Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study

Anonymous ACL submission

Abstract

Recently, ChatGPT has drawn great attention from both the research community and the public. We are particularly interested in whether it can serve as a universal sentiment analyzer. To this end, in this work, we provide a preliminary evaluation of ChatGPT on the understanding of *opinions*, *sentiments*, and *emotions* contained in the text. Specifically, we evaluate it in three settings, including *standard* evaluation, *polarity shift* evaluation and *open-domain* evaluation. We conduct an evaluation on 7 representative sentiment analysis tasks covering 17 benchmark datasets and compare ChatGPT with fine-tuned BERT and corresponding state-of-the-art (SOTA) models on them. We also attempt several popular prompting techniques to elicit the ability further. Moreover, we conduct human evaluation and present some qualitative case studies to gain a deep comprehension of its sentiment analysis capabilities.

1 Introduction

Recently, Large language models (LLMs) have profoundly affected the whole NLP community with their amazing zero-shot ability on various NLP tasks (Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022; Zhang et al., 2022a, *inter alia*). More recently, ChatGPT[1] has appeared out of the blue via interacting with people conversationally. It can conduct fluent conversations with people, write code as well as poetry, solve mathematical problems (Frieder et al., 2023) and so on, which has attracted widespread public attention.

However, despite its huge success, we still know little about the capability boundaries, i.e., where it does well and fails. In this work, we are interested in how ChatGPT performs on the sentiment analysis tasks, i.e., *can it understand the opinions, sentiments, and emotions contained in the text?* To answer this question, we conduct a preliminary evaluation on 7 representative sentiment analysis tasks[2] and 17 benchmark datasets, which involves three different settings including *standard* evaluation, *polarity shift* evaluation and *open-domain* evaluation (refer to Figure 1). We compare ChatGPT with fine-tuned small language models like BERT (Devlin et al., 2019) and corresponding SOTA models (if any) on each task for reference. We also attempt several popular prompting techniques, such as *chain-of-thought (CoT)* (Wei et al., 2022) and *self-consistency* (Wang et al., 2022), to induce the ability of ChatGPT. The main findings of this work are as follows:

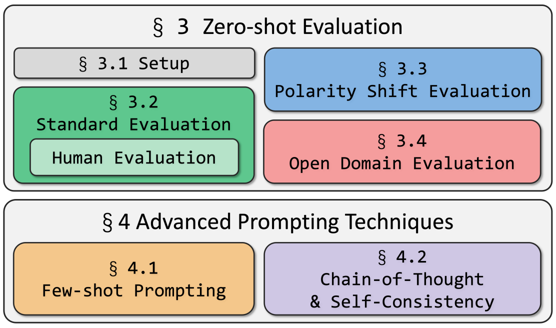

§ 3 Zero-shot Evaluation
§ 3.1 Setup
§ 3.2 Standard Evaluation
Human Evaluation
§ 3.3 Polarity Shift Evaluation
§ 3.4 Open Domain Evaluation
§ 4 Advanced Prompting Techniques
§ 4.1 Few-shot Prompting
§ 4.2 Chain-of-Thought & Self-Consistency

Figure 1: The overview of our evaluation.

❶ ChatGPT demonstrates impressive zero-shot capabilities in sentiment classification tasks and can rival fine-tuned BERT, although it still trails behind the domain-specific fully-supervised SOTA models (☞ § 3.2).

❷ Compared to fully-supervised highly competitive baselines we setup, ChatGPT achieves reasonable zero-shot performance on CSI but struggles on CEE (☞ § 3.2).

---

[1] `https://chat.openai.com/`

[2] They are Sentiment Classification (SC), Aspect-Based Sentiment Classification (ABSC), End-to-End Aspect-Based Sentiment Analysis (E2E-ABSA), Comparative Sentences Identification (CSI), Comparative Element Extraction (CEE), Emotion Cause Extraction (ECE), and Emotion-Cause Pair Extraction (ECPE).

❸ Compared to fully-supervised strong baselines, ChatGPT demonstrates impressive emotion cause analysis ability with significantly higher performance on ECE but lower performance on ECPE (☞ § 3.2).

❹ ChatGPT seems less accurate on sentiment information extraction tasks like E2E-ABSA and CEE. We observe that ChatGPT can often make reasonable predictions but can not strictly match the dataset annotations. Our human evaluation finds that ChatGPT actually performs more desirable, not as poor as metrics indicate. (☞ § 3.2)

❺ When coping with the *polarity shift* phenomenon (e.g., negation and speculation), a challenging problem in sentiment analysis, ChatGPT can make more accurate predictions than fine-tuned BERT. (☞ § 3.3)

❻ Compared to training domain-specific models, which typically perform poorly when generalized to unseen domains, ChatGPT demonstrates its powerful *open-domain* sentiment analysis ability in general, though its performance is quite limited in a few specific domains. (☞ § 3.4)

❼ Few-shot prompting (i.e., equipping with a few random examples in the input) can significantly improve performance across tasks and domains, surpassing fine-tuned BERT in some cases, though still inferior to SOTA models (☞ § 4.1). Applying CoT to the evaluated tasks does not yield gains but diminishes performance. In contrast, self-consistency reliably improves results (☞ § 4.2).

In summary, compared to training a specialized sentiment analysis system for each domain or dataset, **ChatGPT can already serve as a universal and well-behaved sentiment analyzer**.

2 Background and Related Work

2.1 Large Language Models

With the emergence of GPT-3 (Brown et al., 2020), Large language models (LLMs) were spotlighted. They typically have lots of model parameters and are trained on massive volumes of unstructured data at huge computational costs, including but not limited to Gopher (Rae et al., 2021), Megatron-Turing NLG 530B (Smith et al., 2022), LaMDA (Thoppilan et al., 2022), Chinchilla (Hoffmann et al., 2022), PaLM (Chowdhery et al., 2022), OPT (Zhang et al., 2022a), LLaMA (Touvron et al., 2023), and GPT-4 (OpenAI, 2023). As a result, given a simple task instruction, they are able to adapt directly to a new task in a training-free manner. In addition to the task instruction, the predictions will be more accurate and controllable if LLMs could be provided some demonstration examples, an ability known as *in-context learning* (Brown et al., 2020).

Lately, OpenAI has released ChatGPT, a chatbot fine-tuned from GPT-3.5 via reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), drawing increasingly great attention. Next, researchers start exploring its abilities and limitations, testing it on various benchmarks (Gilson et al., 2022; Frieder et al., 2023; Guo et al., 2023; Jiao et al., 2023; Zhuo et al., 2023; Zhong et al., 2023; Ye et al., 2023; Laskar et al., 2023). For example, Bang et al. (2023) evaluate the multitask, multilingual, and multimodal aspects of ChatGPT, Wang et al. (2023) conduct a robustness evaluation from the adversarial and out-of-domain perspective, and Borji (2023) summarizes 11 categories of failures towards ChatGPT. Related to our work, Zhong et al. (2023) analyze the language understanding ability of ChatGPT on GLUE (Wang et al., 2018). In this work, we especially concentrate on analyzing its sentiment analysis ability, aiming to answer the question via a rigorous and comprehensive evaluation, i.e., *whether ChatGPT can be a good sentiment analyzer.* We also discuss some closely related work in § A.1.

2.2 Sentiment Analysis

Sentiment analysis seeks to identify people's *opinions*, *sentiments*, and *emotions* in the text, such as customer reviews, social media posts, and news articles (Liu et al., 2005; Liu, 2015). As one of the most active fields in Natural Language Processing (NLP), it has made rapid progress with the help of *deep learning* (Zhang et al., 2018; Yadav and Vishwakarma, 2020). Among the myriad of tasks associated with sentiment analysis, this paper is primarily concerned with 4 representative task categories, including (sentence-level) sentiment classification (SC), aspect-based sentiment analysis (ABSA), comparative opinion mining (COM), and emotion cause analysis (ECA). For ease of understanding, we will briefly introduce these tasks next.

SC aims to identify the sentiment polarity of a given text, whether it is positive or negative. ABSA is designed to mine fine-grained aspect terms in the review and determine the sentiment polarity toward each aspect (Liu, 2012; Pontiki et al., 2014; Zhang et al., 2022b). We mainly focus on aspect-based sentiment classification (ABSC) and End-to-End Aspect Based Sentiment Analysis (E2E-ABSA) among many subtasks in ABSA. COM seeks to identify comparative sentences, extract the comparative elements, and obtain the corresponding comparative opinion tuples (Jindal and Liu, 2006; Liu et al., 2021). We mainly concentrate on comparative sentences identification (CSI) and comparative element extraction (CEE), i.e., extracting the tuple of (subject, object, comparative aspect, comparison type). The purpose of ECA is to extract the potential *cause clauses* given the emotion clause or extract the potential pair of *emotion clause* and *cause clause* in the text, which correspond to emotion cause extraction (ECE) (Gui et al., 2016) and emotion cause pair extraction (ECPE) (Xia and Ding, 2019), respectively.

In this paper, we are also concerned with two challenging problems in sentiment analysis: *polarity shift* and *open-domain* (Zong et al., 2021). Polarity shift refers to the linguistic phenomenon where the sentiment polarity (positive or negative) of a text shifts over time, context, or with respect to other texts (Li et al., 2010; Xia et al., 2016). Understanding sentiment *polarity shift* is crucial for building accurate sentiment analysis systems. As another challenging issue, *open-domain sentiment analysis* aims to understand the general sentiment of text regardless of the domain, whereas existing sentiment analysis systems typically focus on analyzing the sentiment of texts related to a particular domain (Cambria et al., 2012; Zhang et al., 2015; Luo et al., 2022). Addressing the above two issues is essential to building robust and effective sentiment analysis systems. In this work, we will examine whether ChatGPT can solve them.

3 Evaluation

In this section, we will first introduce the evaluation setup (§ 3.1) followed by *standard* evaluation (§ 3.2), *polarity shift* evaluation (§ 3.3) and *open-domain* evaluation (§ 3.4), as illustrated in Figure 1. As mentioned earlier, the tasks involved in our evaluation are SC, ABSC, E2E-ABSA, CSI, CEE, ECE, and ECPE.

