Targeted Distillation for Sentiment Analysis

Anonymous ACL submission

Abstract

This paper presents a compact model that achieves strong sentiment analysis capabilities through targeted distillation from advanced large language models (LLMs). Our methodology decouples the distillation target into two key components: sentiment-related knowledge and task alignment. To transfer these components, we propose a two-stage distillation framework. The first stage, knowledge-driven distillation (KNOWDIST), transfers sentiment-related knowledge to enhance fundamental sentiment analysis capabilities. The second stage, in-context learning distillation (ICLDIST), transfers task-specific prompt-following abilities to optimize task alignment. For evaluation, we introduce SENTIBENCH, a comprehensive sentiment analysis benchmark comprising 3 task categories across 12 datasets. Experiments on this benchmark demonstrate that our model effectively balances model size and performance, showing strong competitiveness compared to existing small-scale LLMs.[1]

1 Introduction

Sentiment analysis, aiming to identify and extract subjective information from user-generated content (Liu, 2012), has emerged as a significant research area in natural language processing, garnering widespread attention (Zhang et al., 2018; Wankhade et al., 2022). Recent studies demonstrate that large language models (LLMs) exhibit remarkable capabilities and achieve state-of-the-art performance in sentiment analysis tasks (Zhang et al., 2024b; Wang et al., 2024c; Šmíd et al., 2024). Despite these advancements, the practical application of LLMs faces significant challenges. Deploying these models incurs considerable computational costs, and fine-tuning them for enhanced task-specific performance demands greater computational resources.

[1]We will release our code, data, and model weights via Github.

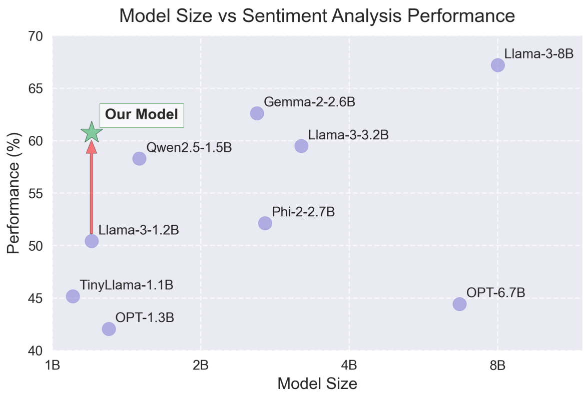

Figure 1: The comparison of our distilled model with other small-scale models in terms of the average performance on SENTIBENCH (F_1-score, %).

To reduce computational overhead, researchers are increasingly turning to knowledge distillation techniques (Hinton et al., 2015). These works focus on transferring general capabilities from advanced LLMs to their more cost-efficient counterparts through carefully curated instructions (Taori et al., 2023; Chiang et al., 2023; Wu et al., 2024). However, when substantial size gaps exist between teacher and student models, such generic distillation is challenging due to the difficulty in developing instructions with sufficient diversity and scale. Consequently, students often merely mimic the output style of teacher LLMs while performing poorly on specialized downstream tasks (Gudibande et al., 2023). In contrast, existing works demonstrate that for a specific application class, LLMs can be potentially approximated by a much smaller model (Xu et al., 2023b; Kim et al., 2024; Zhou et al., 2024). This suggests that targeted distillation towards specialized capabilities offers a more practical and promising direction.

Motivated by these insights, this paper explores targeted distillation specifically for sentiment analysis. We decouple the distillation target into knowledge and alignment and propose a two-stage distillation framework. The first stage, termed **knowledge-driven distillation** (KNOWDIST), focuses on transferring fundamental sentiment analysis capabilities, thereby improving the student model's potential performance. In KNOWDIST, we devise a multi-perspective prompting strategy to elicit comprehensive sentiment-related knowledge from the teacher LLM and systematically transfer this knowledge to the student model. The second stage, termed **in-context learning distillation** (ICLDIST), transfers prompt-following capabilities in sentiment analysis to optimize the student model's task alignment. In ICLDIST, we enable the student model to follow task-specific instructions and demonstrations by mimicking the teacher LLM's responses on few-shot samples. When constructing few-shot samples, we implement format and task diversification strategies to strengthen the generalization of ICLDIST.

To facilitate a systematic evaluation, we develop SENTIBENCH, a comprehensive sentiment analysis benchmark. Our extensive experimentation on this benchmark reveals several key findings: (1) Our approach demonstrates substantial advantages over generic distillation methods, achieving effective distillation of LLMs' sentiment analysis capabilities. Specifically, the student model achieves a 10% improvement in the average F_1-score across various tasks, with a particularly remarkable gain of 38% in irony detection. (2) Leveraging our approach, Llama-3-1.2B outperforms Llama-3-3.2B and exhibits strong competitiveness against other small-scale models (see Figure 1). (3) Further analysis reveals the complementary nature of KNOWDIST and ICLDIST and validate the effectiveness of each component in our approach.

2 Two-stage Distillation Framework

Following Taori et al. (2023); Chiang et al. (2023); Wu et al. (2024), we distill the capabilities of LLMs by making the student model learn from the teacher LLM's output $\boldsymbol{y}$ for specific prompts. Our prompts are composed of instructions $\boldsymbol{i}$, demonstrations $\boldsymbol{d}$ (which may be empty), and input texts $\boldsymbol{x}$. This process can be formulated as follows:

$$\boldsymbol{y} = \mathcal{M}(\boldsymbol{i}, \boldsymbol{d}, \boldsymbol{x}; \theta_T), \quad (1)$$

$$\mathcal{L}(\theta_S) = -\sum_{\boldsymbol{i},\boldsymbol{d},\boldsymbol{x},\boldsymbol{y}} \log P_{\mathcal{M}}(\boldsymbol{y} \mid \boldsymbol{i}, \boldsymbol{d}, \boldsymbol{x}; \theta_S), \quad (2)$$

$$\hat{\theta}_S = \operatorname*{argmin}_{\theta_S} \mathcal{L}(\theta_S), \quad (3)$$

where $\mathcal{M}$ denotes the teacher or student model, and θ_T and θ_S denote their respective parameters.

In contrast to prior research, this paper focuses on distilling the LLMs' capability specifically for sentiment analysis. Prior to distillation, we decouple the target into *sentiment-related knowledge* and *task alignment*. (1) The knowledge reflects a model's ability to comprehend the sentiments expressed in text, including accurate interpretation of sentiment expressions, precise targeting, and possession of the requisite background knowledge. The capacity of this knowledge within the model shapes its potential performance in sentiment analysis tasks. (2) The alignment refers to the model's ability to follow task-specific instructions and demonstrations, *i.e.*, its in-context learning ability. Such alignment capability determines the model's observable performance in sentiment analysis tasks. Based on this decoupling, we develop a distillation framework consisting of two stages: knowledge-driven distillation (KNOWDIST) and in-context learning distillation (ICLDIST).

2.1 Knowledge-Driven Distillation

At this stage, we develop two distinct prompting methods to elicit sentiment-related knowledge from LLMs. The first directs LLMs to *analyze* the sentiments embedded within the given text, while the second instructs LLMs to *rewrite* the text while maintaining its original sentiment. Crucially, both methods require LLMs to provide their reasoning process before generating the final output.

To enhance the effectiveness of these prompting methods, we devise a **multi-perspective prompting strategy**. This strategy defines four different perspectives: (1) EXPRESSION: centering on subjective words and phrases during analyzing or rewriting; (2) TARGET: focusing on the specific entities and their associated aspects being evaluated; (3) EMOTION: highlighting the emotional states and psychological reactions expressed in the text; (4) BACKGROUND: incorporating contextual information and domain knowledge necessary for understanding the sentiment. This strategy guides the analyzing and rewriting process from these four perspectives, thereby eliciting a more comprehensive range of sentiment-related knowledge. The specific prompts can be found in Appendix A.

