Sentiment Analysis in the Era of Large Language Models: A Reality Check

Wenxuan Zhang[*1,2] **Yue Deng**[*1,3] **Bing Liu**[4] **Sinno Jialin Pan**[3,5] **Lidong Bing**[1,2]
[1]DAMO Academy, Alibaba Group, Singapore [2]Hupan Lab, 310023, Hangzhou, China
[3]Nanyang Technological University, Singapore [4]University of Illinois at Chicago
[5]The Chinese University of Hong Kong
{saike.zwx, yue.deng, l.bing}@alibaba-inc.com
liub@uic.edu, sinnopan@cuhk.edu.hk

Abstract

Sentiment analysis (SA) has been a long-standing research area in natural language processing. With the recent advent of large language models (LLMs), there is great potential for their employment on SA problems. However, the extent to which current LLMs can be leveraged for different sentiment analysis tasks remains unclear. This paper aims to provide a comprehensive investigation into the capabilities of LLMs in performing various sentiment analysis tasks, from conventional sentiment classification to aspect-based sentiment analysis and multifaceted analysis of subjective texts. We evaluate performance across 13 tasks on 26 datasets and compare the results against small language models (SLMs) trained on domain-specific datasets. Our study reveals that while LLMs demonstrate satisfactory performance in simpler tasks, they lag behind in more complex tasks requiring a deeper understanding of specific sentiment phenomena or structured sentiment information. However, LLMs significantly outperform SLMs in few-shot learning settings, suggesting their potential when annotation resources are limited. We also highlight the limitations of current evaluation practices in assessing LLMs' SA abilities and propose a novel benchmark, SENTIEVAL, for a more comprehensive and realistic evaluation. Data and code are available at https://github.com/DAMO-NLP-SG/LLM-Sentiment.

1 Introduction

Sentiment analysis[1] (SA) has been a long-established area of research in natural language processing (NLP), which aims to study people's opinions, sentiments, emotions, etc, through computational methods (Liu, 2015; Poria et al., 2020). Since its inception (Turney, 2002; Hu and Liu, 2004), this field has attracted significant interest from both academia and industry given its wide range of applications, such as analyzing product reviews and gaining insights from social media posts (Barbieri et al., 2020; Zhang et al., 2022). Furthermore, achieving a deep understanding of human subjective feeling through sentiment analysis is undoubtedly an important step toward developing artificial general intelligence (Bubeck et al., 2023).

In recent years, large language models (LLMs) have demonstrated impressive performance on various NLP tasks (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023, inter alia). They can directly perform tasks in zero-shot or few-shot in-context learning manner and achieve strong performance without the need for any in-domain supervised training (Bang et al., 2023; Ye et al., 2023; Zhong et al., 2023; Yang et al., 2023). Although there have been some initial attempts to apply LLMs to sentiment analysis (Deng et al., 2023; Zhong et al., 2023; Wang et al., 2023), these studies are often limited to some specific tasks and adopt different models, datasets, and settings in experiments. As such, the extent to which existing large language models can be leveraged for sentiment analysis problems remains unclear.

In this work, we aim to conduct a reality check on the current state of sentiment analysis in the era of large language models. Specifically, we seek to answer the following research questions: 1) *What is the current maturity of various sentiment analysis problems?* 2) *Compared to small specialized models trained on domain-specific data, how do large models fare in both zero-shot and few-shot settings?* 3) *Are current SA evaluation practices still suitable to assess models in the era of LLMs?*

To this end, we first conduct a systematic review of various sentiment analysis-related tasks, from

[*] Equal contribution. Yue Deng is under the Joint PhD Program between DAMO Academy and Nanyang Technological University.

[1]There are many related terminologies including sentiment analysis, opinion mining, affect analysis, opinion extraction, etc. We collectively refer to them as sentiment analysis in this paper, following the convention in Liu (2015).

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3881–3906
June 16-21, 2024

conventional sentiment classification (SC, classifying the sentiment orientation of a given text (Socher et al., 2013)) to aspect-based sentiment analysis (ABSA, analyzing sentiment and opinion information at the more fine-grained aspect level (Zhang et al., 2022)) and the multifaceted analysis of subjective texts (MAST, focusing on specific sentiment or opinion phenomena such as hate speech detection and comparative opinion mining (Barbieri et al., 2020)). In total, we consider 13 sentiment analysis tasks across 26 datasets. These tasks were often studied in isolation in the past due to their unique characteristics. This fragmentation, while reasonable before, offered a somewhat incomplete understanding of how well models could comprehend human subjective information.

For LLMs, we consider both open-source models including Flan-T5 (Chung et al., 2022) and Flan-UL2 (Tay et al., 2022), along with GPT-3.5 model series, namely ChatGPT (`gpt-3.5-turbo`) and InstructGPT (`text-davinci-003`) (Brown et al., 2020; Ouyang et al., 2022). We also establish comparison baselines using smaller language models[2] (SLMs) such as T5 (Raffel et al., 2020), which allows us to measure the performance of LLMs against these specialized models trained with in-domain labeled data.

Our investigation yields several insights: Firstly, LLMs already show strong sentiment analysis ability in zero-shot settings. On some simple SA tasks such as sentiment classification, they can perform on par with SLMs trained with full training data. Secondly, when it comes to more complex tasks such as ABSA tasks that require structured sentiment information, or MAST tasks requiring a deep understanding of specific sentiment phenomena, LLMs still lag behind SLMs trained with in-domain data. Moreover, LLMs appear to be sensitive to prompt design when encountering tasks with complex input and output formats. Thirdly, with a limited quantity of annotated data under the few-shot setting, LLMs with in-context learning consistently outperform SLMs trained with the same amount of data for all types of tasks. This suggests that the application of LLMs is advantageous when annotation resources are scarce.

During the investigation, we also identify several limitations of current practice in evaluating a model's SA capability. For example, the evaluations often only involve specific tasks or datasets; and inconsistent prompts are utilized across different studies to evaluate models. While these evaluation practices might have been appropriate in the past, they fall short of accurately assessing LLMs' SA abilities. To address these issues, we propose a novel benchmark called SentiEval. It breaks the boundary of a wide range of SA tasks, enabling a more comprehensive evaluation of models. It also employs varied task instructions, paired with the corresponding text, alleviating the sensitivities associated with prompt design during the evaluation of different LLMs. Furthermore, by framing these tasks as natural language instructions, we create a more realistic evaluation environment akin to a real-world practical use case.

2 Background

Sentiment Analysis SA has received lots of attention since its early appearance (Turney, 2002; Yu and Hatzivassiloglou, 2003; Hu and Liu, 2004) and remained an active research area in the field of NLP nowadays (Liu, 2015; Poria et al., 2020; Yadav and Vishwakarma, 2020). Such enduring interest stems from both the importance of comprehending human subjective sentiments and opinions toward achieving human-level intelligence (Bubeck et al., 2023), and its broad practical applications, such as analyzing customer reviews (Keung et al., 2020; Zhang et al., 2022) and digesting social media opinions (Yue et al., 2019; Barbieri et al., 2020). SA comprises a broad spectrum of tasks, from sentiment classification that determines the overall sentiment polarity of a given text (Turney, 2002), to aspect-based sentiment analysis (ABSA) (Hu and Liu, 2004; Zhang et al., 2022) and multifaceted analysis of subjective texts (MAST) (Liu, 2015) in recent years. All these tasks collectively contribute to a holistic understanding of sentiment in language and demonstrate the wide range of tasks falling under the umbrella of sentiment analysis.

Large Language Models (LLMs) Recently, there has been a remarkable advancement in the development of LLMs, such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), Flan-UL2 (Tay et al., 2022), LLaMA (Touvron et al., 2023) and ChatGPT. There are some initial attempts on evaluating LLMs for SA tasks. Zhong et al. (2023) observe that the zero-shot performance of LLMs is comparable to fine-tuned BERT model. Wang

[2] So far, there is no clear definition of what models can be counted as small or large language models. In this work, we consider model parameters less than 3B as small, and larger than 3B as large for simplified demonstration.

et al. (2023) conduct a preliminary study with ChatGPT for some SA tasks, specifically investigating its ability to handle polarity shifts, open-domain scenarios, and sentiment inference problems. In addition, Zhao et al. (2023) focus on ChatGPT's emotional conversation capability and indicate it exhibits promising results in generating emotional responses. Moreover, Deng et al. (2023) explore the fine-tuning of a small student model with an LLM to generate weak labels, and the final model performs on par with existing supervised models. Despite those existing efforts, their scope is often limited to specific tasks and involves different datasets and experimental designs. The true capacity of LLMs for SA remains unclear.

3 Investigated Tasks and Datasets

We conduct an extensive survey of a wide range of SA tasks and categorize different tasks into three types: sentiment classification (SC), aspect-based sentiment analysis (ABSA), and multifaceted analysis of subjective texts (MAST). We briefly describe investigated tasks of each type, along with the datasets and evaluation metrics in this section. The detailed descriptions are in Appendix A.1. For each dataset, we sample a maximum of 500 examples from its original test set, to ensure balance across various tasks and datasets.

3.1 Sentiment Classification

Sentiment classification (SC) aims at assigning predefined sentiment classes (e.g., positive, negative, or neutral) to given texts (Liu, 2015). Depending on the level of granularity at which sentiment can be analyzed, SC can be further categorized into three tasks, including document-level, sentence-level, and aspect-level SC. For document-level SC, we take three widely used datasets, including IMDb (Maas et al., 2011), Yelp-2, and Yelp-5 (Zhang et al., 2015), which contain movie reviews and business reviews respectively. For sentence-level SC, we select multiple datasets for evaluation, including MR (Pang and Lee, 2005), SST2, SST5 (Socher et al., 2013), and Twitter (Rosenthal et al., 2017), covering different types of opinionated texts. Aspect-level SC focuses on identifying sentiment towards specific aspects or entities mentioned. There are two widely used datasets including Lap14 and Rest14 (Pontiki et al., 2014) which consist of laptop and restaurant reviews.

These datasets involve a varying number of sentiment classes. We take accuracy scores as the evaluation metric for these SC tasks.

3.2 Aspect-based Sentiment Analysis

Aspect-based sentiment analysis (ABSA) refers to the process of analyzing people's sentiments at a more fine-grained aspect level. It encompasses the analysis of various sentiment elements, such as aspect terms, aspect categories, opinions, and sentiment polarities (Zhang et al., 2022).