3.1 Setup

Comparison Systems. We compare ChatGPT with the state-of-the-art (SOTA) (if any) models on end-tasks. Since SOTA models typically have some task-specific designs, we also provide the results of a commonly used baseline (e.g., fine-tuned BERT[3]) on each task for reference. For SC, we adopt the most common practice, i.e., using the final hidden representation of the `[CLS]` token as the sentence embedding and feeding it into a linear layer for classification. As for ABSC, we concatenate the review sentence and the aspect term via the special token `[SEP]` and classify the sentiment polarity based on the final hidden representation of `[CLS]`. We employ the joint tagging scheme (Li et al., 2019) to perform the E2E-ABSA task. For CSI, we report the performance of Multi-Stage$_{\text{BERT}}$ derived from (Liu et al., 2021) for reference. For CEE, given the complexity of modeling this task, we reformulate it into a text generation task based on `T5-Base` similar to GAS (Zhang et al., 2021), i.e., predicting the sequences of comparison tuples given the input review. We employ PAE-DGL (Ding et al., 2019) and ECPE-2D (Ding et al., 2020) as comparison models for ECE and ECPE, respectively. Unless otherwise specified, the above baseline models are rerun and repeated three times based on our evaluation settings.

Usage of ChatGPT. We mainly use ChatGPT with a specific version of `gpt-3.5-turbo-0301` for evaluation in this work, given its lower cost and improved performance (as stated in the OpenAI documentation[4]). We set the temperature to 0, making the outputs mostly deterministic for the identical inputs. Following Jiao et al. (2023), we ask ChatGPT to generate the task instruction for each task to elicit its ability to the corresponding task. For example, the prompt for E2E-ABSA is “`Given a review, extract the aspect term(s) and determine their corresponding sentiment polarity. Review: {sentence}`”. Due to limited space, please refer to Table 6 and Appendix A.2 for complete prompts and prompts details, respectively. We report the zero-shot results of ChatGPT unless otherwise specified. We manually observe and record the predictions as the responses of ChatGPT do not always follow a cer-

---

[3] All models use `BERT-base-uncased` version and are coupled with a linear layer if necessary.

[4] `https://platform.openai.com/docs/models/gpt-3-5`

tain pattern under the zero-shot setting.

Evaluation Metrics. We use accuracy and macro F1 score to evaluate sentiment classification tasks. We employ accuracy as the metric for CSI. For tasks involving elements extraction such as E2E-ABSA and CEE, we employ micro F1 score, i.e., a tuple is regarded as correct if and only if all elements inside it are exactly the same as the corresponding gold label. For ECE and ECPE, we compute the F1 score of cause clauses and emotion-cause clause pairs for evaluation, respectively.

3.2 Standard Evaluation

In this part, we evaluate ChatGPT on 7 representative sentiment analysis tasks and report its results on related benchmark datasets.

Datasets. We choose SST-2 (Socher et al., 2013) as the testbed of SC. Since the test set of SST-2 is not public, we use its validation set for evaluation. We employ the SemEval 2014-ABSA Challenge Datasets (Pontiki et al., 2014) to evaluate the ability of ChatGPT to ABSA. For CSI and CEE, we employ the Camera dataset (Kessler and Kuhn, 2014; Liu et al., 2021). For ECE and ECPE, we adopt the Emotion Cause Dataset (Gui et al., 2016; Xia and Ding, 2019) and sample 100 examples from this. Except as noted above, we evaluate the remaining datasets on the full test set. The statistics are shown in the third column of Table 1.

Results. The comparison results are shown in Table 1. Overall, ChatGPT demonstrates highly competitive sentiment analysis performance compared with baseline models, albeit often being far inferior to SOTA models. Specifically, we observe that ChatGPT is on par with fine-tuned small language models (i.e., BERT) in sentiment classification tasks, despite being inferior to SOTA models. Secondly, when evaluated on E2E-ABSA, the performance of ChatGPT is indeed inferior to fine-tuned BERT, and the performance gap varies across domains. We speculate that the poorer performance on 14-Laptop is due to the presence of more proprietary terms and specific expressions in this domain. Thirdly, for the challenging COM tasks (i.e., CSI and CEE), which typically involve implicit expressions, although achieving reasonable performance on CSI, it exhibits extremely undesirable performance on CEE. These results are far from satisfactory compared with fine-tuned baselines. Finally, ChatGPT exhibits reasonably good emotion analysis ability. We find that ChatGPT can comprehend the given document thoroughly, for instance, being capable of identifying multiple reasons and extracting emotion clauses and cause clauses even when they are distant. We also observe that ChatGPT can make some reasonable predictions, whereas the corresponding annotations are not in the dataset.

Human Evaluation. In light of the poor performance on certain tasks, we naturally raise a question: *are the predictions of ChatGPT truly unreasonable?* To acquire a more profound comprehension of the prediction results from ChatGPT, we conduct a human evaluation on E2E-ABSA and CEE owing to their unsatisfactory performance. Upon observation of the predicted results, ChatGPT has made many plausible predictions. However, these either did not exactly match the ground truth, or there are no corresponding annotations in the dataset, leading to a subpar performance on the exact-match evaluation. For E2E-ABSA, even though the predictions of ChatGPT are not accurate based on exact-match evaluation, it can still infer some highly reasonable aspect categories for the aspect terms thanks to its text generation paradigm. This also demonstrates its ability to identify implicit expressions to some extent. For instance, given the sentence "*Runs real quick.*", the ground truth is "(*Runs*, positive)" whereas the prediction of ChatGPT is "(*Speed*, positive)". For CEE, the predictions of ChatGPT express the same meaning as the ground truth but in an inconsistent form. As an example, the meaning expressed by ChatGPT is "*The SD800 is better than the SD700.*", whereas the ground truth meaning is "*The SD700 is worse than the SD800.*", where the "SD700" and "SD800" refer to the products being compared. From the perspective of sentiment analysis application, this is equally effective. Therefore, to align the predictions of ChatGPT with the annotation standard of existing datasets, we follow a few simple rules for human evaluation[5]:

☞ For any extra generated tuples, if they are reasonable but absent from the annotations, we will remove them from the prediction results. Otherwise, we will keep them.

☞ We also consider an aspect-sentiment or comparative opinion tuple correct if the boundary of aspect or entity is predicted incorrectly but unambiguously, and the predicted sentiment or preference is also correct.

---

[5] See Appendix A.4 for examples

Task	Datasets	#Test	Metric	Fine-tuned		Zero-shot	
				Baseline	SOTA	ChatGPT	+ Human
SC	SST-2	872	Acc	95.47†	**97.50**$^{\alpha}$	93.12	-
ABSC	14-Restaurant	1119	Acc / F1	83.94† / 75.28†	**89.54 / 84.86**$^{\beta}$	83.85 / 70.57	-
	14-Laptop	632	Acc / F1	77.85† / 73.20†	**83.70 / 80.13**$^{\gamma}$	76.42 / 66.79	-
E2E-ABSA	14-Restaurant	496	F1	77.75†	**78.68**$^{\delta}$	69.14	83.86
	14-Laptop	339	F1	66.05†	**70.32**$^{\delta}$	49.11	72.77
CSI	Camera	661	F1	**93.04**§	-	74.89	-
CEE	Camera	341	F1	**34.41**$^{\flat}$	-	9.10	51.28
ECE	Emotion Cause Dataset	100	F1	69.46‡	-	**74.01**	-
ECPE	Emotion Cause Dataset	100	F1	**65.20**$^{\natural}$	-	52.44	-

Table 1: Performance comparison among ChatGPT, fine-tuned baselines, and SOTA models on 9 datasets. #Test denotes the number of examples used for evaluation. † denotes the performance of fine-tuned BERT we implement. ‡ and ♮ denote the performance of PAE-DGL (Ding et al., 2019) and ECPE-2D (Ding et al., 2020) obtained by re-running experiments. § denotes the model performance of Multi-Stage$_{\text{BERT}}$ derived from Liu et al. (2021) while ♭ represents the results of our implemented GAS-Extraction-style baseline (Zhang et al., 2021). α, β, γ, and δ denote the results derived from T5-11B (Raffel et al., 2020), DPL (Zhang et al., 2022c), RILGNet (Li et al., 2022) and SyMux (Fei et al., 2022), respectively. "+ Human" denotes the performance with human evaluation. The best results are in **bold** except for human evaluation results.

☞ We also regard a prediction that paraphrases the ground truth to be correct, given the text generation paradigm.

The human evaluation results are shown in the last column of Table 1. It is surprising but reasonable to observe that the zero-shot performance of ChatGPT is boosted by 19% (average) and 42% on E2E-ABSA and CEE, respectively, compared to the original results. Moreover, it also significantly surpasses the previous performance of the baseline and SOTA. Although this human evaluation is very lenient for ChatGPT and may not be fair to baselines, at least it can demonstrate that the predictions of ChatGPT indeed align with human preferences (although not align with the annotation standard of the dataset) owing to RLHF and prove the potential of ChatGPT as a universal sentiment analyzer.

Case Study. We also conduct the qualitative analysis for the predictions of ChatGPT. Due to the limited space, please refer to Appendix A.5.

3.3 Polarity Shift Evaluation

Comprehending the phenomenon of *polarity shift* in sentiment analysis is crucial for developing robust and reliable sentiment analysis systems. In this part, we evaluate the ability of ChatGPT to cope with the *polarity shift* problem. Specifically, we mainly focus on the situations of negation and speculation and consider two sentiment classification tasks, SC and ABSC.

Datasets. Since there are few datasets tailored to *polarity shift* for SC, we derive two subsets from SST-2 validation set using a heuristic rule for the evaluation of negation and speculation, namely SST-2-Negation and SST-2-Speculation. In short, it entails identifying whether a sentence contains any negation or speculation words. For instance, we assign a sentence to the negation evaluation subset if it includes the word "*never*". More details are provided in Appendix A.3. As for ABSC, we adopt the 14-Res-Negation, 14-Lap-Negation, 14-Res-Speculation, and 14-Lap-Speculation introduced by Moore and Barnes (2021), which are annotated for negation and speculation, respectively. The statistics are shown in Table 7.

Baseline Details. Generally, we fine-tune BERT on the original training set (e.g., SST-2) and evaluate on polarity-shifting test sets, e.g., SST-2-Negation and SST-2-Speculation.