We employ these prompting methods to perform KNOWDIST, as illustrated in Figure 2. Firstly, we collect a large and diverse set of user-generated content, including movie, product, and restaurant

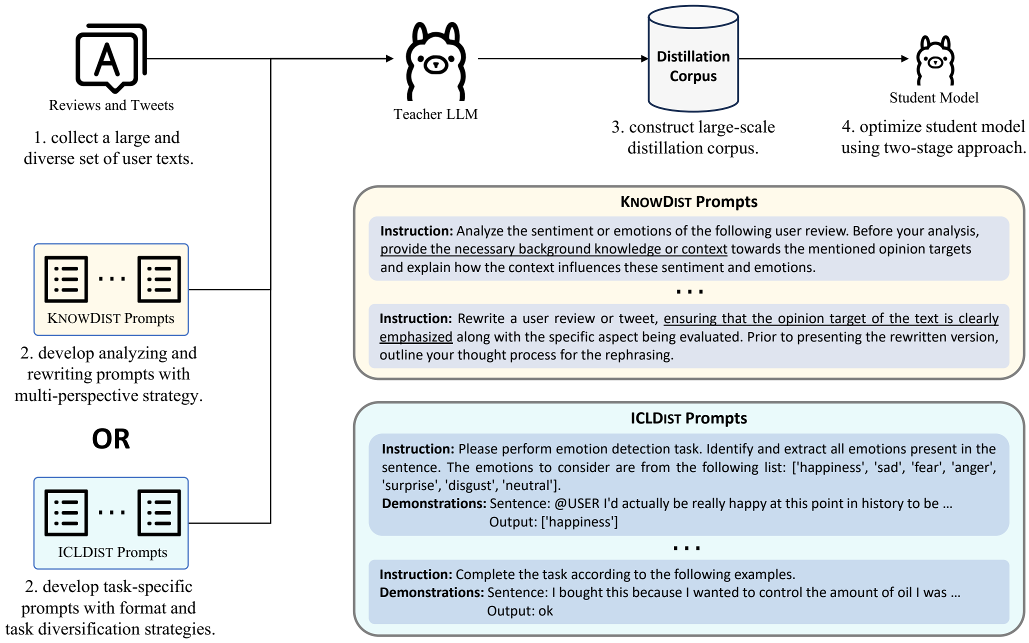

Figure 2: Illustration of our distillation process, consisting of four steps: data collection, prompt construction, corpus generation, and student model optimization.

reviews, and tweets. Secondly, we construct various analyzing and rewriting prompts following our multi-perspective prompting strategy. Thirdly, we apply these prompts to guide the teacher LLM in interpreting existing sentiments within these texts and actively exploring and generating diverse sentiment expression patterns. This process yields a large-scale corpus enriched with sentiment-related knowledge. Finally, we leverage this corpus to optimize the student model, thereby enhancing its fundamental sentiment analysis capabilities.

2.2 In-Context Learning Distillation

After the KnowDist stage, we optimize the student model's alignment in specific sentiment analysis tasks. To achieve this, we construct task-specific prompts comprising instructions, demonstrations, and input text. We then train the student model to mimic the teacher LLM's output on these prompts, aiming to enhance its ability to follow task-specific instructions and demonstrations. However, this method faces a major challenge: we cannot anticipate all potential downstream tasks, making it impossible to prepare corresponding prompts in the ICLDist stage. Consequently, the student model may underperform on previously unseen tasks. For example, when using sentiment classification and emotion recognition as distillation tasks, the student model performs poorly on unseen tasks such as irony detection.

To enhance generalization on unseen tasks, we maximize the diversity of the distillation prompts, introducing **format and task diversification strategies**. Format diversification refers to using varied prompt formats for the same task to mitigate overfitting. We devise three specific strategies to achieve this. The first is to alter label word formats, replacing standard labels like `positive/negative/neutral` with alternatives like `good/bad/ok` or `+1/-1/0`. The second is to diversify label taxonomies, for the emotion recognition task, employing various classification systems, such as Ekman's taxonomy (Ekman, 1992) or the GoEmotions taxonomy (Demszky et al., 2020). The third is to utilize minimized instructions, placing task information within demonstrations, exemplified by prompts like "*Complete the task according to the following examples*".

Task diversification refers to incorporating a variety of tasks other than sentiment analysis during the ICLDist stage. To this end, we select about 100 natural language understanding tasks from the Super-NaturalInstructions dataset (Wang et al., 2022) and construct corresponding prompts. We intentionally exclude sentiment analysis tasks from this selection to prevent overlap with downstream evaluation tasks. While these tasks are not directly related to sentiment analysis, we hypothesize that they can enhance the model's general prompt-following capability.

Task	Dataset	Train	Dev	Test	#Class	Metric
BASIC SENTIMENT ANALYSIS						
Document-level sentiment classification	IMDb	3000	300	1000	2	macro_f1
	Yelp2	3000	300	1000	2	macro_f1
Sentence-level sentiment classification	SST2	3000	300	1821	2	macro_f1
	Twitter17	3000	300	1000	3	macro_f1
MULTIFACETED SENTIMENT ANALYSIS						
Irony detection	Irony18	3000	300	784	2	macro_f1
Emotion recognition	Emotion20	3000	300	1421	4	macro_f1
Stance detection	P-Stance	3000	300	2157	3	macro_f1
Intimacy analysis	MINT-English	1287	300	396	3	macro_f1
FINE-GRAINED SENTIMENT ANALYSIS						
Aspect term sentiment analysis	Rest16	1600	400	676	-	micro_f1
Aspect category sentiment analysis	Rest16	1600	400	676	-	micro_f1
Aspect sentiment quad prediction	Rest16	1264	316	544	-	micro_f1
Structured sentiment analysis	Opener	1744	249	499	-	micro_f1

Table 1: Task overview and dataset statistics in SENTIBENCH. We perform downsampling on some datasets to ensure computational efficiency. For sampling details, please refer to Appendix B.

The ICLDIST process is illustrated in Figure 2. Similar to knowledge collection, we first gather a large volume of user-generated content. Next, we select sentiment classification and emotion recognition as distillation tasks and construct prompts by randomly applying our format diversification strategies. Additionally, we incorporate the task diversification strategy to generate supplementary prompts. We then collect the teacher LLM's responses to these prompts, resulting in a task-alignment corpus. Finally, we optimize the student model on this corpus to enhance its task alignment.

3 SENTIBENCH

To systematically assess LLMs' sentiment analysis capabilities, we develop a comprehensive benchmark. This benchmark encompasses three typical categories: basic sentiment analysis, multifaceted sentiment analysis, and fine-grained sentiment analysis. Multifaceted and fine-grained analyses extend the breadth and depth of evaluation, respectively. For each category, we carefully curate representative tasks and their corresponding datasets. Table 1 provides a comprehensive overview of these tasks along with detailed dataset statistics.

Basic sentiment analysis (BSA) aims to classify the overall sentiment polarity expressed in texts. We collect and curate four widely-adopted sentiment classification datasets, covering both document and sentence levels. For document-level sentiment classification, we incorporate IMDb (Maas et al., 2011) and Yelp2 (Zhang et al., 2015), while for sentence-level classification, we utilize SST2 (Socher et al., 2013) and Twitter17 (Rosenthal et al., 2017).

Multifaceted sentiment analysis (MSA) extends beyond merely identifying sentiment polarity, focusing instead on recognizing a broader range of human emotional states (Zhang et al., 2024b). Our benchmark incorporate four MSA tasks: (1) Irony detection identifies instances whether the intended meaning contradicts the literal expression; (2) Emotion recognition categorizes text into discrete emotional categories, such as anger, joy, sadness, and optimism; (3) Stance detection determines the position or attitude towards a specific target or topic; (4) Intimacy analysis assesses the degree of interpersonal closeness reflected in the text, examining the model's understanding of social information. For these tasks, we curate the following datasets: Irony18 (Van Hee et al., 2018) for irony detection, Emotion20 (Mohammad et al., 2018; Barbieri et al., 2020) for emotion recognition, P-Stance (Li et al., 2021) for stance detection, and MINT-English (Pei et al., 2023) for intimacy analysis.