We focus on three compound ABSA tasks here for investigation, which aim to jointly extract multiple sentiment elements: (1) Unified Aspect-based Sentiment Analysis (UABSA) is the task of extracting both the aspect and its corresponding sentiment polarity simultaneously. We evaluate UABSA on four datasets originally from SemEval-2014 (Pontiki et al., 2014), SemEval-2015 (Pontiki et al., 2015), and SemEval-2016 (Pontiki et al., 2016) shared tasks. (2) Aspect Sentiment Triplet Extraction (ASTE) further extracts the opinion terms on the basis of the UABSA task, which provides an explanation for the predicted sentiment on certain aspects. The datasets we utilized were introduced by Xu et al. (2020), which were built upon the four UABSA datasets. (3) Aspect Sentiment Quadruple Prediction (ASQP) task (Zhang et al., 2021; Cai et al., 2021) was introduced to provide a complete aspect-level sentiment structure, namely (category, aspect, opinion, sentiment) quadruple. Two restaurant datasets are used for the ASQP task.

Following previous studies, we use the Micro-F1 score as the metric for evaluation. A predicted tuple would be counted as correct only if all sentiment elements match exactly with the gold labels.

3.3 Multifaceted Analysis of Subjective Text

Multifaceted analysis of subjective text (MAST) are tasks that involve different aspects of human subjective feeling reflected in the text (Liu, 2015; Poria et al., 2020). These tasks expand SA beyond merely identifying positive or negative feelings but focus on recognizing and understanding a broader range of human emotional states.

We adopt multiple datasets for investigation, including: (1) Implicit sentiment analysis (Li et al., 2021); (2) SemEval2019 HatEval challenge (Basile et al., 2019) for hate speech detection; (3) Subtask 3A of the SemEval2018 (Hee et al., 2018) for irony detection; (4) SemEval2019 OffensEval dataset (Zampieri et al., 2019) for offensive language identification; (5) SemEval2016 shared task

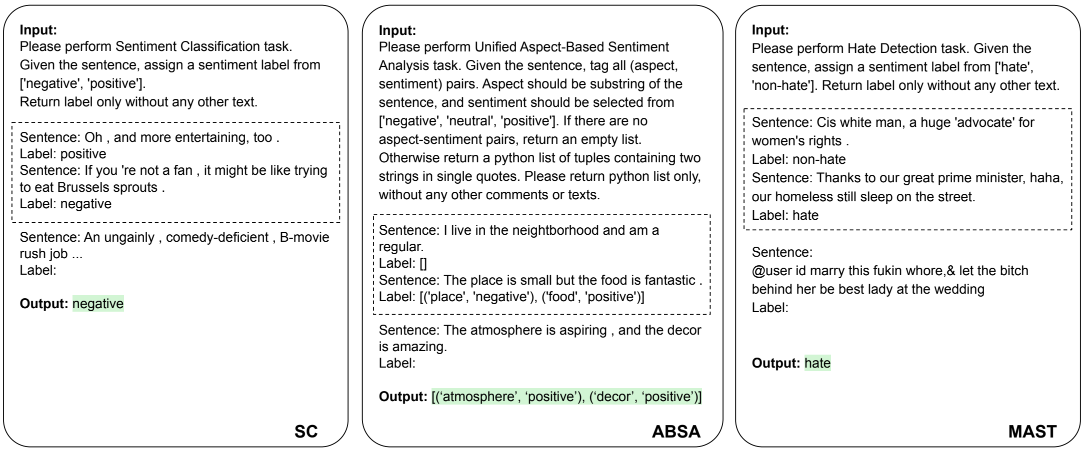

Figure 1: Prompt examples for SC, ABSA, and MAST respectively. The text inside the dashed box are demonstrations of the few-shot setting and would be removed under the zero-shot setting.

on Detection Stance in Tweets (Mohammad et al., 2016) for stance detection task; (6) CS19 dataset (Panchenko et al., 2019) for comparative opinion mining task; (7) TweetEval benchmark (Barbieri et al., 2020) for emotion recognition task.

For the evaluation, we follow previous studies to utilize the most common metrics for each task respectively. Details are given in Appendix A.1 and metrics for each task are summarized in Table 4.

4 Evaluation Setup

4.1 Models

Large Language Models (LLMs) We adopt two models from the Flan model family since they are open-sourced and showed strong zero-shot and few-shot performance, namely `Flan-T5` (XXL version, 13B) (Chung et al., 2022) and `Flan-UL2` (20B) (Tay et al., 2022). We use their checkpoints hosted on Huggingface for the inference. We also take two models from OpenAI, including `ChatGPT` (`gpt-3.5-turbo`[3]) and the text-davinci-003 model (`text-003`, 175B) of the GPT-3.5 family.

Small Language Models (SLMs) For SLMs, we take `T5` (large version, 770M) (Raffel et al., 2020), which shows great performance in tackling multiple SA tasks in a unified text-to-text format. This allows us to utilize a single, consistent SLM for all SA tasks without task-specific designs, enabling us to make a coherent and relatively fair comparison with LLMs. We train the `T5` model with domain-specific data on each dataset, with either the full training set (statistics detailed in Table 4) or sampled data in the few-shot setting. We use the Adam optimizer with a learning rate of 1e-4 and a fixed batch size of 4 for all tasks. We set 3 epochs for the full training setting and 100 epochs for the few-shot training setting. We conduct three runs with different random seeds for SLMs in both settings and report the average results for more stable comparisons.

4.2 Prompting Strategy

LLMs may produce very different responses even when the prompts are semantically similar (Perez et al., 2021; Lu et al., 2022). Furthermore, the preference for prompts varies from one LLM to another. Therefore, we aim to provide relatively consistent prompts for all datasets across different models in this study, rather than specific designs, in order to evaluate the general performance of LLMs. Our goal is to design prompts that are simple, clear, and straightforward.

As shown in Figure 1, we include only essential components in the prompt, namely the task name, task definition, and output format. The task name mentions the name of a specific task. The task definition is constructed based on each task's definition and annotation guidelines and also incorporates the label space as a set of options for the model to output its response. The output format defines the expected structure of the output, enabling us to decode the model's responses into our desired format. For few-shot learning, an additional "demonstration" part is added (contents in the dashed boxes). This includes k examples for each class, each accompanied by their respective gold labels in the desired format. For more detailed information and examples, please refer to Appendix A.6.

[3]May 12 version of ChatGPT is used for the experiments.

Task	Dataset	Baseline		LLM				SLM
		random -	majority -	Flan-T5 (11B)	Flan-UL2 (20B)	text-003 (175B)	ChatGPT (NA)	T5$_{large}$ (770M)
Sentiment Classification (SC)								
Document-Level	IMDb	52.40	46.80	86.60	**97.40**	90.60	94.20	93.93
	Yelp-2	52.80	48.00	92.20	**98.20**	93.20	97.80	96.33
	Yelp-5	19.80	18.60	34.60	51.60	48.60	52.40	**65.60**
Sentence-Level	MR	47.40	49.60	66.00	**92.20**	86.80	89.20	90.00
	SST2	49.20	48.60	72.00	**96.40**	92.80	93.60	93.20
	Twitter	34.20	45.40	43.60	47.40	59.40	**69.40**	67.73
	SST5	21.40	22.20	15.00	**57.00**	45.20	48.00	56.80
Aspect-Level	Lap14	34.80	53.80	69.00	73.20	74.60	76.80	**78.60**
	Rest14	34.00	65.60	80.80	82.40	80.00	82.80	**83.67**
Average		38.44	44.29	62.20	77.31	74.58	78.24	**80.65**
Aspect-Based Sentiment Analysis (ABSA)								
UABSA	Rest14	NA	NA	0.00	0.00	47.56	54.46	**75.31**
	Rest15	NA	NA	0.00	0.00	35.63	40.03	**65.46**
	Rest16	NA	NA	0.00	0.00	40.85	49.61	**73.23**
	Laptop14	NA	NA	0.00	0.00	28.63	33.14	**62.35**
ASTE	Rest14	NA	NA	0.00	0.00	41.43	40.04	**65.20**
	Rest15	NA	NA	0.00	0.00	37.53	33.51	**57.78**
	Rest16	NA	NA	0.00	0.00	41.03	42.18	**65.94**
	Laptop14	NA	NA	0.00	0.00	27.05	27.30	**53.69**
ASQP	Rest15	NA	NA	0.00	0.00	13.73	10.46	**41.08**
	Rest15	NA	NA	0.00	0.00	18.18	14.02	**50.58**
Average		NA	NA	0.00	0.00	33.16	34.47	**61.06**
Multifaceted Analysis of Subjective Text (MAST)								
Implicit	Lap+Res	35.75	56.11	33.03	42.53	45.25	54.98	**67.12**
Hate	HatEval	48.00	36.31	56.09	**70.80**	67.79	50.92	46.94
Irony	Irony18	50.96	58.96	27.31	73.84	76.61	68.66	**79.44**
Offensive	OffensEval	46.67	41.86	32.78	74.44	73.31	64.88	**80.76**
Stance	Stance16	33.94	35.82	20.74	61.10	39.96	50.25	**67.33**
Comparative	CS19	49.36	73.89	54.46	85.67	74.52	75.80	**89.49**
Emotion	Emotion20	22.87	13.92	44.34	69.92	70.51	72.80	**80.35**
Average		41.08	45.27	38.39	68.33	63.99	62.61	**73.05**

Table 1: Zero-shot performance of various sentiment analysis tasks. The best results on each dataset are in bold. Similar to GLUE (Wang et al., 2019), "Average" rows show the average of all dataset-specific metrics. We present the full training set fine-tuned SLM performance as a reference.

5 Evaluation Results and Analysis

5.1 Zero-shot Results

We summarize the zero-shot performance of various LLMs in Table 1. Two baselines are further included for better comparisons: `random` assigns a random label to each sample, and `majority` takes the most common label from the training set's label distribution as the prediction. For SLMs, we report the performance by employing the complete training set to train the model before proceeding to conduct inference on the same test set. The following observations can be made.

LLMs such as ChatGPT demonstrate strong zero-shot performance. As can be observed in the top and bottom parts of Table 1, LLMs have demonstrated a strong ability to tackle simple SC tasks such as binary sentiment classification and MAST tasks without any prior in-domain training. For example, `ChatGPT` achieves comparable results to the `T5` model, which has been specifically fine-tuned with the full training set for each dataset. On average, `ChatGPT`'s performance reaches 97% of the `T5`'s prediction on SC tasks, and 85% on MAST tasks, respectively. Moreover, `Flan-UL2`, despite not being the largest model, is able to achieve comparable, and in some cases, superior performance to larger models like `text-003` across multiple tasks, possibly due to the advantage of both reasonable model size and large-scale instruction tuning. Overall, these results suggest a superior sentiment analysis ability already inherent in these models.