Results. We conduct experiments on six evaluation datasets, and the comparison results are shown in Table 2. Compared to fine-tuned BERT, ChatGPT exhibits greater robustness in *polarity shift* scenarios. Essentially speaking, the *polarity shift* evaluation we conduct can be characterized as an *out-of-distribution* (OOD) evaluation scenario. Not surprisingly, we observe that fine-tuned BERT experiences varying degrees of performance degradation across datasets compared to standard evaluation results. In comparison, ChatGPT is more

Task	Shifting Type	Dataset	Fine-tuned	Zero-shot
			BERT	ChatGPT
SC	Negation	SST-2-Neg.	**90.68**	**90.68**
	Speculation	SST-2-Spec.	**92.05**	**92.05**
ABSC	Negation	14-Res-Neg.	70.93	**79.66**
			61.90	**69.12**
		14-Lap-Neg.	60.25	**72.73**
			53.97	**67.27**
	Speculation	14-Res-Spec.	64.29	**77.01**
			60.53	**68.45**
		14-Lap-Spec.	40.86	**47.47**
			39.40	**46.96**

Table 2: Performance comparison between ChatGPT and BERT on six datasets when dealing with negation and Speculation linguistic phenomena, measured by accuracy (top) and macro F1 score (bottom). The best results are in **bold**.

robust, especially on ABSC, where ChatGPT outperforms fine-tuned BERT by 10% in terms of average accuracy and 8% in terms of average F1 score. Furthermore, we also find that the speculation case in *polarity shift* appears more challenging than the negation case, as the results of the former is poorer.

Case Study. We conduct qualitative analysis for the predictions of ChatGPT in the case of *polarity shift*. Refer to Appendix A.6 for details.

3.4 Open Domain Evaluation

Existing systems are typically trained on specific domains or datasets, leading to suboptimal generalization performance when dealing with unseen domains. However, an ideal sentiment analysis system could be applied to data from diverse domains. In this part, we evaluate the capability of ChatGPT to handle *open-domain* sentiment analysis tasks (i.e., ABSC and E2E-ABSA).

Datasets. As there is currently no widely used *open-domain* evaluation dataset, we sample 30 examples from each domain of existing 10 ABSA datasets according to the original data distribution, resulting in a total of 300 samples both for ABSC and E2E-ABSA. The ten datasets involved are Restaurant (Pontiki et al., 2014), Laptop (Pontiki et al., 2014), Device (Hu and Liu, 2004), Service (Toprak et al., 2010), Books, Clothing, Hotel (Luo et al., 2022), Twitter (Dong et al., 2014), Financial News Headlines (Sinha et al., 2022), METS-CoV (Zhou et al., 2022), covering various domains such as restaurant reviews, product reviews, social media, finance, and medicine. Note that Books, Hotel, and Clothing are originally document-level ABSA datasets with hierarchical entity-aspect-sentiment annotations. We randomly sample 30 sentences from each dataset and only use the aspect-sentiment annotations.

Baseline Details. To simulate the *open-domain* setting, we hold out some datasets, fine-tune BERT on the remaining datasets, and select checkpoints based on the mixture of the corresponding validation sets. Specifically, we set the following settings: (1) *single-source*: the model is trained on one dataset then evaluated on all datasets. Here, we choose Restaurant and Laptop as the testbed; (2) *multi-source*: the model is trained sequentially on nine datasets and then evaluated on the remaining one. Finally, we also fully-supervisedly fine-tune BERT and report the results for reference.

Results. In terms of ABSC, ChatGPT demonstrates a more compelling *open-domain* ability than BERT despite being fine-tuned on this task. As shown in Table 3, ChatGPT matches or even outperforms multi-domain fine-tuned BERT on 7 out of 10 domains in sentiment classification metrics (accuracy or macro-F1) while surpassing it by 8% in accuracy and 7% in F1 score on average across 10 datasets. It is worth mentioning that ChatGPT even performs comparably to full-supervised BERT, which shows its compelling generalization ability. Interestingly, fine-tuning on multiple domains does not necessarily lead to improved performance. For example, we observe that it results in a significant decrease in performance in certain datasets such as Finance and METS-Cov. Table 4 shows ChatGPT exhibits moderate performance on E2E-ABSA under the exact-match evaluation despite in the zero-shot manner. For example, it even beat BERT models on some domains (e.g., restaurant, service, and finance), which are fine-tuned on the nine domains.

Despite its success, we can observe that the performance of ChatGPT is quite poor in some domains, especially social media relevant domains (i.e., twitter, finance, METS-Cov), which suggests that improving performance on these domains remains challenging. It should be noted that due to the use of exact-match evaluation, the actual results of ChatGPT may not be as poor as they appear. Similarly, through our human evaluation (as introduced in § 3.2), we can observe that ChatGPT has achieved an average performance improvement of 18% across domains, surpassing even BERT fine-tuned on nine domains. Again, although the

Model	Metric	Rest.	Lap.	Books	Cloth.	Hotel	Device	Service	Twitter	Finance	METS	Ave.
Fine-tuned on the Rest. domain												
BERT	Acc.	81.11	**77.78**	57.78	74.44	**86.67**	86.67	71.11	62.22	75.56	53.33	72.67
	F1	74.99	**70.60**	41.91	55.00	77.59	85.35	67.91	54.11	62.75	47.06	61.14
Fine-tuned on the Lap. domain												
BERT	Acc.	**84.44**	77.78	57.78	76.67	**86.67**	86.67	71.11	62.22	74.44	50.00	72.78
	F1	**78.76**	72.84	42.84	56.21	76.94	88.92	67.59	56.16	55.59	37.56	60.78
Fine-tuned on the 9 out-of-domains each time												
BERT	Acc.	80.00	76.67	**62.22**	**76.67**	85.56	94.44	**81.11**	**70.00**	31.11	38.89	69.67
	F1	69.63	59.83	46.11	**61.66**	75.34	98.11	**79.29**	**67.83**	31.58	35.65	59.99
Fully-supvised results												
BERT	Acc.	81.11	77.78	71.11	80.00	87.78	100.00	74.44	62.22	82.22	61.11	77.78
	F1	74.99	72.84	57.17	58.15	77.98	100.00	62.69	60.99	79.07	58.53	67.64
Zero-shot results												
ChatGPT	Acc.	83.33	73.33	60.00	70.00	**86.67**	**96.67**	76.67	66.67	**86.67**	**76.67**	**77.67**
	F1	61.16	53.41	**51.25**	59.65	**83.18**	**98.89**	65.30	64.22	**72.35**	**55.56**	**66.50**

Table 3: Performance comparison between ChatGPT and fine-tuned BERT for ABSC task on open-domain evaluation. We also report the domain-specific fully-supervised results (in gray) of BERT for reference. The best results (except for fully-supervised results) are in **bold**.

Model	Rest.	Lap.	Books	Cloth.	Hotel	Device	Service	Twitter	Finance	Mets-Cov	Ave.
Fine-tuned on the Rest. domain											
BERT	76.55	43.57	38.35	29.57	64.07	50.74	27.01	1.67	7.74	3.27	34.25
Fine-tuned on the Lap. domain											
BERT	55.06	68.02	25.93	26.28	53.21	**60.19**	27.03	3.43	7.11	5.14	33.14
Fine-tuned on the 9 out-of-domains each time											
BERT	71.10	**59.36**	**46.64**	**50.72**	**74.85**	58.87	**47.67**	**42.90**	14.21	**10.27**	**47.66**
Fully-supvised results											
BERT	76.55	68.02	61.17	67.97	88.67	75.39	57.83	78.84	79.32	71.71	72.55
Zero-shot results											
ChatGPT	**72.73**	45.45	21.92	25.71	50.60	41.86	45.78	19.18	**38.36**	3.92	36.55
+ Human	82.22	64.00	29.41	34.78	62.5	69.23	63.89	52.63	76.92	9.88	54.55

Table 4: Performance comparison between ChatGPT and BERT for E2E-ABSA task on the open-domain evaluation. We report the domain-specific fully-supervised results (in gray) of BERT for reference. We also report the human evaluation results ("+ Human") of ChatGPT for reference. The best results (except for fully-supervised results and human evaluation results) are in **bold**.

comparison may not be entirely fair, it can demonstrate decent *open-domain* capabilities of ChatGPT, albeit with poor results in a few domains.

Case Study. We conduct qualitative analysis through four examples of ChatGPT on Books and METS-Cov, corresponding to the books and medicine domain, as shown in Figure 7. We also provided a detailed analysis in Appendix A.7.

4 Advanced Prompting Techniques

Given that ChatGPT still lags behind fine-tuned small language models (e.g., BERT) in some tasks and domains to a certain extent, we endeavor to seek help from some advanced prompting techniques to further elicit the capabilities of ChatGPT. Here, we adopt the ABSA tasks as the testbed.

4.1 Few-shot Prompting

We randomly select a few examples from the training dataset used for demonstration and concatenate them with the target input to prompt ChatGPT, a technique also known as *in-context learning* (Brown et al., 2020). We conduct few-shot prompting experiments on ABSC and ASPE with k (i.e., 1, 3, 9 and 27) examples. To reduce the variance caused by the sampling of demonstration examples, we adopt three random seeds for sampling to conduct experiments and report the average performance. We compare the resulting performance with fully-supervised BERT and SOTA.

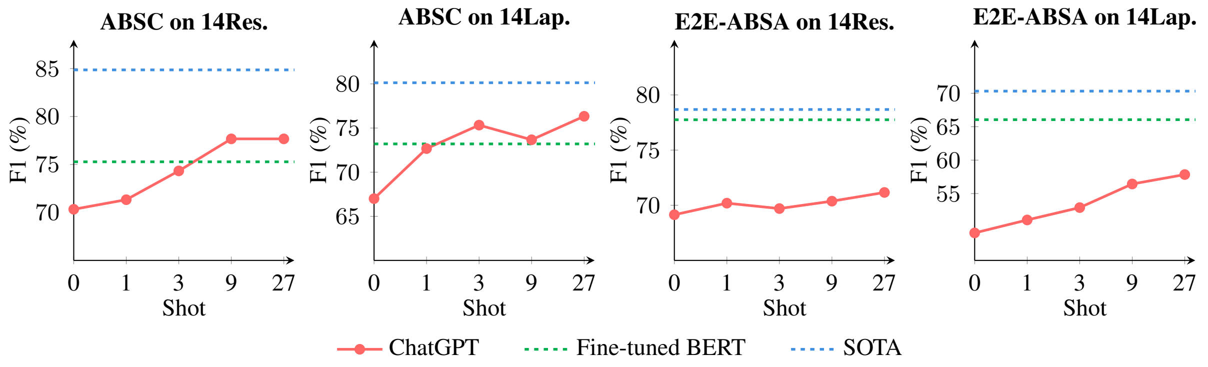


Figure 2: Few-shot prompting results on ABSC and E2E-ABSA tasks.