Fine-grained sentiment analysis (FSA) transcends basic sentiment analysis, aiming to recognize a spectrum of sentiment elements, thereby providing a more complete picture of opinions. Our benchmark incorporates four FSA tasks: (1) Aspect term sentiment analysis (ATSA) extracts aspect terms from the text and determining their sentiment polarities; (2) Aspect category sentiment analysis (ACSA) identifies the evaluated aspect categories and their sentiment polarities; (3) Aspect sentiment quad prediction (ASQP) structures opinions into fine-grained quadruples comprising category, aspect, opinion, and polarity; (4) Structured sentiment analysis (SSA) formalizes opinions as quadruples containing a sentiment holder, target, expression, and polarity. For these tasks, we curate the following datasets: Rest16 (Pontiki et al., 2016; Zhang et al., 2021) for ATSA, ACSA, and ASQP, and Opener (Barnes et al., 2022) for SSA.

Our benchmark is partially inspired by Zhang et al. (2024b). Our work differs from theirs in the following aspects: (1) We develop a reorganized evaluation task taxonomy; (2) Following the revised taxonomy, we refine the tasks and datasets; (3) We conduct comprehensive evaluations across a range of LLMs, with a particular attention to small-scale models.

4 Experiments

4.1 Experimental Setup

Implementation Details. We employ Llama-3.1-70B-Instruct (Grattafiori et al., 2024) as the teacher LLM and Llama-3.2-1.2B-Instruct[2] as the student model. For distillation, we curate a large and diverse corpus of user-generated texts from IMDb (Nguyen et al., 2014), Yelp[3], Amazon[4], and Twitter[5]. We preprocess this corpus by removing texts that overlap with downstream datasets, eliminating duplicates using `simhash`, and filtering out too short or long texts. We then apply the proposed prompting methods to these user texts and obtain 1M KnowDist samples and 400K ICLDist samples. We further supplement the ICLDist corpus with 100K general task samples from the SUPER-NATURALINSTRUCTION (Wang et al., 2022) dataset. Subsequently, we conduct a two-stage optimization of the student model using these two corpora. For the main experiments, we utilize the complete dataset, while for ablation studies, we leverage a subset consisting of 200K KnowDist samples and 100K ICLDist samples. The hyperparameter settings are provided in Appendix C.

After distillation, we evaluate the student model on SentiBench using in-context learning, with dataset statistics shown in Table 1. The specific prompts for each task are provided in Appendix B. During evaluation, we randomly sample 4 examples from the training set as demonstrations. To ensure generation stability, we set the temperature parameter to 0 during model inference. To mitigate the impact of randomness, we conduct each evaluation using 3 different random seeds and report the averaged results.

Baselines. We compare our approach with generic distillation methods. Specifically, we train the student model using existing instruction-following datasets, including the 52K data constructed by Taori et al. (2023) (alpaca-data), and the 2.58M data developed by Wu et al. (2024) (lamini-data). Furthermore, we evaluate a diverse set of models for reference: (1) a fine-tuned T5-base model (Raffel et al., 2020); (2) models from the Llama-3 family, spanning different scales (3.2B, 8B, and 70B variants); (3) several small-scale models ranging from 1B to 3B parameters, including OPT-1.3B (Zhang et al., 2022), TinyLlama-1.1B-Chat-v1.0 (Zhang et al., 2024a), Phi-2-2.7B[6], Qwen-2.5-1.5B-Instruct[7], and Gemma-2-2.6B-it (Team, 2024); and (4) GPT-3.5[8].

4.2 Main Results

Table 2 presents the comparison results on SentiBench. We observe that two generic distillation methods yield only marginal gains in sentiment analysis performance, with the student model showing average F_1-score improvements of 0.81% and 1.43% respectively. These limited improvements suggest that utilizing generic distillation methods to transfer sentiment analysis capabilities is challenging. In contrast, our approach, namely Know & ICLDist, significantly enhances the sentiment analysis performance of the student model. Specifically, our approach achieves an average improvement of over 10%. The most striking improvement is observed in irony detection, where the F_1-score increases dramatically from 35.80% to 73.80% - an improvement of 38.00%. These results demonstrate the effectiveness of our approach in transferring sentiment analysis capabilities from the LLM to its more efficient counterparts.

Furthermore, the experimental results in Table 2 reveal several additional insights. Firstly, within

[2] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/
[3] https://www.yelp.com/dataset
[4] https://nijianmo.github.io/amazon/index.html
[5] https://archive.org/details/twitterstream
[6] https://huggingface.co/microsoft/phi-2
[7] https://qwenlm.github.io/blog/qwen2.5/
[8] Available at https://chat.openai.com/. The specific model used is gpt-3.5-turbo-0125.

Models	BSA				MSA				FSA				Avg
	IMDb	Yelp2	SST2	Twitter	Irony	Emoti.	Stance	Intim.	ATSA	ACSA	ASQP	SSA	
T5-base (Fine-tuned)	92.80	96.62	92.20	65.95	75.18	79.45	72.51	52.43	74.19	72.42	59.45	57.63	74.24
Llama-3-3.2B	92.57	96.53	93.59	61.45	64.00	68.88	71.43	33.32	46.37	51.66	11.09	23.10	59.50
Llama-3-8B	94.17	98.07	95.90	66.58	82.63	73.00	75.86	49.85	54.41	64.57	19.67	31.91	67.22
Llama-3-70B	95.30	98.10	97.14	68.75	83.99	75.87	85.21	53.68	63.78	75.21	31.03	45.29	72.78
GPT-3.5	93.40	97.50	93.57	67.55	65.25	78.14	75.84	55.82	39.96	64.58	30.42	3.41	63.79
OPT-1.3B	78.94	91.37	77.10	39.32	51.18	43.98	53.93	32.65	11.39	19.06	1.72	3.92	42.05
TinyLlama-1.1B	71.27	84.13	78.01	34.21	56.15	50.05	57.25	36.95	26.76	29.42	4.24	13.68	45.18
Phi-2-2.7B	87.03	96.10	90.63	59.59	47.52	45.53	55.36	31.61	39.71	46.54	9.60	16.31	52.13
Qwen-2.5-1.5B	91.92	97.30	92.33	52.39	65.80	63.61	70.90	35.73	37.66	53.25	18.47	20.08	58.29
Gemma-2-2.6B	92.39	97.40	94.17	56.02	70.68	68.85	73.99	42.57	48.00	50.27	18.03	39.08	62.62
Llama-3-1.2B	87.65	94.80	88.93	58.78	35.80	58.07	60.78	25.60	33.80	36.09	8.05	16.91	50.44
+ Distill. *w/* Alpaca-data	89.13	94.37	91.08	58.02	33.01	60.24	64.02	26.10	36.18	37.71	8.72	16.44	51.25(+0.81)
+ Distill. *w/* Lamini-data	89.26	94.63	91.14	62.90	38.05	50.61	63.92	27.90	35.03	41.89	8.30	18.80	51.87(+1.43)
+ KNOWDIST	88.53	95.37	90.80	61.54	44.01	63.49	63.59	31.11	38.75	41.20	10.30	19.62	54.03(+3.59)
+ ICLDIST	92.90	97.63	94.51	68.91	65.35	76.27	70.17	35.15	37.71	48.06	9.76	20.16	59.72(+9.28)
+ KNOW & ICLDIST	93.07	97.70	94.53	68.37	73.80	76.79	69.94	35.39	39.01	47.82	11.69	21.18	60.77(+10.33)

Table 2: Experimental results on SENTIBENCH (F_1-score, %). All results except T5-base are obtained through in-context learning with 4 demonstrations. Our distillation uses 1M KNOWDIST and 500K ICLDIST samples.

the Llama-3 family, we observe a clear positive correlation between model size and performance, with Llama-3-70B achieving the best results, surpassing GPT-3.5 and approaching the fine-tuned T5-base model. Secondly, with Llama-3-70B as the teacher LLM, our approach enables Llama-3-1.2B to achieve comparable performance to the teacher on sentiment classification and emotion recognition. Thirdly, our approach empowers Llama-3-1.2B to outperform Llama-3-3.2B. Moreover, the distilled model demonstrates strong competitive performance compared to other small-scale models and GPT-3.5, also illustrated in Figure 1. Finally, both the distilled model and other small-scale models show inferior performance on intimacy analysis and tuple extraction tasks (*i.e.,* ASQP and SSA). These tasks require a deep understanding of social context and advanced structured extraction capabilities, presenting promising directions for future research.