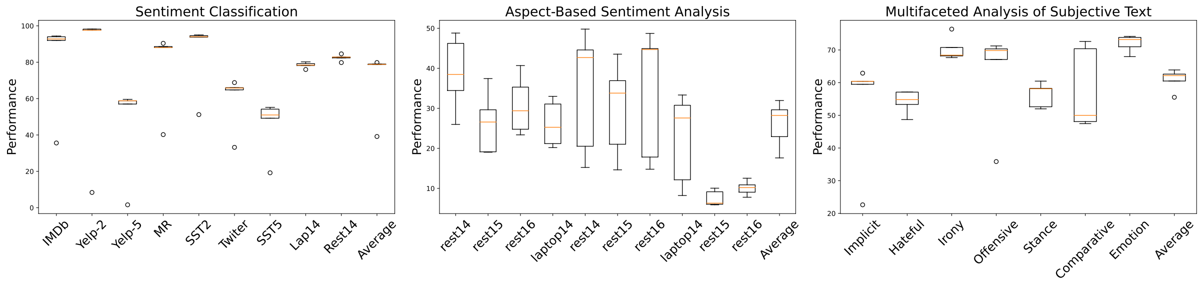

Figure 2: Sensitivity of different prompt designs on three types of SA tasks. The performance variance of each dataset is from five different prompts given by GPT-4. The circles depicted in the figure represent outlier data points.

LLMs still struggle with extracting fine-grained structured sentiment information or tasks requiring a deep understanding of specific sentiment phenomena. While LLMs have shown proficiency in many SA tasks, they fall short when it comes to extracting structured and fine-grained sentiment and opinion information. For instance, `Flan-T5` and `Flan-UL2` were unable to achieve any notable performance on any ABSA tasks across all datasets, as can be noted from the middle part of Table 1. Although they have gone through instruction tuning, they can hardly follow the format required in the instructions and generate meaningless predictions. `text-003` and `ChatGPT` provide better results but were still significantly outperformed by fine-tuned smaller language models. For example, `text-003` reaches only around 54% of the performance of a fine-tuned `T5` model on ABSA tasks, though being more than 200 times larger. Similarly, for more complicated MAST tasks, it also lags behind the fine-tuned `T5` models, e.g., 45.25% v.s. 67.12% accuracy scores on the implicit sentiment analysis task.

Some SA tasks have reached certain maturity Overall, we can see that satisfactory performance of some SA tasks such as binary sentiment classification (e.g., IMDb, Yelp-2, MR, SST2) or simple MAST tasks (e.g., emotion recognition), can be achieved with either LLMs under a zero-shot setting or SLMs trained with in-domain labeled dataset. This observation implies that these SA tasks have reached a level of maturity and can be considered as effectively solved, thereby shifting the focus in the field toward addressing more complex challenges that LLMs still struggle with.

5.2 Analysis of Sensitivity on Prompt Design

The design of suitable prompts is critical when leveraging large language models for specific tasks. Different prompt designs have been shown to even lead to large performance variance in some tasks (Perez et al., 2021; Lu et al., 2022). To investigate the impact of such sensitivity on SA tasks, we further construct an additional five prompts for each task, then conduct experiments with `ChatGPT` to evaluate the variations in performance. We take GPT-4 (OpenAI, 2023) for such prompt generation, which has shown to be effective to generate prompts or instruction-following data (Peng et al., 2023).[4] This can also alleviate the potential bias of manually written prompts. Details of such prompt generation are given in Appendix A.2.

The results of `ChatGPT` with the five different prompts are depicted in Figure 2, in the format of the boxplot. It can be noticed that the impact of different prompts on performance varies from task to task. For SC tasks, the choice of prompt appears to have less effect, e.g., the boxes in the top figure are usually quite concentrated. However, for tasks necessitating structured, fine-grained output, the performance can vary significantly depending on the design of the prompt, as illustrated in the middle figure for ABSA tasks. Interestingly, despite the simplicity of SC tasks, the model still demonstrates sensitivity to certain prompts, with noticeable outliers for some SC datasets (i.e., circles in the figure). With a detailed investigation, we find models tend to be sensitive to certain words, e.g., "*analyze*", where it may generate long explanations even explicitly instructed not to do so.

5.3 Few-shot Results

We also conduct few-shot experiments to assess whether LLMs or SLMs perform better when only a limited number of examples for a sentiment analysis task are available. We consider three K-shot settings: 1-shot, 5-shot, and 10-shot. For each set-

[4]We also conduct preliminary experiments with `ChatGPT`, however, it struggles to understand such complicated instructions, thus failing to produce satisfactory prompts.

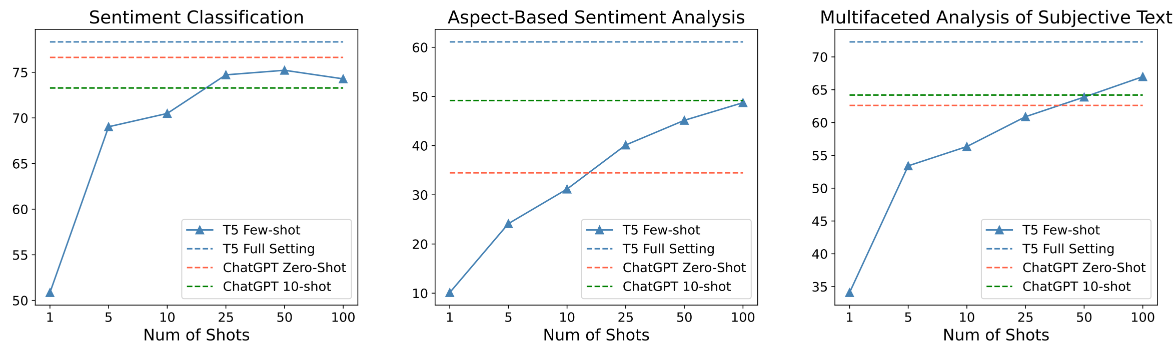

Figure 3: Averaged few-shot results on all datasets for each task type with an increasing number of different shots. Results of ChatGPT zero-shot and T5 full setting are also shown for easy comparison.

Task	1-shot		5-shot		10-shot	
	ChatGPT	T5	ChatGPT	T5	ChatGPT	T5
Doc-SC	81.47	66.76	NA	75.64	NA	77.76
Sent-SC	76.20	46.80	75.20	67.32	72.20	69.52
Aspect-SC	81.57	58.97	75.57	72.47	75.43	72.43
UABSA	52.57	15.70	53.75	29.71	55.02	39.51
ASTE	44.45	6.81	48.65	23.60	50.14	29.89
ASQP	31.07	5.61	34.61	14.08	35.54	17.05
MAST	68.46	34.09	66.21	53.40	64.19	56.34

Table 2: Few-shot performance of various sentiment analysis tasks. All the results are reported with average scores in 3 runs. "NA" denotes infeasible experiments due to limited sequence length.

ting, we sample K examples for each sentiment type (with the exception of the ASQP task, where we sample K examples for each aspect category). These sampled examples serve as in-context learning samples for LLMs and training data for SLMs. The results of these experiments are summarized in Table 2. More detailed results as well as the standard deviation are provided in Table 6.

We can see that LLMs surpass SLMs under varied few-shot settings. Across all three few-shot settings, LLMs consistently outperform SLMs such as T5 in almost all cases. This advantage becomes more obvious for three ABSA tasks, which require the model to output structured sentiment information. SLMs significantly lag behind LLMs under such requirements, possibly due to the difficulty of learning such patterns with limited data. To delve deeper into their respective strengths and limitations, we gradually increase the value of K in the few-shot settings[5], and present the results for T5 in Figure 3. It becomes apparent that even with a 10-shot setting, ChatGPT sets a robust baseline that requires T5 to utilize nearly five to ten times (i.e., 50-shot or 100-shot) more data to achieve comparable performance.

In addition, Table 2 demonstrates that as the number of shots increases, SLMs consistently exhibit substantial improvements in various SA tasks. However, the impact of increasing shots on LLMs' performance varies from task to task. For relatively easier tasks like SC, the incremental benefit of additional shots for LLMs is less obvious. While for ABSA tasks, which demand a deeper understanding and precise output format, increasing the number of shots greatly boosts LLM performance. Moreover, including additional examples for MAST tasks can even lead to a decrease in performance, possibly due to biases introduced by the demonstration examples. This suggests that the utility of extra examples is not a silver bullet for all tasks but varies depending on the complexity of the task.

6 SentiEval Benchmark

6.1 Rethinking SA Capability Evaluation

We have conducted extensive experiments to evaluate LLMs' SA capability in the above sections, where we notice some common flaws regarding the current evaluation practice

Call for more comprehensive evaluation Most of the current evaluations tend to focus narrowly on specific SA tasks or datasets (Zhong et al., 2023; Wang et al., 2023). While these assessments can provide useful insights into certain aspects of an LLM's sentiment analysis competence, they inherently fall short of capturing the full breadth and depth of the model's capabilities. Such limitation not only reduces the overall reliability of the assessment results but also limits the scope of understanding the model's adaptability to diverse SA

[5]We only report results for SLMs here, as LLMs frequently encounter a context length limit, making them unsuitable for larger K values without specific handling.

scenarios. For example, a model with satisfactory sentiment classification ability does not guarantee its performance in detecting hateful speech.

Appeal for natural ways to interact with models Conventional sentiment analysis tasks are often structured as a single sentence paired with its corresponding sentiment label. This format, while facilitating the learning of the mapping relationship between the text and its sentiment label, may not optimally suit LLMs, which are typically text-generation models. In practice, users exhibit varied writing styles, leading to diverse ways of communicating their requirements to LLMs to solve their SA tasks. It is thus critical to account for these diverse expressions in the evaluation process to reflect more realistic use cases.

Sensitivity on Prompt Design As shown in Sec 5.2, variations in prompt design can substantially influence the performance of `ChatGPT`, even on some seemingly simple sentiment classification tasks. Such nuanced sensitivity associated with prompt design introduces challenges when attempting to fairly and stably test the SA capabilities of LLMs. This challenge is further amplified when various studies employ distinct prompts for different SA tasks across a range of LLMs. The inherent bias associated with prompt design complicates the fair comparison of different models using the same prompt, as a single prompt may not be universally appropriate to reflect all models' capabilities.

6.2 SENTIEVAL: Construction

To mitigate the limitations when assessing models' SA capability discussed above, we propose a new benchmark named SENTIEVAL for better **senti**ment analysis **eval**uation in the era of LLMs.