Prompting Methods	14-Res.	14-Lap.
Zero-shot prompting	69.14	49.11
Few-shot prompting (3 shot)	69.70	52.90
Few-shot prompting (9 shot)	70.37	56.43
Few-shot prompting (3 shot) + CoT	67.24	46.28
Few-shot prompting (9 shot) + CoT	64.98	50.19
3-shot + Self-Consist. ($N = 5$)	72.51	53.45
3-shot + Self-Consist. ($N = 10$)	72.87	54.22
3-shot + Self-Consist. ($N = 15$)	**73.22**	**55.01**
3-shot + CoT + Self-Consist. ($N = 5$)	69.12	48.73
3-shot + CoT + Self-Consist. ($N = 10$)	69.17	49.17
3-shot + CoT + Self-Consist. ($N = 15$)	70.39	49.77
Fine-tuned BERT	**77.75**	**66.05**

Table 5: Results of advanced prompting techniques on E2E-ABSA. N denotes the number of outputs sampled for the same input in the self-consistency technique.

Results. As presented in Figure 2, few-shot prompting can significantly improve the performance across tasks and datasets, even surpassing fine-tuned BERT in some cases. It improves the classification performance by 7% and 10% F1 score for ABSC on 14-Restaurant and 14-Laptop, respectively, with 27 demonstration examples. We can also observe certain improvements on ASPE, although the improvement curve is relatively flat. We also provide a case study, as shown in Figure 3.

4.2 Chain-of-Thought and Self-Consistency

Although few-shot prompting clearly improves the performance on ABSC, the performance on E2E-ABSA still lags far behind fine-tuned BERT. We attempt more advanced techniques, i.e., *manual few-shot chain-of-thought (CoT) prompting* (Wei et al., 2022) and *self-consistency* (Wang et al., 2022) on this task, to further elicit the ability. More details are provided in the Appendix A.8

Results. As shown in Table 5, we observe that equipping standard few-shot prompting with chain-of-thought does not bring the expected gains, but rather lead to a noticeable drop. This similar phenomenon was also observed in Ye and Durrett (2022) and Wang et al. (2022) but contrary to the observations in Zhong et al. (2023). We speculate that this may depend on the evaluation tasks. In contrast, self-consistency clearly improves the performance of few-shot prompting, regardless of whether CoT is equipped, once again confirming the effectiveness of this technique (albeit at the cost of increased inference complexity). Regrettably, while effective, it is still inferior to fine-tuned BERT. Future work could explore more efficient prompting methods, such as retrieval-based ones (Liu et al., 2022; Shi et al., 2023, *inter alia*).

5 Conclusion

In this work, we evaluate ChatGPT on a range of test sets and evaluation scenarios and compare its performance to fine-tuned BERT, exploring its capacity boundaries in various sentiment analysis tasks. ChatGPT exhibits magnificent zero-shot sentiment analysis abilities (e.g., sentiment classification, comparative opinion mining and emotion cause analysis), even matching with fine-tuned BERT and SOTA models trained with labeled data in respective domains at times. Compared to fine-tuned BERT, ChatGPT can handle the *polarity shift* problem more effectively in sentiment analysis and exhibits good performance in *open-domain* scenarios. In addition, we also explore some popular prompting techniques to further induce the capability of ChatGPT. Through experiments, we validate the effectiveness of them on sentiment analysis tasks and provide our findings. We aspire to galvanize future research through our empirical insights in sentiment analysis, LLMs and beyond.

Limitations

This work has several limitations as follows: (1) **Data leakage.** Currently, conducting rigorous evaluations for LLMs is extremely challenging. For example, it is difficult for us to determine whether the test set has been seen during the large-scale unsupervised pre-training, especially for models like ChatGPT , which are completely closed-source and can only be accessed through APIs. Nevertheless, in this work, we still find some deficiencies of ChatGPT, such as its sentiment analysis performance in some domains (e.g., medicine and social media) that leaves much to be desired. (2) **Prompt design.** We do not conduct extensive prompt engineering, so there are likely better prompts to obtain better performance. Nevertheless, we believe that ordinary users usually do not do very delicate prompt designs when using LLMs. Therefore, if the ChatGPT can achieve sufficiently robust performance on arbitrary prompts, this would better demonstrate its capability. (3) **Limited evaluation.** Our evaluation is mainly conducted on ChatGPT, without including other equally powerful models. Although we have also supplemented other evaluation results in Appendix A.9, such as `text-davinci-003`. Unfortunately, such models are either completely closed-source and we do not have access to APIs, or we do not have enough GPUs to rigorously evaluate their performance due to their huge model parameters. However, as a representative of currently the most powerful models, evaluation on ChatGPT can also enable us to understand what LLMs currently do well and not well, thereby inspiring future research.

Beyond this work, we believe some promising future directions could include: (1) **New evaluation benchmarks.** We need to propose new and comprehensive benchmarks from real-world scenarios. Meanwhile, evaluation methods are also worth paying attention to. Due to the text generation paradigm, commonly used exact-match may not truly characterize the model performance. In this paper, we adopt human evaluation to alleviate this issue. (2) **Implicit sentiment analysis.** Implicit expression is a very common linguistic phenomenon. For example, “I know real Indian food and this wasn’t it” does not contain explicit opinion words. Moreover, accurate judgment often requires common sense or domain knowledge. Our experiments also confirm that large language models generally perform poorly on implicit sentiment analysis (See Appendix 11 for results). Meanwhile, constructing comprehensive benchmarks for implicit sentiment analysis could be a promising direction. (3) **Enhancing the model capabilities in specific domains.** As shown in Table 3 and Table 4, we can see that the performance of ChatGPT is not satisfactory on many domains (such as books and twitter). Therefore, in the future, we could improve the performance on certain domains through domain-specific training.

Ethics Statement

We honour and support the ACL Ethics Policy. Our work aims to systematically evaluate the sentiment analysis capability of ChatGPT and thus inspire future research in a responsible and ethical manner. The data used for evaluation are from public benchmark datasets. This work does not involve human subjects, and we did not collect or process any personal identification information.

With respect to **the applications of ChatGPT in sentiment analysis**, we present the following broader considerations:

1. If strict accordance with annotations or norms is not required, ChatGPT can be used for sentiment analysis (via zero-shot or few-shot prompting);

2. If strict accordance is desired, fine-tuning a specialized model in a supervised manner is still a better approach;

3. For domain-specific applications, especially those requiring domain knowledge, training specialized models is still advised;

4. For open-domain applications requiring good generalization, ChatGPT is a viable option for sentiment analysis;

5. For domains with abundant labeled data, training a specialized model on the annotations is recommended;

6. For low-resource or even zero-resource domains, ChatGPT is a promising choice.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Ali Borji. 2023. A categorical archive of chatgpt failures. *CoRR*, abs/2302.03494.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Erik Cambria, Yangqiu Song, Haixun Wang, and Newton Howard. 2012. Semantic multidimensional scaling for open-domain sentiment analysis. *IEEE intelligent systems*, 29(2):44–51.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6343–6350. AAAI Press.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. ECPE-2D: emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3161–3170. Association for Computational Linguistics.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. 2022. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4121–4128. ijcai.org.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt. *CoRR*, abs/2301.13867.

Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, R. Andrew Taylor, and David Chartash. 2022. How does chatgpt perform on the medical licensing exams? the implications of large language models for medical education and knowledge assessment. *medRxiv*.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts?

comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Nitin Jindal and Bing Liu. 2006. Mining comparative sentences and relations. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 1331–1336. AAAI Press.

Wiltrud Kessler and Jonas Kuhn. 2014. A corpus of comparisons in product reviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2242–2248, Reykjavik, Iceland. European Language Resources Association (ELRA).

Md. Tahmid Rahman Laskar, M. Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *CoRR*, abs/2305.18486.

Longfeng Li, Haifeng Sun, Qi Qi, Jingyu Wang, Jing Wang, and Jianxin Liao. 2022. Modeling aspect correlation for aspect-based sentiment analysis via recurrent inverse learning guidance. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6887–6896, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shoushan Li, Sophia Y. M. Lee, Ying Chen, Chu-Ren Huang, and Guodong Zhou. 2010. Sentiment classification and polarity shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 635–643, Beijing, China. Coling 2010 Organizing Committee.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Bing Liu. 2015. *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 342–351. ACM.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yun Luo, Hongjie Cai, Linyi Yang, Yanxia Qin, Rui Xia, and Yue Zhang. 2022. Challenges for open-domain targeted sentiment analysis. *CoRR*, abs/2204.06893.

Andrew Moore and Jeremy Barnes. 2021. Multi-task learning of negation and speculation for targeted sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2838–2869, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

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

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. Sentfin 1.0: Entity-aware sentiment analysis for financial news. *J. Assoc. Inf. Sci. Technol.*, 73(9):1314–1335.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zheng, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model. *CoRR*, abs/2201.11990.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *CoRR*, abs/2302.12095.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *CoRR*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Rui Xia, Feng Xu, Jianfei Yu, Yong Qi, and Erik Cambria. 2016. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, 52(1):36–45.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artif. Intell. Rev.*, 53(6):4335–4385.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *CoRR*, abs/2303.10420.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *NeurIPS*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining Knowl. Discov.*, 8(4).

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, Lisbon, Portugal. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *CoRR*, abs/2305.15005.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 504–510. Association for Computational Linguistics.

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

Yiming Zhang, Min Zhang, Sai Wu, and Junbo Zhao. 2022c. Towards unifying the label space for aspect- and sentence-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 20–30, Dublin, Ireland. Association for Computational Linguistics.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? A comparative study on chatgpt and fine-tuned BERT. *CoRR*, abs/2302.10198.

Peilin Zhou, Zeqiang Wang, Dading Chong, Zhijiang Guo, Yining Hua, Zichang Su, Zhiyang Teng, Jiageng Wu, and Jie Yang. 2022. Mets-cov: A dataset of medical entity and targeted sentiment on COVID-19 related tweets. *CoRR*, abs/2209.13773.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of chatgpt: A diagnostic analysis. *CoRR*, abs/2301.12867.

Chengqing Zong, Rui Xia, and Jiajun Zhang. 2021. *Text Data Mining*. Springer Singapore.

A Appendix

A.1 Closely Related Work

Prior to our work, there were some early studies that evaluated ChatGPT's performance on common NLP tasks, but they only covered basic sentiment classification tasks, such as SST-2 (Zhong et al., 2023). Our work, however, is dedicated to a more comprehensive and thorough evaluation for ChatGPT within the sentiment analysis field, representing, to our knowledge, the first of its kind in this field. After our work was completed (over a month later), a new study emerged, evaluating more tasks like ASTE and ASQP within ABSA and also providing few-shot experimental results and human evaluations (Zhang et al., 2023). However, this study did not evaluate critical issues in sentiment analysis such as sentiment polarity shift and open-domain evaluation, which our work does address. Although different works may have their own focuses, we believe they all contribute significantly to providing more comprehensive evaluations in this field.