4.3 Analyis of Two-stage Optimization

Our distillation framework consists of two stages: KNOWDIST and ICLDIST. The results in Table 2 demonstrate that both stages enhance the student model's performance on sentiment analysis tasks, and combining them yields even better results. Below, we conduct an in-depth analysis of these two stages, aiming to distinguish their respective roles and investigate how they complement each other.

Figure 3 illustrates the performance trends of the student model across different volumes of distillation data. We observe that in both stages, model performance generally improves as the data volume increases. Moreover, the improvements brought by ICLDIST are notably more pronounced and efficient. These observations raise two natural questions: (1) Given ICLDIST's superior performance, is the KNOWDIST stage essential to the framework? (2) Could we simplify the framework by merging data from both stages into a unified optimization process?

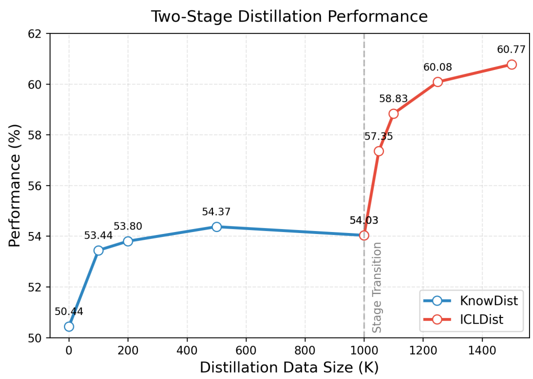

Figure 3: Performance trend of the student model with varying volumes of distillation data (%). Here, performance refers to the average F_1-score on SENTIBENCH.

For the first question, we conduct fine-tuning experiments using the training samples from SENTIBENCH. The results in Table 3 demonstrate that under the fine-tuning setting, both KNOWDIST and ICLDIST can enhance the student model's sentiment analysis performance. Notably, KnowDist achieves more substantial improvements, which contrasts with the in-context learning results in Table 2. These findings support our claims: KnowDist strengthens the student model's fundamental sentiment analysis capabilities, while ICLDist optimizes task alignment. When sufficient labeled samples are available for downstream task alignment, the benefits of ICLDist's task alignment become less significant. However, such labeled data is often scarce in real-world applications. Consequently, both KnowDist and ICLDist are essential components of our framework.

Models	MSA	FSA
Llama-3-1.2B	73.61	68.78
+ KnowDist	76.12(+2.51)	69.70(+0.92)
+ ICLDist	74.64(+1.03)	69.30(+0.52)

Table 3: Experimental results on MSA and FSA categories under fine-tuning settings (F_1-score, %). Models are fine-tuned jointly on all tasks within each category.

For the second question, we conduct experiments to compare unified optimization against two-stage optimization, with results presented in Table 4. The results reveal that unified optimization not only significantly underperforms two-stage optimization but also falls behind using ICLDist alone. This suggests that unified optimization would disrupt the distillation process and impair the learning efficiency of the student model. These findings demonstrate the necessity of two-stage optimization in our framework.

Models	BSA	MSA	FSA
Llama-3-1.2B	82.54	45.06	23.72
+ KnowDist	83.65(+1.01)	50.65(+5.59)	27.11(+3.39)
+ ICLDist	87.83(+5.29)	58.75(+13.69)	27.55(+3.83)
+ Unified	87.21(+4.67)	53.57(+8.51)	27.45(+3.73)
+ Two-stage	88.06(+5.52)	60.70(+15.64)	27.74(+4.02)

Table 4: Comparison results between unified optimization and two-stage optimization (F_1-score, %).

4.4 Ablation Studies

KnowDist. In this stage, we employ two distinct prompting methods (analyzing and rewriting) to elicit sentiment-related knowledge from the teacher LLM and introduce a multi-perspective prompting (MPP) strategy to enhance their effectiveness. As shown in Table 5, the MPP strategy significantly improves the performance of both prompting methods. Specifically, for the analyzing method, MPP yields additional improvements of 3.90% and 1.46% on MSA and FSA, respectively. Among the two prompting methods, the analyzing method achieves more substantial performance gains, while the combination of both methods leads to better overall performance. These results demonstrate the effectiveness of each sub-component within KnowDist.

Dist	Anl	Rw	MPP	BSA	MSA	FSA
✗	-	-	-	82.54	45.06	23.72
✓	✓	✗	✗	83.69(+1.15)	45.72(+0.66)	26.07(+2.35)
✓	✓	✗	✓	83.92(+1.38)	49.62(+4.56)	27.53(+3.81)
✓	✗	✓	✗	83.44(+0.90)	44.98(−0.08)	24.85(+1.13)
✓	✗	✓	✓	82.77(+0.23)	47.90(+2.84)	26.02(+2.30)
✓	✓	✓	✓	83.65(+1.11)	50.65(+5.59)	27.11(+3.39)

Table 5: Ablation results of KnowDist (F_1-score, %). Anl and Rw denote analyzing and rewriting respectively, and MPP stands for multi-perspective prompting.

ICLDist. A key challenge in this stage is the limited generalization to tasks unseen during distillation. To address this challenge, we develop several diversification strategies. As shown in Table 6, without these strategies, the performance improvement on unseen tasks (2.53%) is substantially lower than that on seen tasks (7.71%), confirming our concerns about generalization. After incorporating our diversification strategies, the student model achieves a significant performance gain on unseen tasks (7.79%), reaching a comparable level of improvement to seen tasks. These results demonstrate the effectiveness of our diversification strategies in enhancing model generalization.

Dist	LW	LT	MI	TD	Seen	Unseen
✗	-	-	-	-	77.65	31.00
✓	✗	✗	✗	✗	85.36(+7.71)	33.53(+2.53)
✓	✓	✗	✗	✗	85.18(+7.53)	33.91(+2.91)
✓	✓	✓	✗	✗	85.44(+7.79)	34.07(+3.07)
✓	✓	✓	✓	✗	85.08(+7.43)	35.09(+4.09)
✓	✗	✗	✗	✓	85.64(+7.99)	37.52(+6.52)
✓	✓	✓	✓	✓	85.01(+7.36)	38.79(+7.79)

Table 6: Ablation results of ICLDist (F_1-score, %). LW, LT, and MI denote the format diversification of Label Words, Label Taxonomies, and Minimized Instructions respectively, while TD represents Task Diversification. We divide tasks in SentiBench into seen and unseen categories during distillation, where seen tasks include sentiment classification and emotion recognition, while the rest are considered unseen.

4.5 Discussions

Effect of Teacher LLMs. We experiment with different teacher LLMs in our distillation framework to analyze their impact. The results in Table 7 reveal that teacher quality significantly influences distillation effectiveness, as larger teacher LLMs generally lead to more substantial improvements. Furthermore, we make two noteworthy observations. First, even when using identical models for both teacher and student, distillation has the potential to enhance the student's sentiment analysis performance. This result suggests the potential for leveraging distillation to achieve self-improvement in specialized domains. Second, higher teacher quality does not always correlate with better student performance, as evidenced by the FSA performance when using 8B and 70B models as teachers. This may be related to the capability gap between teachers and students, warranting further exploration in future work.