The main idea of SENTIEVAL is to: 1) break the boundary between individual sentiment analysis tasks to establish a unified testing benchmark, providing a more comprehensive assessment of a model's sentiment analysis proficiency, rather than emphasizing on specific aspects; 2) test the model using natural language instructions presented in various styles. This mimics the real use case when humans interact with the model with natural languages for solving SA tasks, instead of purely learning text-label mapping; 3) equip the benchmark with diverse but fixed instructions, making performance comparisons more stable and reliable across different LLMs and studies. By setting a consistent benchmark, it allows for an equitable comparison that is less subject to prompt variation.

	`Flan-T5`	`Flan-UL2`	`text-003`	`ChatGPT`
SENTIEVAL	29.07	38.82	36.64	**47.55**
SC	54.22	63.13	60.11	**72.73**
ABSA	0.00	0.09	11.66	**14.77**
MAST	34.21	**58.35**	38.48	57.71

Table 3: Results on SENTIEVAL benchmark of different LLMs, measured by the exact match with the label.

Specifically, besides the five prompts generated by GPT-4 in Sec 5.2, we further manually write five additional prompts for each task. Therefore, each task will have ten candidate prompts in total. Then for each data sample of all tasks, we randomly select one prompt and combine it with the text to form a complete query for the model. Additionally, we also randomly decide (with a 50% percent chance) whether to put few-shot examples with the current prompt. In the end, SENTIEVAL contains 12,224 data samples, each containing the original text, the instruction for a specific task, and optional few-shot examples.

6.3 SENTIEVAL: Re-evaluate

After constructing the SENTIEVAL benchmark, we revisit the evaluation of the various LLMs outlined in Sec 4.1 against this benchmark. We report the results in Table 3, which are the exact match scores between the labels and predictions. Although the new benchmark does not treat each task separately, we further report the results of different task types for investigations.

From Table 3, we can see noticeable differences in the relative performance of various models. For example, `Flan-UL2` achieves comparable performance with `ChatGPT` on SC tasks in Table 1, but there is a large gap in Table 3. A potential explanation for this discrepancy is that SENTIEVAL requires the model to comprehend diverse styles of instructions (i.e., varying prompt designs) for optimal performance, where `ChatGPT` exhibits greater robustness. Additionally, it demands the model's compliance with the required format, or adaptation to the pattern set by few-shot examples, thus posing greater challenges. We can see `ChatGPT` sets a strong performance baseline, showing its strong SA capability and instruction-following ability. Overall, there is much room for improvement on this benchmark in the future, especially for more complicated tasks such as ABSA and MAST tasks.

7 Conclusions

In this study, we conduct a systematic evaluation of various sentiment analysis tasks using LLMs, which helps better understand their capabilities in sentiment analysis problems. Experimental results reveal that while LLMs perform quite well on simpler tasks in a zero-shot setting, they struggle with more complex tasks. In a few-shot learning context, LLMs consistently outperform SLMs, suggesting their potential in scenarios where annotation resources are scarce. This work also highlights the limitations of current evaluation practices and then introduces the SENTIEVAL benchmark as a more comprehensive and realistic evaluation tool.

Limitations

In this study, our objective is to conduct a comprehensive evaluation of large language models' capabilities in performing diverse sentiment analysis tasks. We have selected 13 tasks encompassing 26 datasets for this purpose. However, this selection does not represent an exhaustive enumeration of all sentiment analysis-related tasks. Including a broader range of tasks focusing on different sentiment aspects or in different formats would further show the strengths and limitations of LLMs.

Regarding the language, all the datasets included in our investigation are in English. It is worth mentioning that sentiment phenomena are often closely related to the language in which they are expressed, and even to the cultural background. Consequently, extending such investigations to other languages or multilingual settings would yield a more comprehensive understanding of LLMs' performance in sentiment analysis tasks across diverse linguistic and cultural contexts.

Acknowledgements

This work was substantially supported by DAMO Academy through DAMO Academy Research Intern Program. The work of Bing Liu was supported in part by three NSF grants (IIS-1910424, IIS-1838770, and CNS-2225427).

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.

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, pages 1644–1650.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 54–63. Association for Computational Linguistics.

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*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 340–350. Association for Computational Linguistics.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. Llms to the moon? reddit market sentiment analysis with large language models. In *Companion Proceedings of the ACM Web Conference 2023, WWW 2023*, pages 1014–1019.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 39–50. Association for Computational Linguistics.

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*, pages 168–177.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 4563–4568. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 246–256.

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

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 8086–8098. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 1–17. Association for Computational Linguistics.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 31–41. The Association for Computer Linguistics.

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

Long Ouyang, Jeffrey 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. In *NeurIPS*.

Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing comparative sentences. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*, pages 136–145.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 11054–11070.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.*

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A review on offensive language detection. *Advances in Data and Information Sciences: Proceedings of ICDIS 2019*, pages 433–439.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon G. Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Soc. Netw. Anal. Min.*, 8(1):28:1–28:26.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017*, pages 1–10. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. 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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *CoRR*, abs/2205.05131.

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.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*, pages 417–424.

Kasturi Dewi Varathan, Anastasia Giachanou, and Fabio Crestani. 2017. Comparative opinion mining: A review. *J. Assoc. Inf. Sci. Technol.*, 68(4):811–829.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? A preliminary study. *CoRR*, abs/2304.04339.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349.

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

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *CoRR*, abs/2304.13712.

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.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2003, Sapporo, Japan, July 11-12, 2003*.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowl. Inf. Syst.*, 60(2):617–663.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 75–86. Association for Computational Linguistics.

Qingcheng Zeng and An-Ran Li. 2022. A survey in automatic irony processing: Linguistic, cognitive, and multi-x perspectives. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 824–836. International Committee on Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *CoRR*, abs/2203.01054.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *CoRR*, abs/2304.09582.

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.

A appendix

A.1 Details on Investigated Tasks and Datasets

We conduct an extensive survey of a wide range of SA tasks and categorize different tasks into three types: sentiment classification (SC), aspect-based sentiment analysis (ABSA), and multifaceted analysis of subjective texts (MAST). We describe investigated tasks of each type, along with the datasets and evaluation metrics. To ensure balance across various tasks and datasets, we limit our evaluation by sampling a maximum of 500 examples from the test set of each dataset. Detailed statistics on each task and dataset are summarized in Table 4.

A.1.1 Sentiment Classification

Sentiment classification (SC) aims at assigning predefined sentiment classes (e.g., positive, negative, or neutral) to given texts (Liu, 2015). It serves as a fundamental measure of sentiment orientation and is commonly used to analyze customer reviews, social media posts and etc. It can involve a varying number of sentiment classes, ranging from binary classification, where sentiments are categorized as either positive or negative, to more nuanced five-class classification, which grades sentiments on a scale from very negative to very positive. There are also different levels of granularity at which sentiment can be analyzed, including document-level, sentence-level, and aspect-level SC.

Document-Level Sentiment classification at the document level aims to determine the overall sentiment expressed in a text corpus, providing a high-level understanding of the expressed sentiment orientation. We evaluate on three widely used datasets, including IMDb (Maas et al., 2011), Yelp-2, and Yelp-5 (Zhang et al., 2015). The IMDb dataset contains movie reviews, whereas the Yelp-2 dataset includes customer reviews for businesses. Reviews of both datasets are labeled as either *positive* or *negative*. However, the Yelp-5 dataset offers a more fine-grained sentiment classification by introducing three additional sentiment classes: *very positive*, *very negative*, and *neutral*. We employ accuracy as the evaluation metric.

Sentence-Level Sentence-level classification allows for sentiment analysis on a sentence-by-sentence basis. It is particularly useful in analyzing social media posts, customer feedback, or any text where sentiments may change rapidly from sentence to sentence. We select multiple datasets for evaluation, including MR (Pang and Lee, 2005), SST2, SST5 (Socher et al., 2013), and Twitter (Rosenthal et al., 2017). The MR, SST2, and SST5 datasets contain movie reviews, whereas the Twitter dataset consists of social media posts. While the SST2 and MR datasets use binary sentiment

Task	Dataset	train	dev	test	sampled test	class*	metric
Sentiment Classification (SC)							
Document-Level	IMDb	22,500	2,500	25,000	500	2	accuracy
	Yelp-2	504,000	56,000	38,000	500	2	accuracy
	Yelp-5	585,000	65,000	50,000	500	5	accuracy
Sentence-Level	MR	8,534	1,078	1,050	500	2	accuracy
	SST-2	6,920	872	1,821	500	2	accuracy
	Twitter	45,615	2,000	12,284	500	3	accuracy
	SST-5	8,544	1,101	2,210	500	5	accuracy
Aspect-Level	lap14	2,282	283	632	500	3	accuracy
	rest14	3,608	454	1,119	500	3	accuracy
Aspect-based Sentiment Analysis (ABSA)							
UABSA	Rest14	2,736	304	800	500	3	micro_f1
	Rest15	1,183	130	685	500	3	micro_f1
	Rest16	1,799	200	676	500	3	micro_f1
	Laptop14	2,741	304	800	500	3	micro_f1
ASTE	Rest14	1,266	310	492	492	3	micro_f1
	Rest15	605	148	322	322	3	micro_f1
	Rest16	857	210	326	326	3	micro_f1
	Laptop14	906	219	328	328	3	micro_f1
ASQP	Rest15	834	209	537	500	13	micro_f1
	Rest16	1,264	316	544	500	13	micro_f1
Multifaceted Analysis of Subjective Text (MAST)							
Implicit	Lap+Res	1,746	NA	442	442	3	accuracy
Hate	HatEval	9,000	1,000	2,970	500	2	macro_f1
Irony	Irony18	2,862	955	784	500	2	f1(irony)
Offensive	OffensEval	11,916	1,324	860	500	2	macro_f1
Stance	Stance16	2,620	294	1,249	500	3	macro_f1†
Comparative	CS19	1,094	157	314	314	2	accuracy
Emotion	Emotion20	3,257	374	1,421	500	4	macro_f1

Table 4: Investigated tasks and dataset statistics. $*$ represents the number of sentiment classes among each task, except for the two datasets of ASQP, which represent the number of aspect categories. † denotes the macro_f1 score without none class.

labels, Twitter's sentiment analysis introduces an additional *neutral* class. In addition, SST5 provides a wider range of labels including *very positive*, *positive*, *neutral*, *negative*, and *very negative* sentiments. To evaluate the performance on these datasets, we use accuracy as a metric.

Aspect-Level Since sentiment expressed towards different targets might be different even within a single sentence, aspect sentiment classification dives even deeper into the analysis by focusing on identifying sentiment towards specific aspects or entities mentioned. This level of analysis is particularly valuable when the sentiment towards different aspects or entities needs to be assessed individually. There are two widely used datasets including Lap14 and Rest14. These datasets were introduced in the SemEval ABSA challenge 2014 (Pontiki et al., 2014) and consist of laptop and restaurant reviews, respectively. The goal is to determine the sentiment towards a specific aspect mentioned in a review sentence, classifying it as either *positive*, *negative*, or *neutral*. Performance assessment is based on the metric of accuracy.