A.2 Prompts of ChatGPT

Following Jiao et al. (2023), we ask ChatGPT to generate the task instruction for each task to elicit its ability to the corresponding task. Taking the E2E-ABSA task as an example, our query is:

`Please give me three concise prompts for eliciting your ability to perform Aspect-Based Sentiment Analysis (i.e., extract the aspect terms and sentiment polarity). There is no need to give examples and do not limit the prompts to a specific product or domain.`

Then, we examine the generated three prompts on a small-scale (e.g., 50 examples) example set driving from the corresponding training set. We select the best and most reasonable one[6] according to the results[7]. The final prompts adopted for each task are shown in Table 6. During the evaluation process, we feed a prompt and corresponding test example to ChatGPT and obtain a generated response. We manually observe and record the results as the responses do not follow a certain pattern.

A.3 Preparation of Polarity Shift Evaluation Datasets

As previously mentioned, we drive SST-2-Neg and SST-2-Spec from SST-2 by detecting whether a sentence contains any negation or speculation words. The seed words adopted are shown in Table 8. And the statistics of involved datasets are shown in Table 7.

A.4 Examples on Human Evaluation

The exact-match metric has limitations for evaluating generative models like ChatGPT since they can produce reasonable outputs not matched to references. To better characterize ChatGPT's capabilities despite this, we manually refine its outputs before comparing them to those of baselines. We acknowledge this may seem unfair compared to unrefined baselines. However, our goal is to account for the limitations of the exact-match, not to boost ChatGPT's results unfairly. To further illustrate the rules we use as more intuitive and easier to understand, we provide some examples from the E2E-ABSA task, as shown in Table 9.

A.5 Case Study for Standard Evaluation

In this part, we conduct the qualitative analysis on ABSA tasks, COM tasks, and ECA tasks.

Case Study on ABSA. We conduct the qualitative analysis through two examples. Specifically, as shown in Figure 3, we present the results generated by ChatGPT for two test examples under zero-shot and few-shot settings, respectively. Given the example *"I did swap out the hard drive for a Samsung 830 SSD which I highly recommend"*, there are multiple aspect terms with different sentiment polarities in a sentence (e.g., the sentiment polarity of *"hard drive"* is neutral, and that of *"SSD"* is positive). We can observe that ChatGPT can not accurately identify the sentiment polarity of *"hard drive"* under the zero-shot setting. Similarly, in another test example *"I can say that I am fully satisfied with the performance that the computer has supplied."*, the aspect term extracted by ChatGPT is *"computer performance"*, which does not naturally exist in the sentence, indicating that ChatGPT may generate semantically reasonable aspect terms but without being aligned with the annotations in the dataset. However, under the few-shot setting (as introduced in § 4.1), after being equipped with a few demonstration examples, both of the above types of errors can be corrected by ChatGPT.

---

[6] When necessary, we would make minor adjustments to the prompts.

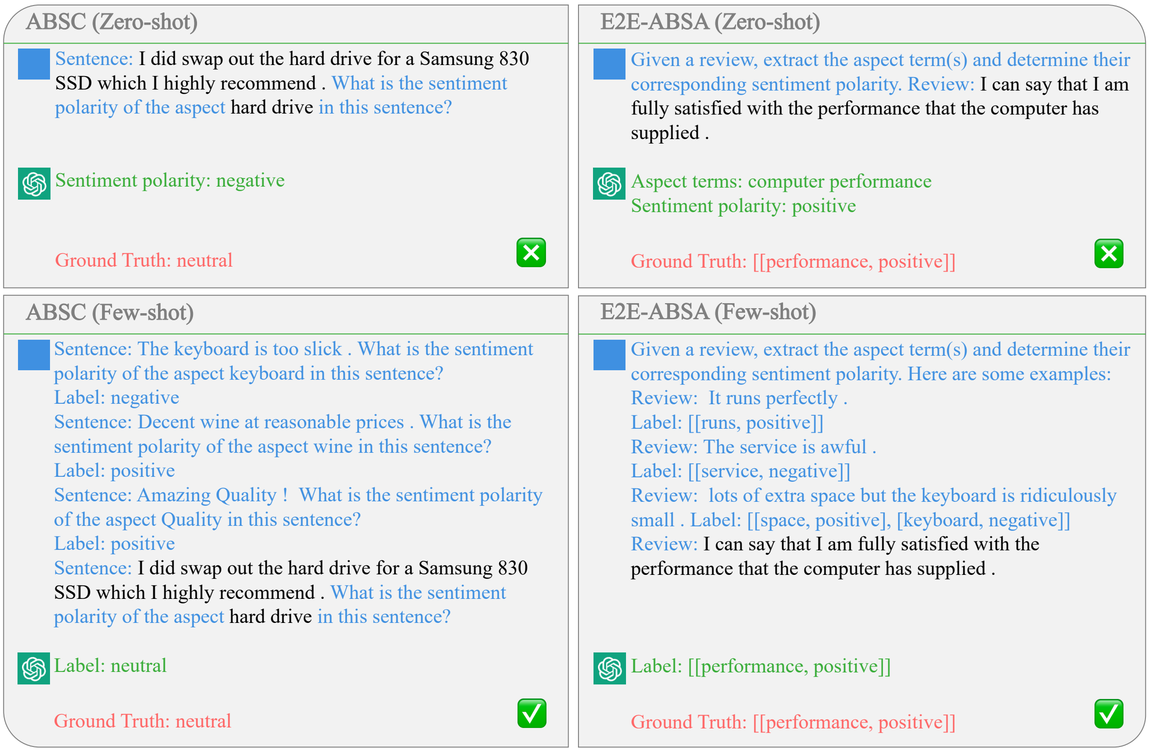
[7] We observe that different prompts have little effect on the performance. We also conducted three experiment repetitions and found minimal deviation in the results. Considering the cost of API calls, we only run the experiment once for the final evaluation unless otherwise specified.

Task	Prompt
SC	Given this text, what is the sentiment conveyed? Is it positive or negative? Text: `{sentence}`
ABSC	Sentence: `{sentence}` What is the sentiment polarity of the aspect `{aspect}` in this sentence?
E2E-ABSA	Given a review, extract the aspect term(s) and determine their corresponding sentiment polarity. Review: `{sentence}`
CSI	Does any comparison of products (including implicit products) exist in the product review: `{sentence}`? If so, outputs ‘TRUE’, else outputs ‘FALSE’.
CEE	The following product review contains comparison of products (including implicit products): `{sentence}`. Extract the subject and object of comparison, tell me which aspect of products is being compared, and tell me if the author of the review thinks the subject is better or worse than or similar to or different from the object.`\n` If multiple comparisons exist, output multiple comparisons.
ECE	Document: `{doc}` `\n` Each line in the above document represents a clause and the number at the beginning of each line indicates the clause ID. Clauses expressing emotions are referred to as “emotion clause” and clauses causing emotions are referred to as “cause clauses”. It has been identified that the clause with ID `{emo_id}`, `{emotion clause}` is an emotion clause, and the corresponding emotion keyword is `{emotion}`. Based on the above information, complete the following tasks: 1. Describe in one sentence the cause of the emotion clause with ID `{emo_id}`. 2. Based on the result of Task 1, output the ID of the cause clause that best fits the requirements. 3. According to the result of Task 2, match clauses with causality into pairs in the form “(emotion clause ID, cause clause ID)” and output all pairs as a set, such as (1,2),(3,4). Note: the emotion clause and the cause clause may be the same clause, and only the most obvious pairs need to be outputted.
ECPE	Document: `{doc}` `\n` Each line in the above document represents a clause and the number at the beginning of each line indicates the clause ID. Clauses expressing emotions are referred to as “emotion clause” and clauses causing emotions are referred to as “cause clauses”. Based on the above information, complete the following tasks: 1. Describe the emotions and their corresponding causes contained in the document in one sentence. 2. Output the ID of the emotion clause in task 1, you only need to find the one with the strongest intensity. 3. For each emotion clause in task 2, find the corresponding cause clause and output the cause clause ID, you only need to find the most suitable one. 4. Match clauses with causality into pairs in the form “(emotion clause ID, cause clause ID)” and output all pairs as a set, such as (1,2),(3,4). Note: the emotion clause and the cause clause may be the same clause, and only the most obvious pairs need to be outputted.

Table 6: The prompts used for prompting ChatGPT for each task. We manually design prompts for emotion cause analysis tasks (i.e., ECE and ECPE) due to the task complexity.

Case Study on COM. We conduct qualitative analysis through two examples of ChatGPT in the case of CSI and CEE tasks, as shown in Figure 4. For the CSI task, it can be seen that ChatGPT is able to accurately identify explicit product comparison sentences. However, when the compared objects are implicit products, ChatGPT often considers the sentence not to be a comparison sentence, such as the sentence *"However, focus accuracy was not as impressive."* ChatGPT assumes that there are no explicitly mentioned products in the comment and therefore determines that it is not a comparison sentence. For CEE task, although ChatGPT is able to correctly identify comparison sentences and extract comparative elements, it tends to exhibit paraphrase phenomena when generating answers. For example, in the example sentence *"It seems to get less light to the sensors than my old 4MP A80."* the annotation indicates that the comparison subject is *"worse"* than the comparison object. However, when replying, ChatGPT expresses it as the comparison object is *"better"* than the comparison subject. This situation causes alignment issues between the generated answers and the annotations during automatic evaluation.

Case Study on ECA. We also conduct qualitative analysis through two examples of ChatGPT in the case of ECE and ECPE tasks, as shown in



ABSC (Zero-shot)

Sentence: I did swap out the hard drive for a Samsung 830 SSD which I highly recommend . What is the sentiment polarity of the aspect hard drive in this sentence?

Sentiment polarity: negative

Ground Truth: neutral ✗

E2E-ABSA (Zero-shot)

Given a review, extract the aspect term(s) and determine their corresponding sentiment polarity. Review: I can say that I am fully satisfied with the performance that the computer has supplied .

Aspect terms: computer performance
Sentiment polarity: positive

Ground Truth: [[performance, positive]] ✗

ABSC (Few-shot)

Sentence: The keyboard is too slick . What is the sentiment polarity of the aspect keyboard in this sentence?
Label: negative
Sentence: Decent wine at reasonable prices . What is the sentiment polarity of the aspect wine in this sentence?
Label: positive
Sentence: Amazing Quality ! What is the sentiment polarity of the aspect Quality in this sentence?
Label: positive
Sentence: I did swap out the hard drive for a Samsung 830 SSD which I highly recommend . What is the sentiment polarity of the aspect hard drive in this sentence?