Teachers	BSA	MSA	FSA
No Distill.	82.54	45.06	23.72
Llama-3-1.2B	80.45(-2.09)	46.33(+1.27)	22.53(-1.19)
Llama-3-3.2B	85.85(+3.31)	51.05(+5.99)	27.59(+3.87)
Llama-3-8B	85.90(+3.36)	57.16(+12.10)	29.02(+5.30)
Llama-3-70B	88.06(+5.52)	60.70(+15.64)	27.74(+4.02)

Table 7: Experimental results using different teacher LLMs in our distillation framework (F_1-score, %).

Results on MMLU. A potential concern of targeted distillation towards specialized capabilities is the possible degradation of the model's general abilities. To investigate this concern, we conduct evaluations on the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021). As shown in Table 8, we find that our distillation approach not only avoids any deterioration but also results in a slight improvement. This indicates that our distillation approach can enhance specialized capabilities without compromising general capabilities.

Models	Human.	Social.	STEM	Other	Avg
Llama-3-1.2B	42.87	51.16	39.68	52.11	46.12
+ KNOW & ICLDIST	43.14	52.62	40.17	53.40	46.94

Table 8: Experimental results on 5-shot MMLU (accuracy, %). Our evaluation is conducted using LM-Evaluation-Harness provided at `https://github.com/meta-llama/llama-cookbook`.

5 Related Work

Applying LLMs for Sentiment Analysis. Many researchers adopt in-context learning methods to harness LLMs for sentiment analysis tasks (Zhang et al., 2024b; Wang et al., 2024c; Bai et al., 2024; Xu et al., 2023a). To enhance the effectiveness of in-context learning, research has branched into (1) selecting semantically relevant examples for demonstrations (Wang et al., 2024b; Xu et al., 2024a; Wang et al., 2024a), (2) utilizing chain-of-thought reasoning to enhance sentiment inference (Fei et al., 2023), and (3) integrating relevant background knowledge to generate more nuanced and informed predictions (Zhang et al., 2023). Furthermore, a range of studies explore fine-tuning methods to better align LLMs with sentiment analysis tasks (Fatemi and Hu, 2023; Šmíd et al., 2024; Simmering and Huoviala, 2023).

Knowledge Distillation from LLMs. In light of the high computational demands or issues of proprietary access, many studies explore knowledge distillation techniques (Hinton et al., 2015) to transfer the capabilities of LLMs into more compact and accessible models (Taori et al., 2023; Chiang et al., 2023; Wu et al., 2024; Chen et al., 2024; Muralidharan et al., 2024). Recent advancements in this field concentrate on optimizing distillation objectives to improve the efficiency and effectiveness of the distillation process (Zhong et al., 2024; Gu et al., 2024; Ko et al., 2024; Agarwal et al., 2024). Besides, there is a growing trend towards distilling specialized capabilities from LLMs, including leveraging LLMs as annotators to generate pseudo-labeled data (Ding et al., 2023; Xu et al., 2023b; Kim et al., 2024; Zhou et al., 2024; He et al., 2024) and synthesizing task-specific data from scratch (Ye et al., 2022; He et al., 2023; Gao et al., 2023; Xu et al., 2024b).

6 Conclusions

This paper explores targeted distillation for sentiment analysis, introducing a two-stage distillation framework. The first stage (KNOWDIST) aims to transfer fundamental sentiment analysis capabilities, while the second stage (ICLDIST) focuses on transfering task-specific prompt-following abilities. Experimental results demonstrate that our framework enables Llama-3-1.2B model to surpass the sentiment analysis performance of Llama-3-3.2B and show strong competitiveness compared to other small-scale models.

Limitations

We list the potential limitations of this paper:

- Our approach transfers knowledge directly from teacher LLMs without filtering or processing their responses. This direct transfer may propagate erroneous or low-quality information to the student model, potentially impacting its performance. Future work could explore quality control mechanisms during the distillation process.
- As shown in Table 2, our model exhibits unsatisfactory performance on tuple extraction tasks (*i.e.*, ASQP and SSA). This suggests the need for specialized optimization of structured extraction capabilities.

We believe that these limitations offer promising directions for future research.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

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

Francesco Barbieri, Jose 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*, pages 1644–1650, Online. Association for Computational Linguistics.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Hongzhan Chen, Ruijun Chen, Yuqi Yi, Xiaojun Quan, Chenliang Li, Ming Yan, and Ji Zhang. 2024. Knowledge distillation of black-box large language models. *Preprint*, arXiv:2401.07013.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

P. Ekman. 1992. Are there basic emotions? *Psychological Review*, 99:550–553.

Sorouralsadat Fatemi and Yuheng Hu. 2023. A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis. *Preprint*, arXiv:2312.08725.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, WEIZHONG ZHANG, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang,

Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara

Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *Preprint*, arXiv:2305.15717.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. *Preprint*, arXiv:2303.16854.

Zexue He, Marco Tulio Ribeiro, and Fereshte Khani. 2023. Targeted data generation: Finding and fixing model weaknesses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8506–8520, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. DistiLLM: Towards streamlined distillation for large language models. In *Forty-first International Conference on Machine Learning*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5.

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 *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Saif 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*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Bhuminand Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Compact language models via pruning and knowledge distillation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dai Quoc Nguyen, Dat Quoc Nguyen, Thanh Vu, and Son Bao Pham. 2014. Sentiment classification on polarity reviews: An empirical study using rating-based features. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 128–135.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozen Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. SemEval-2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2235–2246, Toronto, Canada. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique

Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Paul F. Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *Preprint*, arXiv:2310.18025.

Jakub Šmíd, Pavel Priban, and Pavel Kral. 2024. LLaMA-based models for aspect-based sentiment analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Bangkok, Thailand. Association for Computational Linguistics.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemma Team. 2024. Gemma.

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*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Qianlong Wang, Keyang Ding, Xuan Luo, and Ruifeng Xu. 2024a. Improving in-context learning via sequentially selection and preference alignment for few-shot aspect-based sentiment analysis. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2462–2466, New York, NY, USA. Association for Computing Machinery.

Qianlong Wang, Hongling Xu, Keyang Ding, Bin Liang, and Ruifeng Xu. 2024b. In-context example retrieval from multi-perspectives for few-shot aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8975–8985, Torino, Italia. ELRA and ICCL.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024c. Is chatGPT a good sentiment analyzer? In *First Conference on Language Modeling*.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024. LaMini-LM: A diverse herd of distilled models from large-scale instructions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–964, St. Julian's, Malta. Association for Computational Linguistics.

Hongling Xu, Qianlong Wang, Yice Zhang, Min Yang, Xi Zeng, Bing Qin, and Ruifeng Xu. 2024a. Improving in-context learning with prediction feedback for sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3879–3890, Bangkok, Thailand. Association for Computational Linguistics.

Hongling Xu, Yice Zhang, Qianlong Wang, and Ruifeng Xu. 2024b. Ds^2-absa: Dual-stream data synthesis with label refinement for few-shot aspect-based sentiment analysis. *Preprint*, arXiv:2412.14849.

Xiancai Xu, Jia-Dong Zhang, Rongchang Xiao, and Lei Xiong. 2023a. The limits of chatgpt in extracting aspect-category-opinion-sentiment quadruples: A comparative analysis. *Preprint*, arXiv:2310.06502.

