mile)
- sextuple-6: (Isabella, food truck, customer service, frustrated, negative, I've had a few bad experiences with their customer service)
- sextuple-7: (Lucas, food truck, customer service, unacceptable, negative, **that's really disappointing to hear**)

Subtask-II:

- sentiment-flip: (Lucas, food truck, customer service, positive, negative, participant feedback and interaction)

Output of Sentica+CoS

Subtask-I:

- sextuple-1: (James, **tacos**, [flavors], so delicious, positive, the fresh ingredients make it stand out)
- sextuple-2: (Sophia, **tacos**, [flavors], amazing, positive, they blend traditional and unexpected ingredients perfectly, which really elevates the overall taste)
- sextuple-3: (Oliver, tacos, portion sizes, too small, negative, **[compare with another food truck]**)
- sextuple-4: (James, tacos, portion sizes, just right, **positive**, I wouldn't want them any larger)
- sextuple-5: (Lucas, food truck, customer service, exceptional, positive, they always go the extra mile)
- sextuple-6: (Isabella, food truck, customer service, frustrated, negative, I've had a few bad experiences with their customer service)
- sextuple-7: (Lucas, food truck, customer service, unacceptable, negative, **[hear about the negative experiences shared by Isabella]**)

Subtask-II:

- sentiment-flip: (Lucas, food truck, customer service, positive, negative, participant feedback and interaction)

Output of Sentica+Cos+PpV

Subtask-I:

- sextuple-1: (James, **tacos**, [flavors], so delicious, positive, the fresh ingredients make it stand out)
- sextuple-2: (Sophia, **tacos**, [flavors], amazing, positive, they blend traditional and unexpected ingredients perfectly, which really elevates the overall taste)
- sextuple-3: (Oliver, tacos, portion sizes, too small, negative, **[compare with another food truck]**)
- sextuple-4: (James, tacos, portion sizes, just right, **neutral**, I wouldn't want them any larger)
- sextuple-5: (Lucas, food truck, customer service, exceptional, positive, they always go the extra mile)
- sextuple-6: (Isabella, food truck, customer service, frustrated, negative, I've had a few bad experiences with their customer service)
- sextuple-7: (Lucas, food truck, customer service, unacceptable, negative, **[hear about the negative experiences shared by Isabella]**)

Subtask-II:

- sentiment-flip: (Lucas, food truck, customer service, positive, negative, participant feedback and interaction)

Figure 16: A conversation on domain of food and cuisine.