Label: neutral

Ground Truth: neutral ✓

E2E-ABSA (Few-shot)

Given a review, extract the aspect term(s) and determine their corresponding sentiment polarity. Here are some examples:
Review: It runs perfectly .
Label: [[runs, positive]]
Review: The service is awful .
Label: [[service, negative]]
Review: lots of extra space but the keyboard is ridiculously small . Label: [[space, positive], [keyboard, negative]]
Review: I can say that I am fully satisfied with the performance that the computer has supplied .

Label: [[performance, positive]]

Ground Truth: [[performance, positive]] ✓

Figure 3: Case study for ChatGPT on ABSC and E2E-ABSA in zero-shot and few-shot settings. The text in blue, black, green and red denote the given prompts, the examples to be evaluated, the responses of ChatGPT and the ground truths, respectively.

Task	Dataset	#Test
SC	SST-2-Negation	236
	SST-2-Speculation	88
ABSC	14-Res-Negation	1008
	14-Res-Speculation	448
	14-Lap-Negation	462
	14-Lap-Speculation	217

Table 7: The tasks and datasets involved in the polarity-shifting evaluation. #Test denotes the number of examples used for evaluation.

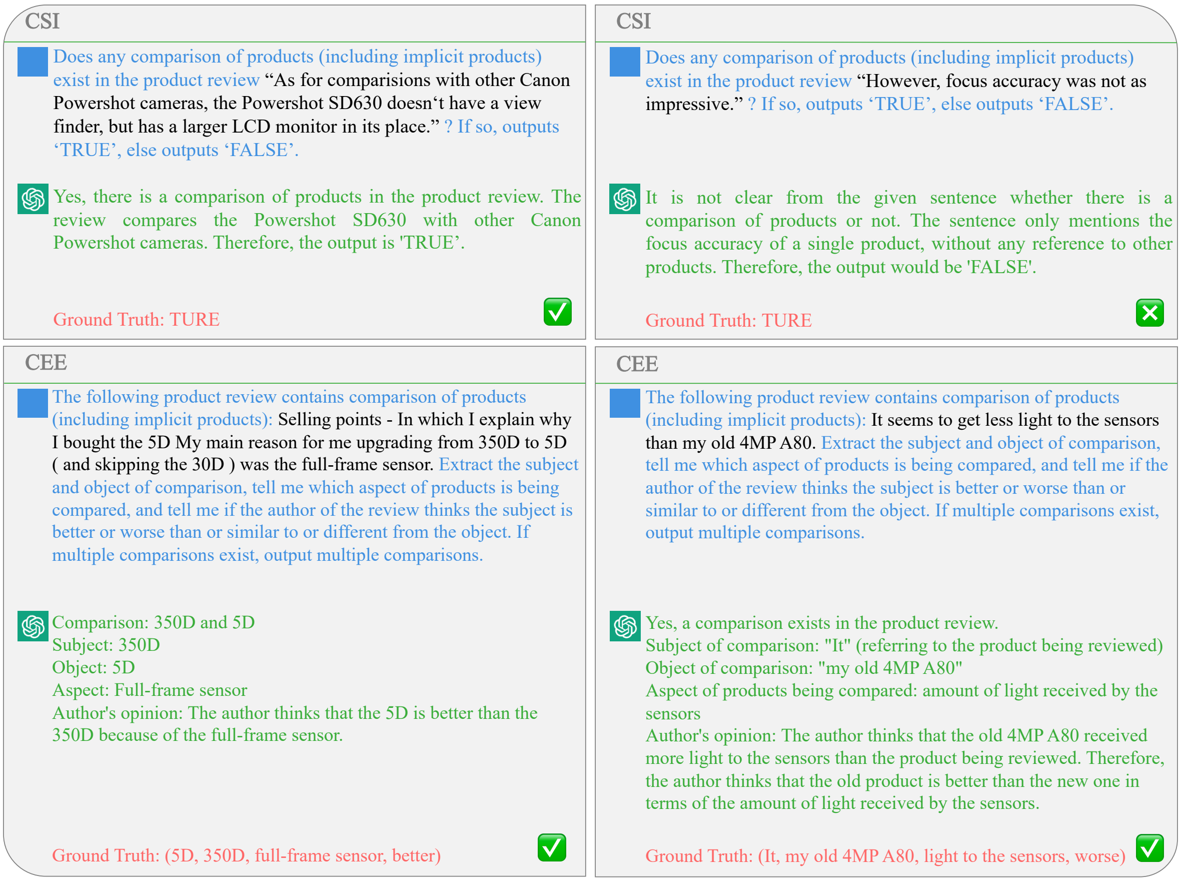
Figure 5. It can be seen that ChatGPT can perfectly follow our given instructions to complete the task. Given the emotion of a document, ChatGPT can accurately analyze its corresponding cause, but the clause ID is not output as required (refer to the upper part of Figure 5). For the same document, without providing the emotion clause beforehand, we find that ChatGPT can still correctly identify it, which corresponds to the ECPE task (refer to the bottom part of Figure 5). Although the identified cause clause may not be the most obvious, the prediction is still reasonable.

Shifting Type	Seed Words
Negation	n't, no, not, never, neither, nor, unless, but, however, rather than, not yet, not only, nonetheless, despite, although, even though, in spite of, unlikely
Speculation	if, would, could, should, seems, might, maybe, whether, unless, even if, if only, can't believe, grant that, guessing, suspect, hope, wish, let's probably

Table 8: Seed words used for deriving SST-2-Neg and SST-2-Spec from SST-2.

A.6 Case Study for Polarity Shift Evaluation

We conduct qualitative analysis through four examples of ChatGPT in the case of *polarity shift* (i.e., negation and speculation), as shown in Figure 6. Observing the two examples of SC in the first row, it becomes apparent that ChatGPT can accurately determine the overall sentiment polarity of sentences accompanied by *polarity shift* due to the presence of negation and speculation expressions.



CSI

Does any comparison of products (including implicit products) exist in the product review "As for comparisions with other Canon Powershot cameras, the Powershot SD630 doesn't have a view finder, but has a larger LCD monitor in its place." ? If so, outputs 'TRUE', else outputs 'FALSE'.

Yes, there is a comparison of products in the product review. The review compares the Powershot SD630 with other Canon Powershot cameras. Therefore, the output is 'TRUE'.

Ground Truth: TURE ✓

CSI

Does any comparison of products (including implicit products) exist in the product review "However, focus accuracy was not as impressive." ? If so, outputs 'TRUE', else outputs 'FALSE'.

It is not clear from the given sentence whether there is a comparison of products or not. The sentence only mentions the focus accuracy of a single product, without any reference to other products. Therefore, the output would be 'FALSE'.

Ground Truth: TURE ✗

CEE

The following product review contains comparison of products (including implicit products): Selling points - In which I explain why I bought the 5D My main reason for me upgrading from 350D to 5D (and skipping the 30D) was the full-frame sensor. Extract the subject and object of comparison, tell me which aspect of products is being compared, and tell me if the author of the review thinks the subject is better or worse than or similar to or different from the object. If multiple comparisons exist, output multiple comparisons.

Comparison: 350D and 5D
Subject: 350D
Object: 5D
Aspect: Full-frame sensor
Author's opinion: The author thinks that the 5D is better than the 350D because of the full-frame sensor.

Ground Truth: (5D, 350D, full-frame sensor, better) ✓

CEE

The following product review contains comparison of products (including implicit products): It seems to get less light to the sensors than my old 4MP A80. Extract the subject and object of comparison, tell me which aspect of products is being compared, and tell me if the author of the review thinks the subject is better or worse than or similar to or different from the object. If multiple comparisons exist, output multiple comparisons.

Yes, a comparison exists in the product review.
Subject of comparison: "It" (referring to the product being reviewed)
Object of comparison: "my old 4MP A80"
Aspect of products being compared: amount of light received by the sensors
Author's opinion: The author thinks that the old 4MP A80 received more light to the sensors than the product being reviewed. Therefore, the author thinks that the old product is better than the new one in terms of the amount of light received by the sensors.

Ground Truth: (It, my old 4MP A80, light to the sensors, worse) ✓

Figure 4: Case study for ChatGPT on CSI and CEE. The text in blue, black, green and red denote the given prompts, the examples to be evaluated, the responses of ChatGPT and the ground truths, respectively.

Regarding the two examples in the second row of the ABSC task, when multiple aspect terms are listed in parallel and involve the linguistic phenomena such as negation and speculation (e.g., "*faster and sleeker looking*" and "*super fast and had outstanding graphics*"), ChatGPT struggles to identify their sentiment polarity accurately, leading to incorrect predictions.

A.7 Case Study for Open-Domain Evaluation

We conduct qualitative analysis through four examples of ChatGPT on Books and METS-Cov, corresponding to the books and medicine domain, as shown in Figure 7. Regarding two examples in the first column, ChatGPT has difficulty accurately predicting sentiment in books domain since this domain usually contains unique expressions and plot descriptions that differ from typical product reviews. We also find that ChatGPT often generates reasonable aspect-sentiment pairs in the E2E-ABSA task, such as the *"(chapter creations, negative)"* (in the bottom left example). However, as we mentioned earlier, they are not originally annotated in the dataset, which is also an important reason affecting the performance of E2E-ABSA. As for the two examples in the second column, when dealing with a rare domain like medicine, ChatGPT can accurately determine the sentiment polarity of the given aspect term, but accurately extracting aspect-sentiment pairs remains a challenge.

A.8 Details for Chain-of-Thought and Self-Consistency Prompting

The chain-of-thought method augments each demonstration example in standard few-shot prompting with a chain of reasoning for the associated answer (Wei et al., 2022). We manually write CoT for randomly sampled examples. Self-consistency seeks to sample a diverse set of candidate outputs from LLMs and then aggregate the answers via a majority vote. We apply the temperature sampling with $T = 0.8$ as self-consistency is generally robust to sampling strategies (Wang et al., 2022). For the aggregation of answers, unlike the arithmetic reasoning task that typically has only one certain answer, the E2E-ABSA task we evaluate usually contains multiple aspect-sentiment tuples in an example. We adopt a heuristic approach by counting the frequency of each tuple in N sampled predictions and filtering by setting a frequency threshold to obtain the final prediction. We can finely control the answer aggregation by setting the threshold. In our experiments, we find that when $N = 15$, a threshold between 7 and 12 performs well.