A.1.2 Aspect-based Sentiment Analysis

Aspect-based sentiment analysis (ABSA) refers to the process of analyzing people's sentiments at a more fine-grained aspect level. It encompasses the analysis of various sentiment elements, such as aspects, opinions, and sentiment polarities (Zhang et al., 2022). ABSA has gained significant attention in recent years, resulting in the emergence of a wide range of tasks. We focus on three compound ABSA tasks here for investigation, which aim to jointly extract multiple sentiment elements.

Unified Aspect-based Sentiment Analysis (UABSA) UABSA is the task of extracting both the aspect and its corresponding sentiment polarity simultaneously. We evaluate UABSA on four datasets originally from SemEval-2014 (Pontiki et al., 2014), SemEval-2015 (Pontiki et al., 2015), and SemEval-2016 (Pontiki et al., 2016) shared tasks, which consist of reviews from Laptops and Restaurants domains. Following previous studies, we use Micro-F1 score as the metric for evaluation. A predicted pair would be counted as correct only if both the aspect term and sentiment polarity match exactly with the gold labels.

Aspect Sentiment Triplet Extraction (ASTE) The ASTE task further extracts the opinion terms on the basis of the UABSA task, which provides an explanation for the predicted sentiment on certain aspects. Therefore, the final target of ASTE is to extract the (aspect, opinion, and sentiment) triplet for a given text. The datasets we utilized were introduced by Xu et al. (2020), which were built upon four UABSA datasets. Likewise, we employ the Micro-F1 metric and consider an exact match prediction of each triplet as correct.

Aspect Sentiment Quadruple Prediction (ASQP) ASQP task was introduced to provide a complete aspect-level sentiment structure, namely (category, aspect, opinion, sentiment) quadruple (Zhang et al., 2021; Cai et al., 2021). By introducing an additional aspect category element, it can still provide useful information when the aspect term is not explicitly mentioned. Our study utilizes two restaurant datasets from Zhang et al. (2021). We adopt the same evaluation metric and standardization with UABSA and ASTE, using Micro-F1 score as the evaluation metric.

A.1.3 Multifaceted Analysis of Subjective Text

Multifaceted analysis of subjective text (MAST) are tasks that involve different aspects of human subjective feeling reflected in the text (Liu, 2015; Poria et al., 2020). These tasks expand SA beyond merely identifying positive or negative feelings but focus on recognizing and understanding a broader range of human emotional states.

Implicit Sentiment Analysis Implicit sentiment analysis focuses on identifying the sentiment expressed indirectly or implicitly in text. It requires uncovering sentiments that are conveyed through subtle cues, such as contextual clues, tone, or linguistic patterns. Li et al. (2021) divided the Laptop and Restaurant reviews from SemEval 2014 (Pontiki et al., 2014) into two parts: implicit and explicit. For our analysis, we only utilized the implicit dataset and merged the data from both domains into a single dataset. To evaluate the performance, we employed accuracy as the metric.

Hate Speech Detection Hate speech detection refers to the process of identifying content that promotes discrimination, hostility, or violence against individuals or groups based on attributes such as race, religion, ethnicity, gender, sexual orientation, or other protected characteristics (Schmidt and Wiegand, 2017). For our analysis, we utilize the dataset from the SemEval2019 HatEval challenge (Basile et al., 2019). This dataset focuses on predicting whether a tweet exhibits hateful content towards two specific target communities: immigrants and women. We calculate the macro-averaged F1 score across the two binary classes: *hate* and *non-hate*.

Irony Detection Irony is a rhetorical device where the intended meaning of a statement is different or opposite to its literal interpretation. Irony detection aims to recognize and understand instances of irony in the text (Zeng and Li, 2022). We choose the Subtask 3A dataset of the SemEval2018 Irony Detection challenge (Hee et al., 2018) (referred to as "Irony18"). The goal is to determine whether a tweet contains ironic intent or not. For evaluation, we follow the convention to specifically consider the F1 score for the *irony* class, while ignoring *non-irony* F1 score.

Offensive Language Identification Offensive language identification involves identifying and flagging text that contains offensive or inappropriate content, including profanity, vulgarities, obscenities, or derogatory remarks (Pradhan et al., 2020). Different from hate speech, offensive language does not necessarily target a specific individual or group. For example, profanity expressions can be considered offensive language even when not directed at anyone in particular. We use the SemEval2019 OffensEval dataset (Zampieri et al., 2019). It involves classifying each given text as either *offensive* or *non-offensive*. We adopt a macro-averaged F1 score as the metric.

Stance Detection Stance detection refers to determining the perspective or stance expressed in a given text towards a particular topic or entity. It helps identify whether the text expresses *favor*, *against*, or *none* opinion towards a subject (Küçük and Can, 2020). We utilize the SemEval2016 shared task on Detection Stance in Tweets (Mohammad et al., 2016), and refer to it as "Stance16". It provides data in five domains (i.e., targets): abortion, atheism, climate change, feminism, and Hillary Clinton. In order to facilitate evaluation, we aggregate these domains into a single dataset. When evaluating the results, we only consider macro-averaged of F1 of *favor* and *against* classes, and ignore *none* class, following previous studies.

Comparative Opinion Mining Comparative opinion mining is the task of analyzing opinions and sentiments expressed in a comparative context (Varathan et al., 2017). It involves comparing different aspects of a product, service, or any other subject to determine preferences or relative opinions. In our study, we take the CS19 dataset (Panchenko et al., 2019), which provides annotated comparative sentences in the field of computer science. These sentences involve comparisons between various targets such as programming languages, database products, and technology standards. The opinions expressed in the dataset are categorized as either *better* or *worse*. To evaluate the performance, we employ accuracy as the metric.

Emotion Recognition Emotion recognition involves the identification and understanding of emotions expressed in text (Sailunaz et al., 2018). It focuses on detecting and categorizing different emotional states. We use the dataset provided by the TweetEval benchmark (Barbieri et al., 2020), which we refer to it as "Emotion20". It transforms the SemEval2018 Affects in Tweets dataset (Mohammad et al., 2018) from multi-class classification into a multi-label dataset, by keeping only the tweets labeled with a single emotion. It selects the most common four emotions, namely *anger*, *joy*, *sadness*, and *optimism*. For evaluation, we utilize the macro-averaged F1 score, which considers the overall performance across all classes.

A.2 Details on Prompt Generation

Specifically, we provide the task description, format requirement (similar to those described in Sec 4.2), and an instruction to require GPT-4 to generate several prompts, representing as Python f-strings. We also optionally provide some input-target pairs to help the model better grasp the goals of the task. We present an example prompt in Figure 4, using the aspect-level SC task for illustration.

A.3 Cost Analysis

We provide a comparison of the average cost per task category when utilizing `ChatGPT` and `T5`$_{large}$ in our experiments, as detailed in Table 5 for reference. In practical applications, costs are influenced by a multitude of factors, such as the availability of training data, the volume of inference requests, and the pricing of cloud services or APIs. Developers are advised to select models based on their specific requirements and use-case scenarios.

Input:
The aspect sentiment classification task is to assign a sentiment label towards a specific aspect from the label space given a text.

To solve this task, a model will be given the original text (\`text\`), and the target aspect (\`aspect\`), and it is supposed to predict the corresponding label which must fall into a predefined label space (\`label_space\` - a list of possible labels).

Based on the above information, please suggest 10 prompts for large language models that instructs the model to solve the task with the given information. Represent the prompt as a Python f-string that uses the provided information as variables in the string.

Output:
```
f"In the following review text, determine the sentiment
expressed towards the given aspect: '{text}'. The aspect
under consideration is '{aspect}'. Choose your answer
from the following options: {label_space}."
...
```

Figure 4: Example prompts generated by GPT-4 for the aspect-level SC task. The first generated prompt is shown for illustrative purposes, and subsequent prompts are not included for brevity.

A.4 Detailed results in few-shot settings

We present detailed few-shot performance of various sentiment analysis tasks in Table 6. All the results are reported with average and standard deviation in 3 runs.

A.5 Discussions

A.5.1 LLMs for SA in Practice

In this study, we carry out a comprehensive evaluation of various large language models across a range of sentiment analysis tasks. The experimental results lead us to several primary findings and recommendations for practical SA application:

Task	0-shot	1-shot		5-shot		10-shot		Full
	ChatGPT	ChatGPT	T5$_{large}$	ChatGPT	T5$_{large}$	ChatGPT	T5$_{large}$	T5$_{large}$
SC	0.10	0.29	0.46	0.30	0.64	0.58	0.88	45.49
ABSA	0.10	0.12	0.46	0.37	0.61	0.65	0.79	0.65
MAST	0.05	0.23	0.49	0.65	0.73	1.19	0.53	1.65
Average	0.09	0.22	0.47	0.46	0.67	0.83	0.72	16.44