Yichong Xu, Ruochen Xu, Dan Iter, Yang Liu, Shuohang Wang, Chenguang Zhu, and Michael Zeng. 2023b. InheritSumm: A general, versatile and compact summarizer by distilling from GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13879–13892, Singapore. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 349–356, New York, NY, USA. Association for Computing Machinery.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024a. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

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

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

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024b. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. 2024. Revisiting knowledge distillation for autoregressive language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10900–10913, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

Analyzing - BASIC Analyze the overall sentiment of the following text. Provide a brief explanation supporting your conclusion. Text: {Text}
Analyzing - TARGET Given a text, list the mentioned opinion targets, analyzing the evaluated aspects and the corresponding sentiments. Provide brief explanations supporting your conclusions. Text: {Text}
Analyzing - EXPRESSION Identify all sentiment expressions in the following text, i.e., those words or phrases that convey sentiment or emotion. For each sentiment expression, provide a clear explanation of how it contributes to the overall sentiment. Text: {Text}
Analyzing - EMOTION Analyze the following text and identify any emotions being expressed, such as happiness, sadness, anger, fear, surprise, or disgust. For each emotion identified, explain how it is reflected in the text. Text: {Text}
Analyzing - BACKGROUND Analyze the sentiment or emotions of the following text. Before your analysis, provide the necessary background knowledge or context towards the mentioned opinion targets and explain how the context influences these sentiment and emotions. Text: {Text}

Table 9: Analyzing prompts in KNOWDIST.

Rewriting - BASIC Rewrite the following text to ensure it retains the original sentiment and tone, but presents it in a rephrased or alternative way. Prior to presenting the rewritten version, outline your thought process for the rephrasing. Text: {Text}
Rewriting - TARGET Rewrite the following text, ensuring that the opinion target of the text is clearly emphasized along with the specific aspect being evaluated. Prior to presenting the rewritten version, outline your thought process for the rephrasing. Text: {Text}
Rewriting - EXPRESSION Rewrite the following text while focusing on the sentiment expressions used. Prior to presenting the rewritten version, outline your thought process for the rephrasing. Text: {Text}
Rewriting - EMOTION Rewrite the following text by highlighting the expressed emotions (such as happiness, sadness, anger, fear, surprise, or disgust). Prior to presenting the rewritten version, outline your thought process for the rephrasing. Text: {Text}
Rewriting - BACKGROUND Rewrite the following text to enhance sentiment clarity by incorporating necessary background knowledge or context. Prior to presenting the rewritten version, outline your thought process for the rephrasing. Text: {Text}

Table 10: Rewriting prompts in KNOWDIST.

Organization of Appendices

We structure the appendix into four sections:

- Appendix A details the complete prompts utilized in our distillation framework;
- Appendix B provides the construction details and evaluation prompts of SENTIBENCH;
- Appendix C outlines the hyperparameter settings of the two-stage optimization; and
- Appendix D presents the case study of the distilled model.

A Distillation Prompts

A.1 Prompts in Knowledge-Driven Distillation

In this stage, we develop two distinct prompting methods (analyzing and rewriting) along with a multi-perspective prompting strategy. The corresponding prompts for these methods are presented in Tables 9 and 10.

A.2 Prompts in In-Context Learning Distillation

In this stage, we employ sentiment classification and emotion recognition as distillation tasks and devise multiple strategies to enhance prompt diversity, including label word (LW) diversification, label taxonomies (LT) diversification, and minimized instruction (MI) strategies. Tables 11 and 12 present the specific prompts. In practice, these prompts contain a random number of demonstrations ranging from 1 to 16. These tables only show examples with one demonstration.

B SENTIBENCH Details

B.1 Dataset Downsampling

For computational efficiency, we sample from the original datasets. Specifically:

- For basic sentiment analysis tasks, we randomly sample 3000 instances from each training set of IMDb, Yelp2, SST2, and Twitter17. For validation, we randomly sample

Sentiment Classification - BASIC
Please perform sentiment classification task. The label should be one of the following: [‘positive’, ‘negative’, ‘neutral’]. In your classification, consider the overall content, tone, emotional language, and any contextual clues that indicate the sentiment behind the sentence. Do not provide any reasoning or explanation and directly output the final answer.

Sentence: I bought this because I wanted to control the amount of oil I was using. I read the other reviews and the ...
Output: neutral

Sentence: A fabulous social commentary is illustrated between the lines that you can enjoy privately in your mind while ...
Output:

Sentiment Classification - LW
Please perform sentiment classification task. The label should be one of the following: [‘+1’, ‘-1’, ‘0’]/[‘POS’, ‘NEG’, ‘NEU’]/[‘good’, ‘bad’, ‘ok’]. In your classification, consider the overall content, tone, emotional language, and any contextual clues that indicate the sentiment behind the sentence. Do not provide any reasoning or explanation and directly output the final answer.

Sentence: I bought this because I wanted to control the amount of oil I was using. I read the other reviews and the ... Output: 0

Sentence: If your planting several rows of garden veggies, ie: corn beans, etc, this is a great time saver. You must make ...
Output:

Sentiment Classification - MI
Please complete the task according to the following examples. Do not provide any reasoning or explanation and directly output the final answer.

Sentence: I couldn’t use this cable. But it is not the fault of the cable. I ordered it to use with my new kodak printer. I ...
Output: neutral

Sentence: This is a good family game, easy to learn, and straightforward to play. Also helpful in teaching US geography ...
Output:

Table 11: Sentiment classification prompts in ICLDIST.

300 instances from each validation set of these datasets. For testing, we randomly sample 1,000 instances each from the test sets of IMDb, Yelp2, and Twitter17, while retaining the original test set for SST2 due to its smaller size.

- For multifaceted sentiment analysis tasks, we randomly sample 3000 instances each from the training sets of Irony18, Emotion20, and P-Stance. For validation, we randomly sample 300 instances from each validation set of these four datasets. Due to their limited sizes, we retained all original test sets for these tasks.

- For fine-grained sentiment analysis tasks, we retain all original datasets due to their limited sizes.

Emotion Recognition - BASIC
Please perform emotion detection task. Identify and extract all emotions present in the sentence. The emotions to consider are from the following list: [‘happiness’, ‘sad’, ‘fear’, ‘anger’, ‘surprise’, ‘disgust’, ‘neutral’]. In your analysis, take into account the language used, context, and any emotional expressions or cues that indicate multiple emotions. Do not provide any reasoning or explanation and directly output the final answer.

Sentence: I just received a pair 38x30 VIP and they were a bit loose around the waste, and the legging was long enough ...
Output: [‘disgust’, ‘neutral’, ‘sadness’]

Sentence: First, the title is misleading. One might expect a book called Štumbling on happinessẗo perhaps provide ...
Output:

Emotion Recognition - LT
Please perform emotion detection task. Identify and extract all emotions present in the sentence. The emotions to consider are from the following list: [‘neutral’, ‘curiosity’, ‘confusion’, ‘amusement’, ‘gratitude’, ‘admiration’, ‘pride’, ‘approval’, ‘realization’, ‘surprise’, ‘excitement’, ‘joy’, ‘relief’, ‘caring’, ‘optimism’, ‘desire’, ‘love’, ‘fear’, ‘nervousness’, ‘grief’, ‘sadness’, ‘remorse’, ‘disapproval’, ‘disappointment’, ‘anger’, ‘annoyance’, ‘embarrassment’, ‘disgust’]. In your analysis, take into account the language used, context, and any emotional expressions or cues that indicate multiple emotions. Do not provide any reasoning or explanation and directly output the final answer.

Sentence: Let me start by saying that I have read as many Agatha Christie books as I possibly could. Sad Cypress ...
Output: [‘curiosity’, ‘admiration’, ‘surprise’, ‘disappointment’, ‘disapproval’]

Sentence: I put this in my Garage and the humidity that comes out of the end is good for the wood in this kind of ...
Output:

Emotion Recognition - MI
Please complete the task according to the following examples. Do not provide any reasoning or explanation and directly output the final answer.

Sentence: I really don’t get how this game got such good ratings. My only guess is that people just like game of ...
Output: [‘disgust’, ‘neutral’, ‘anger’]

Sentence: This wonderful allegory is highly entertaining for a young person and deeply inspiring for an adult who is ...
Output:

Table 12: Emotion recognition prompts in ICLDIST.

BSA - IMDb
Please perform Sentiment Analysis task. Given the sentence, assign a sentiment polarity label from [‘negative’, ‘positive’]. Return label only without any other text.