A.9 Other Evaluation Results

Evaluation on text-davinci-003 Some readers might be curious about the performance of other powerful GPT-3.5 models in comparison to ChatGPT. To address this concern, we evaluate the powerful GPT-3.5 model, `text-davinci-003`, on some benchmarks. We carefully tune the evaluation to be as rigorous and controlled as possible, with temperature 0, top_p of 1, and 3 repeated runs to account for any variability (which is turned out to be negligible). As shown in Table 10, `text-davinci-003` achieves overall performance on par with ChatGPT.

Evaluation on Implicit Sentiment Analysis As an interesting and challenging direction, we also explore the evaluation on implicit sentiment analysis. Following the dataset split of implicit sentiment analysis described in (Li et al., 2021), we evaluate ChatGPT on the ABSC task and report BERT results (derived from (Li et al., 2021)) as a reference. We also evaluate the performance of `text-davinci-003`. Similarly, we run 3 trials and report the average F1 over the implicit subset and the full ABSC dataset (we find that the variance is small). As shown in Table 11, we can observe that these large language models perform poorly on implicit sentiment analysis, although `text-davinci-003` outperforms ChatGPT, both are weaker than fine-tuned BERT. These results suggest ample opportunities for future research.

Rule#1: For any extra generated tuples, if they are reasonable but absent from the annotations, we will remove them from the prediction results. Otherwise, we will keep them.

Example#1
Input: It is super fast and has outstanding graphics .
Output:
Aspect term: speed, graphics
Sentiment polarity: positive, positive
Ground Truth: [(graphics, positive)]
Refined Output:
Aspect term: graphics
Sentiment polarity: positive

Rule#2: We also consider an aspect-sentiment or comparative opinion tuple correct if the boundary of aspect or entity is predicted incorrectly but unambiguously, and the predicted sentiment or preference is also correct.

Example#1
Input: the hardware problems have been so bad , i ca n't wait till it completely dies in 3 years , TOPS !
Output:
Aspect term: hardware problems
Sentiment polarity: negative
Ground Truth: [(hardware, negative)]
Refined Output:
Aspect term: hardware
Sentiment polarity: negative
Example#2
Input: And the fact that it comes with an i5 processor definitely speeds things up.
Output:
Aspect term: processor
Sentiment polarity: positive
Ground Truth: [(i5 processor, positive)]
Refined Output:
Aspect term: i5 processor
Sentiment polarity: positive

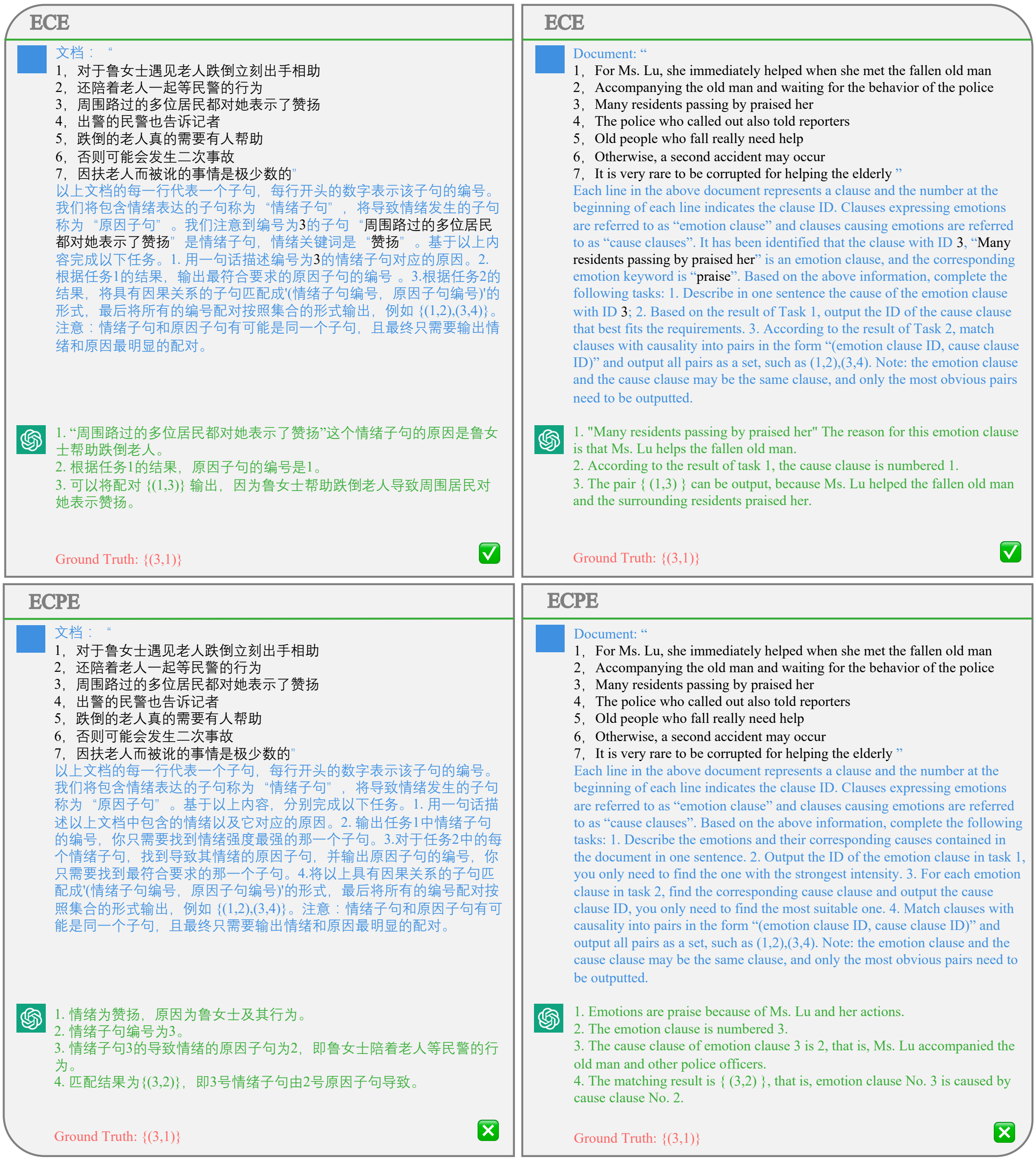
Rule#3: We also regard a prediction that paraphrases the ground truth to be correct, given the text generation paradigm.

Example#1
Input: Shipped very quickly and safely .
Output:
Aspect term: Shipping
Sentiment polarity: Positive
Ground Truth: [(Shipped, positive)]
Refined Output:
Aspect term: Shipped
Sentiment polarity: Positive
Example#2
Input: Runs real quick .
Output:
Aspect term: Speed/Performance
Sentiment polarity: Positive
Ground Truth: [(Runs, positive)]
Refined Output:
Aspect term: Runs
Sentiment polarity: Positive

Table 9: Examples on human evaluation. For simplicity, the task instruction is omitted.

Task	Dataset	Metric	Baseline	SOTA	ChatGPT	text-devinci-003
SC	SST-2	Acc	95.47	**97.50**	93.12	90.52
ABSC	14-Rest.	Acc / F1	83.94 / 75.28	**89.54 / 84.86**	83.85 / 70.57	82.19 / 71.74
ABSC	14-Lap.	Acc / F1	77.85 / 73.20	**83.70 / 80.13**	76.42 / 66.79	75.11 / 70.63
E2E-ABSA	14-Rest.	F1	77.75	**78.68**	69.14	65.06
E2E-ABSA	14-Lap.	F1	66.05	**70.32**	49.11	50.44

Table 10: Performance comparison among ChatGPT, `text-davinci-003`, fine-tuned baselines, and SOTA models on 5 datasets. Most results are derived from Table 1.



ECE

文档： “
1，对于鲁女士遇见老人跌倒立刻出手相助
2，还陪着老人一起等民警的行为
3，周围路过的多位居民都对她表示了赞扬
4，出警的民警也告诉记者
5，跌倒的老人真的需要有人帮助
6，否则可能会发生二次事故
7，因扶老人而被讹的事情是极少数的”
以上文档的每一行代表一个子句，每行开头的数字表示该子句的编号。我们将包含情绪表达的子句称为“情绪子句”，将导致情绪发生的子句称为“原因子句”。我们注意到编号为3的子句“周围路过的多位居民都对她表示了赞扬”是情绪子句，情绪关键词是“赞扬”。基于以上内容完成以下任务。1. 用一句话描述编号为3的情绪子句对应的原因。2. 根据任务1的结果，输出最符合要求的原因子句的编号。3.根据任务2的结果，将具有因果关系的子句匹配成'(情绪子句编号，原因子句编号)'的形式，最后将所有的编号配对按照集合的形式输出，例如 {(1,2),(3,4)}。注意：情绪子句和原因子句有可能是同一个子句，且最终只需要输出情绪和原因最明显的配对。

1. “周围路过的多位居民都对她表示了赞扬”这个情绪子句的原因是鲁女士帮助跌倒老人。
2. 根据任务1的结果，原因子句的编号是1。
3. 可以将配对 {(1,3)} 输出，因为鲁女士帮助跌倒老人导致周围居民对她表示赞扬。

Ground Truth: {(3,1)} ✅

ECE

Document: “
1，For Ms. Lu, she immediately helped when she met the fallen old man
2，Accompanying the old man and waiting for the behavior of the police
3，Many residents passing by praised her
4，The police who called out also told reporters
5，Old people who fall really need help
6，Otherwise, a second accident may occur
7，It is very rare to be corrupted for helping the elderly ”
Each line in the above document represents a clause and the number at the beginning of each line indicates the clause ID. Clauses expressing emotions are referred to as “emotion clause” and clauses causing emotions are referred to as “cause clauses”. It has been identified that the clause with ID 3, “Many residents passing by praised her” is an emotion clause, and the corresponding emotion keyword is “praise”. Based on the above information, complete the following tasks: 1. Describe in one sentence the cause of the emotion clause with ID 3; 2. Based on the result of Task 1, output the ID of the cause clause that best fits the requirements. 3. According to the result of Task 2, match clauses with causality into pairs in the form “(emotion clause ID, cause clause ID)” and output all pairs as a set, such as (1,2),(3,4). Note: the emotion clause and the cause clause may be the same clause, and only the most obvious pairs need to be outputted.

1. "Many residents passing by praised her" The reason for this emotion clause is that Ms. Lu helps the fallen old man.
2. According to the result of task 1, the cause clause is numbered 1.
3. The pair { (1,3) } can be output, because Ms. Lu helped the fallen old man and the surrounding residents praised her.