Table 5: Average Cost Comparison in $USD for ChatGPT and T5$_{large}$

Task	Dataset	1-shot			5-shot			10-shot	
		Flan-UL2	ChatGPT	T5$_{large}$	Flan-UL2	ChatGPT	T5$_{large}$	ChatGPT	T5$_{large}$
				Sentiment Classification (SC)					
Document-Level	IMDb	NA	95.33$_{0.50}$	77.20$_{10.74}$	NA	NA	90.00$_{2.03}$	NA	91.80$_{1.44}$
	Yelp2	NA	97.60$_{0.92}$	86.60$_{5.56}$	NA	NA	92.40$_{0.00}$	NA	90.87$_{1.63}$
	Yelp5	NA	51.47$_{2.50}$	36.47$_{4.40}$	NA	NA	44.53$_{3.19}$	NA	50.60$_{0.53}$
Sentence-Level	MR	92.87$_{0.23}$	91.60$_{0.40}$	72.87$_{9.15}$	93.80$_{0.00}$	90.20$_{0.53}$	85.67$_{1.62}$	87.53$_{3.44}$	86.60$_{1.22}$
	SST2	97.00$_{0.20}$	94.87$_{0.81}$	59.33$_{2.89}$	97.40$_{0.20}$	95.27$_{0.46}$	91.40$_{3.36}$	90.93$_{3.72}$	94.60$_{0.72}$
	Twitter	47.53$_{0.31}$	66.47$_{1.62}$	28.33$_{7.96}$	47.93$_{0.31}$	64.33$_{1.40}$	53.20$_{4.65}$	62.73$_{0.81}$	56.60$_{3.14}$
	SST5	51.80$_{0.92}$	51.87$_{0.76}$	26.67$_{1.10}$	NA	51.00$_{3.27}$	39.00$_{1.25}$	47.60$_{1.25}$	40.27$_{4.84}$
Aspect-Level	Lap14	73.60$_{0.20}$	78.60$_{3.14}$	65.47$_{1.10}$	73.47$_{0.12}$	76.27$_{2.37}$	69.13$_{1.50}$	76.67$_{2.41}$	74.40$_{0.87}$
	Rest14	82.87$_{0.23}$	84.53$_{0.64}$	52.47$_{19.00}$	83.07$_{0.12}$	74.87$_{7.40}$	75.80$_{0.20}$	74.20$_{4.13}$	70.47$_{1.70}$
				Aspect-based Sentiment Analysis (ABSA)					
UABSA	Rest14	16.67$_{2.90}$	63.62$_{0.89}$	18.43$_{4.17}$	NA	62.40$_{1.02}$	36.55$_{1.92}$	63.30$_{1.21}$	44.07$_{2.19}$
	Rest15	16.50$_{1.81}$	49.35$_{2.53}$	18.04$_{3.89}$	NA	52.18$_{1.56}$	29.95$_{0.35}$	52.85$_{0.75}$	38.96$_{1.44}$
	Rest16	17.98$_{2.10}$	56.50$_{2.34}$	15.86$_{4.38}$	NA	57.74$_{0.39}$	32.32$_{3.43}$	59.22$_{2.00}$	46.62$_{4.28}$
	Laptop14	13.29$_{0.88}$	40.82$_{4.61}$	10.47$_{2.30}$	NA	42.67$_{0.12}$	20.00$_{2.22}$	44.70$_{1.36}$	28.38$_{0.89}$
ASTE	Rest14	9.26$_{1.75}$	44.92$_{3.53}$	5.62$_{4.35}$	NA	50.75$_{5.93}$	25.00$_{4.09}$	54.11$_{2.98}$	33.17$_{1.21}$
	Rest15	9.31$_{0.43}$	47.30$_{1.96}$	9.19$_{1.15}$	NA	49.99$_{4.34}$	27.44$_{1.26}$	48.11$_{0.78}$	32.28$_{2.29}$
	Rest16	11.81$_{1.99}$	50.09$_{4.28}$	9.48$_{8.84}$	NA	51.30$_{0.47}$	26.44$_{2.52}$	53.60$_{4.51}$	32.14$_{4.38}$
	Laptop14	5.19$_{1.54}$	35.49$_{3.38}$	2.94$_{2.14}$	NA	42.56$_{1.78}$	15.52$_{3.14}$	44.74$_{2.36}$	21.95$_{3.50}$
ASQP	Rest15	NA	30.15$_{1.48}$	8.69$_{0.95}$	NA	31.21$_{1.94}$	13.75$_{0.78}$	30.92$_{2.78}$	14.87$_{1.06}$
	Rest16	NA	31.98$_{2.06}$	2.53$_{2.14}$	NA	38.01$_{2.28}$	14.40$_{4.76}$	40.15$_{1.49}$	19.23$_{1.42}$
				Multifaceted Analysis of Subjective Text (MAST)					
Implicit	Lap+Res	49.40$_{0.79}$	65.08$_{4.89}$	34.01$_{10.13}$	50.91$_{1.17}$	59.58$_{5.01}$	46.53$_{4.12}$	59.73$_{1.85}$	52.56$_{9.98}$
Hate	HatEval	64.76$_{0.97}$	55.88$_{8.17}$	25.77$_{3.17}$	64.12$_{3.32}$	50.46$_{1.57}$	49.89$_{5.29}$	57.96$_{3.34}$	52.54$_{3.03}$
Irony	Irony18	81.78$_{0.87}$	79.57$_{2.76}$	38.23$_{10.72}$	82.32$_{0.45}$	84.28$_{1.30}$	57.69$_{7.55}$	80.16$_{1.47}$	58.90$_{2.40}$
Offensive	OffensEval	77.29$_{0.47}$	72.75$_{1.63}$	17.67$_{7.35}$	78.01$_{1.14}$	72.54$_{1.34}$	49.19$_{1.26}$	70.21$_{3.33}$	49.97$_{5.66}$
Stance	Stance16	67.75$_{1.96}$	59.31$_{1.81}$	33.37$_{4.22}$	70.49$_{0.80}$	53.53$_{5.04}$	35.15$_{3.78}$	43.15$_{5.33}$	36.94$_{1.75}$
Comparative	CS19	86.62$_{1.10}$	73.99$_{2.96}$	46.39$_{11.98}$	87.26$_{1.10}$	68.79$_{3.32}$	70.28$_{4.03}$	68.26$_{3.83}$	71.87$_{2.07}$
Emotion	Emotion20	71.05$_{0.73}$	72.59$_{2.01}$	43.16$_{9.98}$	69.85$_{2.02}$	74.30$_{2.41}$	65.08$_{4.23}$	69.88$_{1.34}$	71.60$_{0.55}$

Table 6: Few-shot performance of various sentiment analysis tasks. All the results are reported with average and standard deviation in 3 runs. "NA" denotes infeasible experiments due to limited sequence length.

- For simple SA tasks such as binary or trinary sentiment classification, LLMs can already serve as effective solutions. Even in a zero-shot setting, their performance can match or surpass fine-tuned smaller language models, and with little sensitivity to different prompt designs (as shown in Sec 5.2).

- When annotation resources are scarce, LLMs remain a good choice due to their superior few-shot in-context learning performance compared to SLMs trained on the same limited data. However, the restricted context length of LLMs can limit their use case, particularly in document-level tasks where SLMs might be more suitable.

- For tasks requiring structured sentiment output, like aspect-based sentiment analysis tasks, LLMs might not be the best option. They tend to lag behind SLMs in both automatic and human evaluations, and performance can vary significantly with different prompt designs.

- Larger models do not always guarantee superior performance, for instance, `Flan-UL2` often performs comparably to the GPT-3.5 series of models, despite being much smaller in size. This suggests that employing instruction-tuning to attain a reasonably sized model may suffice for practical SA applications.

A.5.2 SA Challenges for LLMs

With the advancement of LLMs, many SA tasks can be claimed to be solved such as binary sentiment classification, as we saw from the experimental results. However, does it mean sentiment analysis in general has reached its maturity in the era of LLMs? We discuss some remaining challenges that we think still pose great difficulties.

Understanding Complex Linguistic Nuances and Cultural Specificity Sentiment is often shaded with nuance and subtlety. Developing models capable of understanding such subtleties in language, such as sarcasm, irony, humor, and specific cultural idioms or expressions is still challenging. They often depend on the context and shared cultural background knowledge or even specific human experiences. For example, on Chinese social media, a comment "您说的都对" (English translation: "You are right about everything you said" with "You" in a respectful tone) may not necessarily indicate agreement but can be used ironically. However, this linguistic phenomenon may require familiarity with social media to interpret correctly.

Extracting fine-grained and structured sentiment information As can be seen from the results, requiring the models to generate structured fine-grained information, i.e., the ABSA tasks, is still challenging for the models. However, such information can be useful to quickly summarize large-scale information to produce a more organized digest, especially since the long context is still a limitation for many LLMs. Also, distinguishing more precise emotional states or intensities of sentiment for more detailed analysis is also challenging but worth exploring.

Real-Time Adaptation for Evolving Sentiment Analysis Sentiments and expressions constantly evolve, particularly on platforms like social media. This leads to the continual emergence of new idioms and sentiment-caring expressions. It thus demands the sentiment analysis models to adapt and learn from these evolving trends to accurately interpret the embedded sentiments. However, one of the major limitations of current LLMs lies in their lack of flexibility in fine-tuning or re-training. This issue restricts their capability to keep up with the fast-paced evolution of language and sentiment, resulting in outdated or inaccurate sentiment analysis. Therefore, a critical research direction involves developing methods for rapid and effective model updates to ensure real-time and accurate sentiment analysis.

A.6 Prompts for Each SA Task

We present a 1-shot prompt for each investigated sentiment analysis task, which is shown on the following pages.

task	Dataset	1-shot Prompt
SC	IMDb	Please perform Sentiment Classification task. Given the sentence, assign a sentiment label from ['negative', 'positive']. Return label only without any other text. Sentence: I 've seen the original English version on video . Disney 's choice of voice actors looks very promising .... Label:positive Sentence: " This is a depressingly shallow , naive and mostly unfunny look at a wildly improbable relationship between Brooks ' psychotic film editor and Harold , his vapid girlfriend .... Label:negative Sentence: " Jack and Kate meet the physician Daniel Farady first and then the psychics Miles Straume and they demonstrate that have not come to the island with the intention of rescuing the survivors . Locke and his group find the anthropologist Charlotte Staples Lewis , and Ben Linus shoots her . Meanwhile , the group of Jack finds the pilot Frank Lapidus , who landed the helicopter with minor damages that can be repaired . Jack forces Miles to tell the real intention why they have come to the island. < br / > < br / > The second episode of the Fourth Season returns to the island , with four new characters , stops the confusing " " flash-forwards " " and it seems that will finally be the beginning of the explanations that I (and most of the fans and viewers) expect to be provided in " " Lost " " . Why the interest of the government in Ben Linus , and how he is informed from the boat are some of the questions that I expect to see in the next episodes . My vote is eight. < br / > < br / > Title (Brazil) : Not Available " Label:
SC	Yelp-2	Please perform Sentiment Classification task. Given the sentence, assign a sentiment label from ['negative', 'positive']. Return label only without any other text. Sentence: Had a great time with my beautiful wife listening to The Instant Classics . Drinks are pricey and menu seems a little limited , but I had a great time .... Label:positive Sentence: I have been to this location multiple times and every time the service is horrendous and the food is mediocre . Not sure if the location being in a mall has to do with it .... Label:negative Sentence: I expected the prices of the entrees to be a little bit higher but the quality of the Chinese food was not worth the money I paid for the dishes . I got the 18 monk noodle and the traditional dimsum . If I could describe the food in one word-terrible ! Making the dimsum look pretty by topping it with gold flakes did not do anything to make up for the flavor of the dimsum . It seemed too starchy and you can hardly taste the meat . The noodles looked like a sad , greasy slop of Mai fun type noodles (noodles were stuck together) saturated with soy sauce for color , and garnished with a few pieces of shitake mushrooms , green onions and fine threads of carrots . And yes , portions were small , but that 's not really the worst part of the whole experience . Just poorly prepared , way overpriced Chinese food ... sorry . Label:

Continued on next page

Continued from previous page

SC	Yelp-5	Please perform Sentiment Classification task. Given the sentence, assign a sentiment label from ['negative', 'neutral', 'positive', 'very negative', 'very positive']. Return label only without any other text. Sentence: The most important thing to me in an airline is that we do not fall out of the sky in an uncontrolled fashion . After all landing is a controlled crash .... Label:neutral Sentence: " Great place to go for hair , nails or massage . Great service in a professional and clean environment . Most places u have to wait even if u have an appt .... Label:very positive Sentence: Loved the atmosphere . Right across from chase field . The pretzel and provolone and shrimp appetizers were plentiful and fantastic . Easily enough for four people to share .... Label:positive Sentence: " 1 star- why ? The food was n't too bad . My husband had the fish tacos which were good . I ordered the Sicilian Stuffed Chicken , but get this .... Label:negative Sentence: " Hello there ! 00a0 00a0 00a0 My name is Naiby Moreno , and the reason why I 'm writing you this email is because last night , around this time .... Label:very negative Sentence: Came a few days ago for a lease , was n't sure of size needed , so I guessed , three times ! Finally got it right , but hey , the store did n't bat a eye lash when I returned the ones that did n't work , they just asked if I needed help picking out a replacement . Since my cat has been loosing weight , I could not get the size down , so after my attempts , finally got the small dog size and sure enough it worked . Now to get the cat used to it before we need it . This store has everything you could need . They is even a new section by Martha Stewart , everything for you little pet . But her stuffs pricey , a lease from here collection , $ 19.99 , boy that 's steep ! The store is clean , neatly kept , well organized and they have grooming services . The employees were friendly and helpful , they looked like they enjoyed their jobs , and I would make this a regular place . Label:
SC	MR	Please perform Sentiment Classification task. Given the sentence, assign a sentiment label from ['negative', 'positive']. Return label only without any other text. Sentence: " it 's the chemistry between the women and the droll scene-stealing wit and wolfish pessimism of anna chancellor that makes this " " two weddings and a funeral " " fun . " Label:positive Sentence: the entire movie is about a boring , sad man being boring and sad . Label:negative Sentence: " if you 're a crocodile hunter fan , you 'll enjoy at least the " " real " " portions of the film . if you 're looking for a story , do n't bother . " Label:
SC	SST2	Please perform Sentiment Classification task. Given the sentence, assign a sentiment label from ['negative', 'positive']. Return label only without any other text. Sentence: Oh , and more entertaining , too . Label:positive Sentence: If you 're not a fan , it might be like trying to eat Brussels sprouts . Label:negative Sentence: An ungainly , comedy-deficient , B-movie rush job ... Label:

Continued on next page

Continued from previous page		
SC	Twitter	Please perform Sentiment Classification task. Given the sentence, assign a sentiment label from ['negative', 'neutral', 'positive']. Return label only without any other text. Sentence: - Just bought my 1st iPad, iPad3, feeling real burned, mad, about iPad4 so soon. Grrr. REALLY mad! Don't even care about mini now," Label:negative Sentence: @user @user @user I think this is the motive of the Yakub's laywers for pursuing the case Label:neutral Sentence: Kanye West was honored in a big way during Sunday night's MTV Video Music Awards by receiving the Michael Jackso... Label:positive Sentence: Do you think Michelle Obama wanted to smack Melania Trump for plagiarizing her convention speech? She has the arms for it. Label:
SC	SST5	Please perform Sentiment Classification task. Given the sentence, assign a sentiment label from ['negative', 'neutral', 'positive', 'very negative', 'very positive']. Return label only without any other text. Sentence: ' Like a child with an important message to tell ... (Skins ') faults are easy to forgive because the intentions are lofty . ' Label:neutral Sentence: That Haynes can both maintain and dismantle the facades that his genre and his character construct is a wonderous accomplishment of veracity and narrative grace . Label:very positive Sentence: Oh , and more entertaining , too . Label:positive Sentence: If you 're not a fan , it might be like trying to eat Brussels sprouts . Label:negative Sentence: When it comes out on video , then it 's the perfect cure for insomnia . Label:very negative Sentence: Everywhere the camera looks there is something worth seeing . Label:
SC	Lap14	Please perform Aspect Sentiment Classification task. Given the sentence, assign a sentiment label towards "Office" from ['negative', 'neutral', 'positive']. Return label only without any other text. Sentence: It even has a great webcam , and Skype works very well . (sentiment towards "webcam") Label:positive Sentence: - Touchpad will take a bit of time to get used to . (sentiment towards "-Touchpad") Label:neutral Sentence:) And printing from either word processor is an adventure . (sentiment towards "word processor") Label:negative Sentence: (but Office can be purchased) IF I ever need a laptop again I am for sure purchasing another Toshiba !! Label:
Continued on next page		

Continued from previous page		
SC	Rest14	Please perform Aspect Sentiment Classification task. Given the sentence, assign a sentiment label towards "garlic knots" from ['negative', 'neutral', 'positive']. Return label only without any other text. Sentence: While the new restaurant still features much of the same classical furniture that made Tiffin so attractive , the menu has been overhauled . (sentiment towards "classical furniture") Label:positive Sentence: And it all comes at a very reasonable price (congee , noodles , and rice dishes are no more than 3-6 each) . (sentiment towards "(congee") Label:neutral Sentence: The Singapore Mai Fun had NO curry flavor whatsoever . (sentiment towards "curry flavor") Label:negative Sentence: I also recommend the garlic knots . Label:
UABSA	Rest14	Please perform Unified Aspect-Based Sentiment Analysis task. Given the sentence, tag all (aspect, sentiment) pairs. Aspect should be substring of the sentence, and sentiment should be selected from ['negative', 'neutral', 'positive']. If there are no aspect-sentiment pairs, return an empty list. Otherwise return a python list of tuples containing two strings in double quotes. Please return python list only, without any other comments or texts. Sentence: also make sure you pay attention to the music being piped in , quite a weird selection . Label:[('music', 'neutral')] Sentence: but I would n't wan na live there . Label:[] Sentence: And their prices are very high , they actually think that they can get away with charging such prices for such terrible food and service ! Label:[('prices', 'negative'), ('prices', 'negative'), ('food', 'negative'), ('service', 'negative')] Sentence: Having not been home in the last 2 years may skew this reviewer a bit , but the food was tasty and spicy sans the oil that comes floating along at similar venues . Label:[('food', 'positive'), ('oil', 'neutral')] Sentence: After I paid for my purchase , I noticed they had not given me utensils so I could eat my pie . Label:
UABSA	Rest15	Please perform Unified Aspect-Based Sentiment Analysis task. Given the sentence, tag all (aspect, sentiment) pairs. Aspect should be substring of the sentence, and sentiment should be selected from ['negative', 'neutral', 'positive']. If there are no aspect-sentiment pairs, return an empty list. Otherwise return a python list of tuples containing two strings in double quotes. Please return python list only, without any other comments or texts. Sentence: The portions are HUGE , so it might be good to order three things to split rather than one appetizer and entree per person for two people . Label:[('portions', 'neutral')] Sentence: No , really . Label:[] Sentence: The food was bland oily . Label:[('food', 'negative')] Sentence: The food 's as good as ever . Label:[('food', 'positive')] Sentence: Need I say more ? Label:
Continued on next page		

Continued from previous page

UABSA	Rest16	Please perform Unified Aspect-Based Sentiment Analysis task. Given the sentence, tag all (aspect, sentiment) pairs. Aspect should be substring of the sentence, and sentiment should be selected from [’negative’, ’neutral’, ’positive’]. If there are no aspect-sentiment pairs, return an empty list. Otherwise return a python list of tuples containing two strings in double quotes. Please return python list only, without any other comments or texts. Sentence: Food was okay , nothing great . Label:[(’Food’, ’neutral’)] Sentence: I live in the neightborhood and am a regular . Label:[] Sentence: The place is small and cramped but the food is fantastic . Label:[(’place’, ’negative’), (’food’, ’positive’)] Sentence: One special roll and one regular roll is enough to fill you up , but save room for dessert ! Label:[(’special roll’, ’positive’), (’regular roll’, ’positive’), (’dessert’, ’positive’)] Sentence: The atmosphere is aspiring , and the decor is festive and amazing . Label:
UABSA	Laptop14	Please perform Unified Aspect-Based Sentiment Analysis task. Given the sentence, tag all (aspect, sentiment) pairs. Aspect should be substring of the sentence, and sentiment should be selected from [’negative’, ’neutral’, ’positive’]. If there are no aspect-sentiment pairs, return an empty list. Otherwise return a python list of tuples containing two strings in double quotes. Please return python list only, without any other comments or texts. Sentence: After that the said it was under warranty . Label:[(’warranty’, ’neutral’)] Sentence: I really wanted a Mac over a pc because I used a Mac in high school . Label:[] Sentence: Another issue I have with it is the battery . Label:[(’battery’, ’negative’)] Sentence: I love the size , keyboard , the functions . Label:[(’size’, ’positive’), (’keyboard’, ’positive’), (’functions’, ’positive’)] Sentence: Hopefully my replacement is brand new . Label:
ASTE	Rest 14	Please perform Aspect Sentiment Triplet Extraction task. Given the sentence, tag all (aspect, opinion, sentiment) triplets. Aspect and opinion should be substring of the sentence, and sentiment should be selected from [’negative’, ’neutral’, ’positive’]. Return a python list of tuples containing three strings in double quotes. Please return python list only, without any other comments or texts. Sentence: Service was slow had to wait to order and get food although not crowded . Label:[(’Service’, ’slow’, ’negative’)] Sentence: The atmosphere is n’t the greatest , but I suppose that ’s how they keep the prices down . Label:[(’atmosphere’, "is n’t the greatest", ’neutral’), (’prices’, ’down’, ’positive’)] Sentence: The fries are yummy . Label:[(’fries’, ’yummy’, ’positive’)] Sentence: Most importantly , it is reasonably priced . Label:
ASTE	Rest 15	Please perform Aspect Sentiment Triplet Extraction task. Given the sentence, tag all (aspect, opinion, sentiment) triplets. Aspect and opinion should be substring of the sentence, and sentiment should be selected from [’negative’, ’neutral’, ’positive’]. Return a python list of tuples containing three strings in double quotes. Please return python list only, without any other comments or texts. Sentence: the only things u could really taste are the very salty soy sauce (even its low sodium) , the vinegar-soaked rice , and the scallion on top of the fish . Label:[(’soy sauce’, ’salty’, ’negative’), (’rice’, ’vinegar-soaked’, ’negative’)] Sentence: Food was okay , nothing great . Label:[(’Food’, ’okay’, ’neutral’), (’Food’, ’nothing great’, ’neutral’)] Sentence: We recently decided to try this location , and to our delight , they have outdoor seating , perfect since I had my yorkie with me . Label:[(’outdoor seating’, ’perfect’, ’positive’)] Sentence: This establishment is the real deal . Label:

Continued on next page

Continued from previous page		
ASTE	Rest 16	Please perform Aspect Sentiment Triplet Extraction task. Given the sentence, tag all (aspect, opinion, sentiment) triplets. Aspect and opinion should be substring of the sentence, and sentiment should be selected from ['negative', 'neutral', 'positive']. Return a python list of tuples containing three strings in double quotes. Please return python list only, without any other comments or texts. Sentence: limited menu , no-so-fresh ingredients , thinly-sliced fish , fall-apart rice . Label:[('menu', 'limited', 'negative'), ('ingredients', 'no-so-fresh', 'negative'), ('fish', 'thinly-sliced', 'negative'), ('rice', 'fall-apart', 'negative')] Sentence: For desserts , we tried the frozen black sesame mousse (interesting but not extraordinary) and matcha (powdered green tea) and blueberry cheesecake , which was phenomenal . Label:[('frozen black sesame mousse', 'interesting', 'neutral'), ('frozen black sesame mousse', 'extraordinary', 'neutral'), ('matcha (powdered green tea) and blueberry cheesecake', 'phenomenal', 'positive')] Sentence: The food was good . Label:[('food', 'good', 'positive')] Sentence: In Grammercy/Union Square/East Village this is my neighbors and my favorite spot . Label:
ASTE	Laptap14	Please perform Aspect Sentiment Triplet Extraction task. Given the sentence, tag all (aspect, opinion, sentiment) triplets. Aspect and opinion should be substring of the sentence, and sentiment should be selected from ['negative', 'neutral', 'positive']. Return a python list of tuples containing three strings in double quotes. Please return python list only, without any other comments or texts. Sentence: Dealing with the support drone on the other end of the chat was sheer torture . Label:[('support', 'sheer torture', 'negative')] Sentence: I did think it had a camera because that was one of my requirements , but forgot to check in the specifications on this one before I purchased . Label:[('specifications', 'check in', 'neutral')] Sentence: A longer battery life would have been great - but it meets it 's spec quite easily . Label:[('spec', 'easily', 'positive')] Sentence: It was important that it was powerful enough to do all of the tasks he needed on the internet , word processing , graphic design and gaming . Label:
Continued on next page		

Continued from previous page

ASQP	Rest15	Please perform Aspect Sentiment Quad Prediction task. Given the sentence, tag all (category, aspect, opinion, sentiment) quadruples. Aspect and opinion should be substring of the sentence. Category should be selected from ['ambience general', 'drinks prices', 'drinks quality', 'drinks style_options', 'food general', 'food prices', 'food quality', 'food style_options', 'location general', 'restaurant general', 'restaurant miscellaneous', 'restaurant prices', 'service general']. Sentiment should be selected from ['negative', 'neutral', 'positive']. Only aspect can be 'NULL', category, opinion and sentiment cannot be 'NULL'. Return a python list of tuples containing four strings in double quotes. Please return python list only, without any other comments or texts. Sentence: The price is reasonable although the service is poor . Label:[('restaurant prices', 'NULL', 'reasonable', 'positive'), ('service general', 'service', 'poor', 'negative')] Sentence: This little place definitely exceeded my expectations and you sure get a lot of food for your money . Label:[('food style_options', 'food', 'lot', 'positive'), ('restaurant general', 'place', 'exceeded my expectations', 'positive'), ('food prices', 'food', 'lot', 'positive')] Sentence: This place is really trendi but they have forgotten about the most important part of a restaurant , the food . Label:[('food quality', 'food', 'forgotten', 'negative'), ('ambience general', 'place', 'trendi', 'positive')] Sentence: The restaurant looks out over beautiful green lawns to the Hudson River and the Statue of Liberty . Label:[('location general', 'restaurant', 'beautiful', 'positive')] Sentence: With so many good restaurants on the UWS , I do n't need overpriced food , absurdly arrogant wait-staff who do n't recognize they work at a glorified diner , clumsy service , and management that does n't care . Label:[('food prices', 'food', 'overpriced', 'negative'), ('service general', 'wait-staff', 'arrogant', 'negative'), ('service general', 'service', 'clumsy', 'negative'), ('service general', 'management', "does n't care", 'negative')] Sentence: the drinks are amazing and half off till 8pm . Label:[('drinks quality', 'drinks', 'amazing', 'positive'), ('drinks prices', 'drinks', 'amazing', 'positive')] Sentence: A cool bar with great food , and tons of excellent beer . Label:[('ambience general', 'bar', 'cool', 'positive'), ('food quality', 'food', 'great', 'positive'), ('drinks quality', 'beer', 'excellent', 'positive'), ('drinks style_options', 'beer', 'excellent', 'positive')] Sentence: The food is great and reasonably priced . Label:[('food quality', 'food', 'great', 'positive'), ('food prices', 'food', 'reasonably priced', 'positive')] .... Sentence: For me dishes a little oily , but overall dining experience good . Label:

Continued on next page

Continued from previous page		
ASQP	Rest16	Please perform Aspect Sentiment Quad Prediction task. Given the sentence, tag all (category, aspect, opinion, sentiment) quadruples. Aspect and opinion should be substring of the sentence. Category should be selected from ['ambience general', 'drinks prices', 'drinks quality', 'drinks style_options', 'food general', 'food prices', 'food quality', 'food style_options', 'location general', 'restaurant general', 'restaurant miscellaneous', 'restaurant prices', 'service general']. Sentiment should be selected from ['negative', 'neutral', 'positive']. Only aspect can be 'NULL', category, opinion and sentiment cannot be 'NULL'. Return a python list of tuples containing four strings in double quotes. Please return python list only, without any other comments or texts. Sentence: The wine list is interesting and has many good values . Label:[('drinks style_options', 'wine list', 'interesting', 'positive'), ('drinks prices', 'wine list', 'good values', 'positive')] Sentence: The food is amazing ... especially if you get the Chef 's tasting menu and your favourite bottle (or two !) of wine from an extensive selection of wines . k Label:[('food quality', 'food', 'amazing', 'positive'), ('drinks style_options', 'selection of wines', 'extensive', 'positive'), ('food quality', "Chef 's tasting menu", 'favourite', 'positive')] Sentence: Gorgeous place ideal for a romantic dinner Label:[('ambience general', 'place', 'Gorgeous', 'positive'), ('restaurant miscellaneous', 'place', 'ideal', 'positive')] Sentence: The drinks are great , especially when made by Raymond . Label:[('drinks quality', 'drinks', 'great', 'positive'), ('service general', 'Raymond', 'great', 'positive')].... Sentence: It was worth the wait . Label:
Implicit	Lap+Res	Please perform Aspect-Based Implicit Sentiment Analysis task. Given the sentence, please infer the sentiment towards the aspect "vintages". Please select a sentiment label from ['negative', 'neutral', 'positive']. Return label only without any other text. Sentence: The steak was excellent and one of the best I have had (I tasted the butter intitally but in no way did it overwhelm the flavor of the meat). (sentiment towards "butter") Label:negative Sentence: Yes, they use fancy ingredients, but even fancy ingredients don't make for good pizza unless someone knows how to get the crust right. (sentiment towards "crust") Label:neutral Sentence: Three page wine menu, one page entree and horedevous. (sentiment towards "wine menu") Label:positive Sentence: Somewhat disappointing wine list (only new vintages. Label:
Hate	HatEval	Please perform Hate Detection task. Given the sentence, assign a sentiment label from ['hate', 'non-hate']. Return label only without any other text. Sentence: My family's idea of a merienda for this moment is siopao. They really hate me. Me: *calls Tim Ho Wan* Do you deliver in elyu? Label:non-hate Sentence: This is horrendous Label:hate Sentence: @user id marry this fukin whore, let the bitch behind her be best lady at the wedding Label:
Irony	Irony18	Please perform Irony Detection task. Given the sentence, please determine wheter or not it contains irony. Assign a sentiment label from ['irony', 'non_irony']. Return label only without any other text. Sentence: @user You truly are my son. Label:non_irony Sentence: Just watched how Pretzels were made. Label:irony Sentence: Fighting over chargers is definitely how I wanted to start my day. Label:
Continued on next page		

Continued from previous page		
Offensive	OffensEval	Please perform Offensive Detection task. Given the sentence, assign a sentiment label from ['non-offensive', 'offensive']. Return label only without any other text. Sentence: user Hi Bernice I hope you are enjoying the xrpcommunity and learning lots about xrp 0589 user Label:non-offensive Sentence: @user this isn't me disagreeing this is me basically saying that i hope you're right but if you are i will spontaneously combust Label:offensive Sentence: MAGA ... got any ideas how she could have done it? Label:
Stance	Stance16	Please perform Stance Detection (abortion) task. Given the sentence, assign a sentiment label expressed by the author towards "abortion" from ['against', 'favor', 'none']. Return label only without any other text. Sentence: user i don't follow the news, is there a new law that ALL gay people have to get married? I'm against that! #SemST (opinion towards "abortion") Label:none Sentence: The natural world is part of our inheritance, we have to protect it user with user on #BBC #Earth #SemST (opinion towards "climate") Label:favor Sentence: user we lost 4,000 of our Military boys when your President pulled out of Iraq. #LiberalConsequences #SemST (opinion towards "hillary") Label:against Sentence: Women have outgrown the common housewife stigma long ago #SemST Label:
Comparative	CS19	Please perform Comparative Opinions task. Given the sentence, compare "Microsoft" to "Sony", and assign an opinion label from ['better', 'worse']. Return label only without any other text. Sentence: Java isn't too bad of a first language, but Python is a little easier to pick up. (compare "Java" to "Python") Label:worse Sentence: In supply-chain conversations, the Pacific Crest semiconductor team learned that Windows 7 inventory is moving faster than Windows 8. (compare "Windows 7" to "Windows 8") Label:better Sentence: And I think Microsoft will have more money to make better games than Sony. Label:
Emotion	Emotion20	Please perform Comparative Opinions task. Given the sentence, compare "Microsoft" to "Sony", and assign an opinion label from ['better', 'worse']. Return label only without any other text. Sentence: the football team is decent but getting better! the basketball teams are awesome!the Label:worse Sentence: Now let's be clear; in this author's humble opinion, Apple is still way better than IBM. Label:better Sentence: And I think Microsoft will have more money to make better games than Sony. Label:

Table 7: Detailed prompts for investigated tasks and datasets. We show 1-shot prompt for illustration.