Sentence: I have to agree with MR. Caruso Jr Lanza,s was the finest voice god had to offer if only he could have ...
Label: positive

Sentence: I watched this film with a bunch of friends at a Halloween party last night. I got to say that the ...
Label:

BSA - Yelp2
Please perform Sentiment Analysis task. Given the sentence, assign a sentiment polarity label from [‘negative’, ‘positive’]. Return label only without any other text.

Sentence: I’m so glad Yelp has added verbal descriptions for the star system as, "Meh. I’ve experienced better." ...
Label: negative

Sentence: We went here yesterday for lunch, it wasnt packed at all and the lunch prices are good. They start you off ...
Label:

BSA - SST2
Please perform Sentiment Analysis task. Given the sentence, assign a sentiment polarity label from [‘negative’, ‘positive’]. Return label only without any other text.

Sentence: as relationships shift , director robert j. siegel allows the characters to inhabit their world without ...
Label: positive

Sentence: this is one of polanski ’s best films .
Label:

BSA - Twitter17
Please perform Sentiment Analysis task. Given the sentence, assign a sentiment polarity label from [‘negative’, ‘positive’, ‘neutral’]. Return label only without any other text.

Sentence: "It’s 4.33am, I can’t sleep. Just bought two pairs of sun glasses online n caught up on Hulk Hogan news ...
Label: positive

Sentence: @user Bull vs Corbin is the gold standard for bad no DQ matches, this was a close second.
Label:

Table 13: The prompts for basic sentiment analysis (BSA) task.

B.2 Task Prompts

SENTIBENCH comprises three task categories: basic sentiment analysis (BSA), multifaceted sentiment analysis (MSA), and fine-grained sentiment analysis (FSA). The corresponding prompts for these tasks are presented in Tables 13, 14, and 15.

MSA - Irony Detection - Irony18
Please perform Irony Detection task. Given the sentence, assign a sentiment label from [‘irony’, ‘non-irony’]. Return label only without any other text.

Sentence: @user I infer that you are besmirching coffee, but that can’t be right
Label: non-irony

Sentence: Just walked in to #Starbucks and asked for a "tall blonde" Hahahaha
Label:

MSA - Emotion Recognition - Emotion20
Please perform Emotion Detection task. Given the sentence, assign a emotion label from [‘anger’, ‘joy’, ‘sadness’, ‘optimism’]. Return the label only without any other text.

Sentence: it’s pretty depressing when u hit pan on ur favourite highlighter
Label: sadness

Sentence: @user Interesting choice of words... Are you confirming that governments fund #terrorism? Bit of an open door, but still...
Label:

MSA - Stance Detection - P-Stance
Please perform Stance Detection task. Given the sentence, assign a sentiment label expressed by the author towards "Bernie Sanders" from [‘against’, ‘favor’]. Return label only without any other text.

Sentence: ? seriously - no hate but what leadership . dude is loosing sensibility and MIA. Bernie though has ...
Label: favor (opinion towards ‘Bernie Sanders’)

Sentence: He’s the ONLY ONE Where have I heard that before? No, Bernie is NOT the only one The Democrats ...
Label:

MSA - Intimacy Analysis - MINT-English
Please perform Intimacy Detection task. Given the sentence, assign an intimacy label from [‘not intimate’, ‘moderately intimate’, ‘highly intimate’]. Return label only without any other text.

Sentence: Would God be pleased if you were working to hasten the apocalypse?
Label: not intimate

Sentence: @tessavirtue Happy new year!!!! Love u
Label:

Table 14: The prompts for multifaceted sentiment analysis (MSA) task.

C Hyperparameter Settings of Distillation

The detailed hyperparameters of two-stage optimization are listed in Tables 16 and 17.

FSA - ATSA - Rest16
Please perform Aspect Term Sentiment Analysis task. Given the sentence, extract all (aspect term, sentiment polarity) pairs.

Sentence: I had the best ravioli ever.
Label: [(‘ravioli’, ‘positive’)]

Sentence: Green Tea creme brulee is a must!
Label:

FSA - ACSA - Rest16
Please perform aspect-level sentiment analysis task. Given the sentence, tag all (aspect category, sentiment) pairs. Aspect category should be selected from [‘ambience general’, ‘drinks prices’, ‘drinks quality’, ‘drinks style_options’, ‘food prices’, ‘food quality’, ‘food style_options’, ‘location general’, ‘restaurant general’, ‘restaurant miscellaneous’, ‘restaurant prices’, ‘service general’], and sentiment should be selected from [‘negative’, ‘neutral’, ‘positive’]. If there are no target-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: I pray it stays open forever.
Label: [(‘restaurant general’, ‘positive’)]

Sentence: Serves really good sushi.
Label:

FSA - ASQP - Rest16
Please perform Aspect Sentiment Quad Prediction task. Given the sentence, extract all (aspect term, aspect category, opinion, sentiment polarity) quadruples.
1. Aspect 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’].
2. Sentiment polarity should be selected from [‘negative’, ‘neutral’, ‘positive’].
3. If there is no aspect term, use ‘NULL’ as the aspect term. Only aspect term can be ‘NULL’, aspect category, opinion and sentiment polarity CANNOT be ‘NULL’.
4. Please return python list only, without any other comments or texts.

Sentence: Make sure you try this place as often as you can .
Label: [(‘restaurant general’, ‘place’, ‘try’, ‘positive’)]

Sentence: All their menu items are a hit , and they serve mimosas
Label:

FSA - SSA - Opener
Please perform the Structured Sentiment Analysis task. Given a sentence, extract all opinion tuples in the format (holder, target, sentiment expression, sentiment polarity).
Each tuple should contain:
- Holder: The entity expressing the sentiment, if there is no explicit holder, use ‘NULL’ as the holder.
- Target: The entity being evaluated, if there is no explicit target, use ‘NULL’ as the target.
- Sentiment Expression: The phrase conveying the sentiment.
- Sentiment Polarity: The polarity of the sentiment, either positive, negative, or neutral.

Follow these rules:
1. If there is no sentiment expression, return ‘NULL’ for all fields.
2. Please return python list only, without any other comments or texts.

Sentence: A beautiful wellness hotel
Label: [(‘NULL’, ‘wellness hotel’, ‘beautiful’, ‘positive’]

Sentence: We went foor a cheap city trip and that ’s what we have got .
Label:

Table 15: The prompts for fine-grained sentiment analysis (FSA) task.

D Case Study

Figures 4 and 5 present representative examples to demonstrate the basic sentiment analysis capa-

Prompt

Given a user review, list the mentioned opinion targets, analyzing the evaluated aspects and the corresponding sentiments. Provide brief explanations supporting your conclusions.

Review: My wife and wanted some silverware with slightly rounded edge. We had viewed these in person at a store first. They feel great in the hands and are easy to rotate. They have survived several times through the dishwasher so far. The also stack pretty good in our draw even though they are slightly rounded.

Llama-3-70B's Response

Here are the opinion targets, evaluated aspects, and corresponding sentiments extracted from the review:

1. **Opinion Target: Silverware**
   * Evaluated Aspect: Edge shape
   * Sentiment: Positive (The reviewer and their wife wanted silverware with slightly rounded edges, indicating a preference for this feature)
2. **Opinion Target: Silverware**
   * Evaluated Aspect: Ergonomics
   * Sentiment: Positive (The reviewer mentions that the silverware "feel great in the hands", indicating comfort and ease of use)
3. **Opinion Target: Silverware**
   * Evaluated Aspect: Ease of rotation
   * Sentiment: Positive (The reviewer states that the silverware is "easy to rotate", suggesting that it is convenient to use)
4. **Opinion Target: Silverware**
   * Evaluated Aspect: Durability
   * Sentiment: Positive (The reviewer mentions that the silverware has "survived several times through the dishwasher", indicating that it is resistant to wear and tear)
5. **Opinion Target: Silverware**
   * Evaluated Aspect: Storage
   * Sentiment: Positive (The reviewer notes that the silverware "stack pretty good in our draw", despite being slightly rounded, indicating that it is convenient to store)

Overall, the reviewer is very satisfied with the silverware, praising its design, ergonomics, durability, and storage convenience.