Ground Truth: {(3,1)} ✅

ECPE

文档： “
1，对于鲁女士遇见老人跌倒立刻出手相助
2，还陪着老人一起等民警的行为
3，周围路过的多位居民都对她表示了赞扬
4，出警的民警也告诉记者
5，跌倒的老人真的需要有人帮助
6，否则可能会发生二次事故
7，因扶老人而被讹的事情是极少数的”
以上文档的每一行代表一个子句，每行开头的数字表示该子句的编号。我们将包含情绪表达的子句称为“情绪子句”，将导致情绪发生的子句称为“原因子句”。基于以上内容，分别完成以下任务。1. 用一句话描述以上文档中包含的情绪以及它对应的原因。2. 输出任务1中情绪子句的编号，你只需要找到情绪强度最强的那一个子句。3.对于任务2中的每个情绪子句，找到导致其情绪的原因子句，并输出原因子句的编号，你只需要找到最符合要求的那一个子句。4.将以上具有因果关系的子句匹配成'(情绪子句编号，原因子句编号)'的形式，最后将所有的编号配对按照集合的形式输出，例如 {(1,2),(3,4)}。注意：情绪子句和原因子句有可能是同一个子句，且最终只需要输出情绪和原因最明显的配对。

1. 情绪为赞扬，原因为鲁女士及其行为。
2. 情绪子句编号为3。
3. 情绪子句3的导致情绪的原因子句为2，即鲁女士陪着老人等民警的行为。
4. 匹配结果为{(3,2)}，即3号情绪子句由2号原因子句导致。

Ground Truth: {(3,1)} ❎

ECPE

Document: “
1，For Ms. Lu, she immediately helped when she met the fallen old man
2，Accompanying the old man and waiting for the behavior of the police
3，Many residents passing by praised her
4，The police who called out also told reporters
5，Old people who fall really need help
6，Otherwise, a second accident may occur
7，It is very rare to be corrupted for helping the elderly ”
Each line in the above document represents a clause and the number at the beginning of each line indicates the clause ID. Clauses expressing emotions are referred to as “emotion clause” and clauses causing emotions are referred to as “cause clauses”. Based on the above information, complete the following tasks: 1. Describe the emotions and their corresponding causes contained in the document in one sentence. 2. Output the ID of the emotion clause in task 1, you only need to find the one with the strongest intensity. 3. For each emotion clause in task 2, find the corresponding cause clause and output the cause clause ID, you only need to find the most suitable one. 4. Match clauses with causality into pairs in the form “(emotion clause ID, cause clause ID)” and output all pairs as a set, such as (1,2),(3,4). Note: the emotion clause and the cause clause may be the same clause, and only the most obvious pairs need to be outputted.

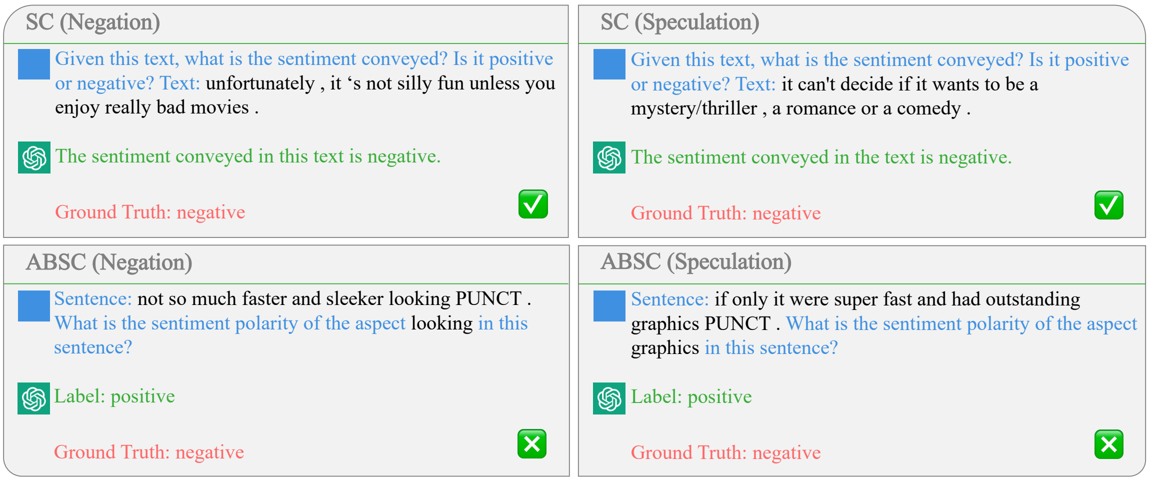
1. Emotions are praise because of Ms. Lu and her actions.
2. The emotion clause is numbered 3.
3. The cause clause of emotion clause 3 is 2, that is, Ms. Lu accompanied the old man and other police officers.
4. The matching result is { (3,2) }, that is, emotion clause No. 3 is caused by cause clause No. 2.

Ground Truth: {(3,1)} ❎

Figure 5: Case study for ChatGPT on ECE and ECPE in both Chinese (left) and English (right). The text in blue, black, green and red denote the given prompts, the examples to be evaluated, the responses of ChatGPT and the ground truths, respectively.

Model	14-Rest.		14-Lap.	
	Implicit-split	All	Implicit-split	All
Fine-tuned BERT	**65.54**	**77.16**	**69.54**	**73.45**
ChatGPT	56.31	69.72	52.68	65.92
text-devinci-003	56.85	71.09	57.17	71.09

Table 11: Evaluation results on implicit sentiment analysis among fine-tuned BERT, ChatGPT and `text-davinci-003`.



SC (Negation)

Given this text, what is the sentiment conveyed? Is it positive or negative? Text: unfortunately , it ‘s not silly fun unless you enjoy really bad movies .

The sentiment conveyed in this text is negative.

Ground Truth: negative ✓

SC (Speculation)

Given this text, what is the sentiment conveyed? Is it positive or negative? Text: it can't decide if it wants to be a mystery/thriller , a romance or a comedy .

The sentiment conveyed in the text is negative.

Ground Truth: negative ✓

ABSC (Negation)

Sentence: not so much faster and sleeker looking PUNCT . What is the sentiment polarity of the aspect looking in this sentence?

Label: positive

Ground Truth: negative ✗

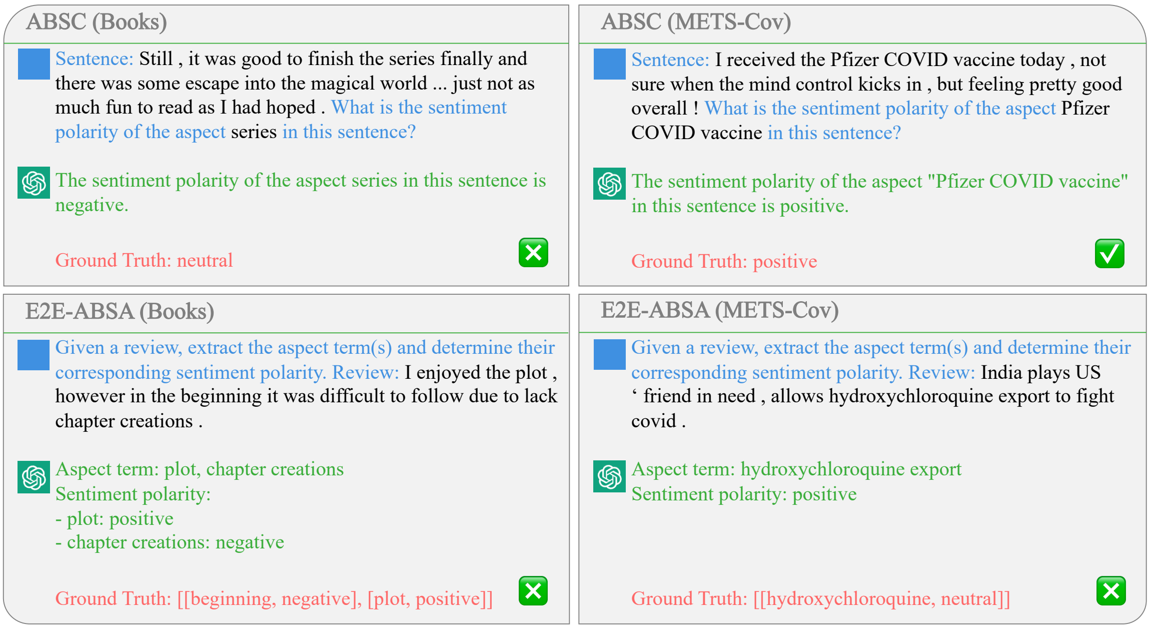
ABSC (Speculation)

Sentence: if only it were super fast and had outstanding graphics PUNCT . What is the sentiment polarity of the aspect graphics in this sentence?

Label: positive

Ground Truth: negative ✗

Figure 6: Case study for ChatGPT on SC and ABSC in case of the linguistic phenomena such as negation and speculation. The text in blue, black, green and red denote the given prompts, the examples to be evaluated, the responses of ChatGPT and the ground truths, respectively.



ABSC (Books)

Sentence: Still , it was good to finish the series finally and there was some escape into the magical world ... just not as much fun to read as I had hoped . What is the sentiment polarity of the aspect series in this sentence?

The sentiment polarity of the aspect series in this sentence is negative.

Ground Truth: neutral ✗

ABSC (METS-Cov)

Sentence: I received the Pfizer COVID vaccine today , not sure when the mind control kicks in , but feeling pretty good overall ! What is the sentiment polarity of the aspect Pfizer COVID vaccine in this sentence?

The sentiment polarity of the aspect "Pfizer COVID vaccine" in this sentence is positive.

Ground Truth: positive ✓

E2E-ABSA (Books)

Given a review, extract the aspect term(s) and determine their corresponding sentiment polarity. Review: I enjoyed the plot , however in the beginning it was difficult to follow due to lack chapter creations .

Aspect term: plot, chapter creations
Sentiment polarity:
- plot: positive
- chapter creations: negative

Ground Truth: [[beginning, negative], [plot, positive]] ✗

E2E-ABSA (METS-Cov)

Given a review, extract the aspect term(s) and determine their corresponding sentiment polarity. Review: India plays US ‘ friend in need , allows hydroxychloroquine export to fight covid .

Aspect term: hydroxychloroquine export
Sentiment polarity: positive

Ground Truth: [[hydroxychloroquine, neutral]] ✗

Figure 7: Case study for ChatGPT on ABSC and E2E-ABSA on books and medicine domains. The text in blue, black, green and red denote the given prompts, the examples to be evaluated, the responses of ChatGPT and the ground truths, respectively.