Our KnowDist Model's Response

Here are the opinion targets, evaluated aspects, and corresponding sentiments extracted from the review:

1. **Opinion Target: Silverware**
   * Evaluated Aspect: Edge shape
   * Sentiment: Positive
   * Explanation: The reviewer mentions that the silverware has a "slightly rounded edge", indicating a positive sentiment towards the shape of the edge.
2. **Opinion Target: Silverware**
   * Evaluated Aspect: Comfort and ergonomics
   * Sentiment: Positive
   * Explanation: The reviewer states that the silverware "feel great in the hands", indicating a positive sentiment towards the comfort and ergonomics of the product.
3. **Opinion Target: Silverware**
   * Evaluated Aspect: Durability
   * Sentiment: Positive
   * Explanation: The reviewer mentions that the silverware "survived several times through the dishwasher", indicating a positive sentiment towards the durability of the product.
4. **Opinion Target: Silverware**
   * Evaluated Aspect: Stackability
   * Sentiment: Positive
   * Explanation: The reviewer states that the silverware "stack pretty good in our draw", indicating a positive sentiment towards the stackability of the product.
5. **Opinion Target: Silverware**
   * Evaluated Aspect: Overall quality
   * Sentiment: Positive
   * Explanation: The reviewer concludes that the silverware "feel great in the hands and are easy to rotate", indicating a positive sentiment towards the overall quality of the product.

Overall, the reviewer is very satisfied with the silverware, praising its comfort, durability, and stackability.

Llama-3-1.2B's Response

Based on the review, the following opinion targets are mentioned:

1. **Silverware quality**: The reviewer mentions the quality of the silverware, specifically its feel, durability, and performance in the dishwasher.
2. **Design and aesthetics**: The reviewer notes the design of the silverware, mentioning its rounded edge and ease of rotation.
3. **Durability**: The reviewer highlights the silverware's ability to withstand multiple wash cycles and its ability to stack well in the drawer.
4. **Ease of use**: The reviewer mentions the ease of use of the silverware, particularly in terms of its comfort in the hands and ease of rotation.

Analyzing the evaluated aspects and corresponding sentiments, we can conclude that the reviewer is generally satisfied with the silverware's quality, design, and durability. The reviewer's use of positive adjectives such as "great", "easy", and "comfortable" suggests a high level of satisfaction with these aspects.

The reviewer's sentiment towards the silverware is overwhelmingly positive, with no negative comments or criticisms mentioned. The reviewer's tone is also enthusiastic and encouraging, with phrases such as "We had viewed these in person at a store first" and "They feel great in the hands" indicating a strong endorsement of the product.

Overall, the review suggests that the reviewer is very pleased with the silverware and would recommend it to others.

Figure 4: Case study (1).

Hyper-parameter	Value
Batch Size	128
Learning Rate	5e-6
Training Epoch	4
Learning Rate Deacy	Cosine
Warmup Step Ratio	0.01
Weight Decay	0.1
Adam β_1	0.9
Adam β_2	0.95

Table 16: Hyperparameters for KnowDist's optimization.

bilities of Llama-3-70B, our model, and Llama-3-1.2B. Among the three models, Llama-3-70B achieves the best analysis results, followed by our model, while Llama-3-1.2B shows the weakest performance. The key differences are reflected in three aspects: the accuracy and comprehensiveness of the analysis results, as well as the depth of reasoning.

Hyper-parameter	Value
Batch Size	128
Learning Rate	1e-5
Training Epoch	4
Learning Rate Deacy	Linear
Warmup Step Ratio	0.02
Weight Decay	0.01
Adam β_1	0.9
Adam β_2	0.999

Table 17: Hyperparameters for ICLDist's optimization.

Prompt

Identify all sentiment expressions in a user review, i.e., those words or phrases that convey sentiment or emotion. For each sentiment expression, provide a clear explanation of how it contributes to the overall sentiment.

Review: Thrillers involving serial killers are usually my favorites, but this movie was God awful. I watched it on the strength of the description and Robert Downey, Jr being one of my favorite actors. It is a convoluted mess and not nearly as clever as it tries to be. Only watch if you need a sleep aid.

Llama-3-70B's Response

Here are the sentiment expressions in the review with explanations of how they contribute to the overall sentiment:

1. **"God awful"**: This phrase is a strong negative sentiment expression that indicates the reviewer's extreme dislike for the movie. The use of "God" adds emphasis to the intensity of their negative emotions.
2. **"Convuluted mess"**: This phrase is a negative sentiment expression that conveys the reviewer's frustration and disappointment with the movie's plot. The word "convoluted" implies that the plot is overly complicated and difficult to follow, while "mess" suggests a lack of coherence or organization.
3. **"Not nearly as clever as it tries to be"**: This phrase is a negative sentiment expression that implies the reviewer was unimpressed by the movie's attempts to be intelligent or clever. The phrase suggests that the movie failed to meet the reviewer's expectations and came across as pretentious.
4. **"Only watch if you need a sleep aid"**: This phrase is a sarcastic sentiment expression that implies the movie is so boring that it could put someone to sleep. The reviewer is using humor to express their strong negative opinion of the movie.

The overall sentiment of the review is overwhelmingly negative. The reviewer's use of strong negative language, such as "God awful" and "convoluted mess", emphasizes their dislike for the movie. The only positive sentiment expression in the review is the mention of Robert Downey, Jr. being one of their favorite actors, but this is not enough to redeem the movie in their opinion.

Our KNOWDIST Model's Response

Here are the sentiment expressions in the review with explanations of how they contribute to the overall sentiment:

1. **"God awful"**: This phrase is a strong negative sentiment expression that sets the tone for the rest of the review. It implies that the movie is extremely bad, which contributes to the overall negative sentiment.
2. **"convoluted mess"**: This phrase is a negative sentiment expression that describes the movie's plot as confusing and difficult to follow. It reinforces the idea that the movie is not well-made.
3. **"not nearly as clever as it tries to be"**: This phrase is a negative sentiment expression that implies that the movie's attempts to be clever are unsuccessful. It suggests that the movie's writers or directors did not put enough thought into the plot or characters.
4. **"Only watch if you need a sleep aid"**: This phrase is a negative sentiment expression that implies that the movie is so bad that it's almost a necessity to watch it just to get a good night's sleep. It's a sarcastic comment that reinforces the reviewer's negative opinion of the movie.

Overall, the sentiment of the review is extremely negative. The reviewer uses strong language to express their dislike for the movie, calling it "God awful" and describing it as a "convoluted mess" and a "sleep aid". The only positive sentiment expressed is a nod to Robert Downey, Jr. as a favorite actor, but even that is not enough to redeem the movie.

Llama-3-1.2B's Response

Here are the sentiment expressions in the review, along with explanations of how they contribute to the overall sentiment:

1. **"God awful"** - This is an extremely negative expression that conveys strong disgust and dislike. It's a strong insult, implying that the movie is completely terrible and not worth watching.
2. **"convoluted mess"** - This word conveys a sense of confusion and frustration, implying that the movie is overly complicated and difficult to follow. It's a negative expression that suggests the movie is not well-made or engaging.
3. **"clever"** - This word is the opposite of "convoluted", implying that the movie is not clever or sophisticated. It's a negative expression that suggests the movie is not well-written or thought-provoking.
4. **"Only watch if you need a sleep aid"** - This is a sarcastic expression that implies the movie is so bad that it's almost a necessity to watch it just to get some rest. It's a negative expression that suggests the movie is not worth watching, but is necessary in a desperate situation.

Overall, the review is extremely negative, with a strong emphasis on criticizing the movie's plot, acting, and overall quality. The reviewer's use of strong negative language and sarcastic tone suggests that they are not a fan of the movie.

Figure 5: Case study (2).