

# RVISA: Reasoning and Verification for Implicit Sentiment Analysis

Wenna Lai , *Student Member, IEEE*, Haoran Xie , *Senior Member, IEEE*, Guandong Xu , *Member, IEEE*, and Qing Li , *Fellow, IEEE*

**Abstract**—Under the context of the increasing social demand for fine-grained sentiment analysis (SA), implicit sentiment analysis (ISA) poses a significant challenge owing to the absence of salient cue words in expressions. Thus, reliable reasoning is required to understand how sentiment is evoked, enabling the identification of implicit sentiments. In the era of large language models (LLMs), encoder-decoder (ED) LLMs have emerged as popular backbone models for SA applications, given their impressive text comprehension and reasoning capabilities across diverse tasks. In comparison, decoder-only (DO) LLMs exhibit superior natural language generation and in-context learning capabilities. However, their responses may contain misleading or inaccurate information. To accurately identify implicit sentiments with reliable reasoning, this study introduces a two-stage reasoning framework named Reasoning and Verification for Implicit Sentiment Analysis (RVISA), which leverages the generation ability of DO LLMs and reasoning ability of ED LLMs to train an enhanced reasoner. The framework involves three-hop reasoning prompting to explicitly furnish sentiment elements as cues. The generated rationales are then used to fine-tune an ED LLM into a skilled reasoner. Additionally, we develop a straightforward yet effective answer-based verification mechanism to ensure the reliability of reasoning learning. Evaluation of the proposed method on two benchmark datasets demonstrates that it achieves state-of-the-art performance in ISA.

**Index Terms**—Implicit sentiment analysis, large language models, multi-task learning, chain-of-thought.

## I. INTRODUCTION

**S**ENTIMENT analysis (SA) aims to extract opinions, sentiments, and emotions through computational methods [1]. Nowadays, people are more inclined to express and share their ideas about day-to-day activities and global issues online. With the increasing demand on social media, SA has gained significant interest considering great commercial value in exploring customer opinions or sentiments from user reviews or

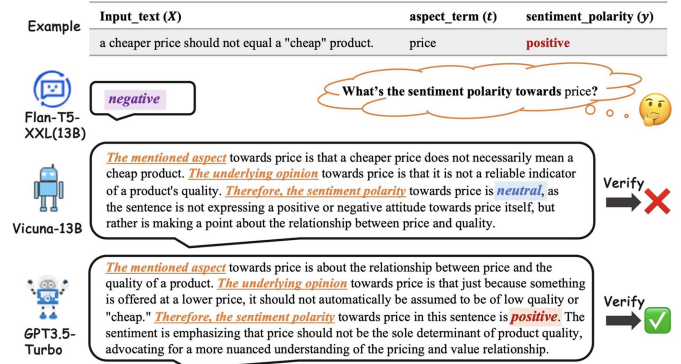


Fig. 1. Different LLMs demonstrate diverse reasoning abilities for implicit sentiment analysis. It is conducive to detecting implicit sentiment polarities by explicitly inferring sentiment elements as rationale but verification is required to ensure reliability.

other sources, while also assisting learning, communication, decision-making, and situation awareness in human-centric environments [2]. Traditionally, SA is classified into three levels: document-level, sentence-level, and aspect-level [3]. Among these, aspect-based sentiment analysis (ABSA) can extract fine-grained opinions regarding a given aspect or entity. Considering context information, ABSA can be further classified into implicit sentiment analysis (ISA) and explicit sentiment analysis (ESA), where ISA involves expressions that lack explicit polarity markers but still deliver human-aware sentiment polarity [4].

In 2021, Li et al. [5] split the SemEval-2014 Restaurant and Laptop benchmarks into explicit sentiment expression and implicit sentiment expression slices based on the presence of opinion words, drawing attention to ISA in ABSA tasks. Fei et al. [6] conducted pre-experiments on 20 existing sentiment classifiers and found that traditional methods exhibit inferior performance for the same implicit case. Most traditional classifiers tend to overlook the ISA problem and address it superficially. While humans can easily grasp real intent and perceive changes in mood through common sense and reasoning, models typically find it more challenging to address ISA than ESA due to limited contextual information and insufficient reasoning skills.

With recent advancements, large language models (LLMs) have demonstrated impressive complex reasoning abilities with chain-of-thought (CoT) prompting [7], [8] and in-context learning [9], driving their increased use in downstream applications [10], [11], [12]. Zhang et al. [13] investigated the

Received 4 July 2024; revised 9 December 2024; accepted 29 January 2025. Date of publication 3 February 2025; date of current version 15 September 2025. This work was supported by the Faculty Research at Lingnan University, Hong Kong under Grant DB24A4 and Grant DB24C5. Recommended for acceptance by R. Mao. (Corresponding author: Haoran Xie.)

Wenna Lai and Qing Li are with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, SAR, China (e-mail: winnelai05@gmail.com; qing-prof.li@polyu.edu.hk).

Haoran Xie is with the School of Data Science, Lingnan University, Hong Kong, SAR, China (e-mail: hrxie@iee.org).

Guandong Xu is with the School of Computer Science and the Data Science Institute, University of Technology Sydney, Sydney, NSW 2007, Australia, and also with the Education University of Hong Kong, Hong Kong, SAR, China (e-mail: gdxu@eduhk.hk).

Digital Object Identifier 10.1109/TAFFC.2025.3537799

performance of LLMs in prompt-based inference and observed that for tasks requiring structured sentiment outputs, such as ABSA tasks, both DO LLMs (e.g., GPT-3.5-turbo [14]) and ED LLMs (e.g., Flan-T5-XXL [15]) tend to underperform in automatic and human evaluations compared to ED backbone models (e.g., T5-Large [16]) trained with domain-specific data. Additionally, the performance can vary significantly with the prompt design. These observations indicate that when LLMs are directly deployed for ISA without training, their reasoning capacity may not be fully harnessed, resulting in suboptimal results. THOR [6] applied CoT fine-tuning to Flan-T5 for ISA and achieved improved performance. However, the intermediate steps generated by Flan-T5 often lack trustworthiness, with insufficient or duplicated content constrained by weak generation capabilities. As illustrated in Fig. 1, different LLMs exhibit varying performance in analyzing implicit sentiment toward the aspect term “price,” given the text “a cheaper price should not equal a ‘cheap’ product.” Inferior LLMs, such as Flan-T5 in the group of encoder-decoder (ED) LLM, display excellent comprehension and reasoning in tasks with diverse input information but limited generation and prompt-based inference capabilities in open-text scenarios [17]. They were predisposed to inaccurately predict implicit sentiment in the absence of explicit cues. Conversely, decoder-only (DO) LLMs, such as Vicuna-13B [18] and GPT-3.5-turbo, exhibit more advanced generation abilities and can effectively deduce sentiment elements pertinent to the context under reasoning prompts. However, their reliability in consistently producing accurate or correct responses remains a concern. Furthermore, LLMs often showcase superior performance with emergent abilities when scaling up at a certain level [19], the direct deployment or fine-tuning of large-scale models may be hindered by prohibitive computational costs. To effectively discern implicit sentiment polarities toward a specific aspect, it is essential to exploit reliable reasoning methods for applicable backbone models.

With this motivation, this study is aimed at enhancing the reasoning ability of ED backbone models by explicitly learning from convincing rationales provided by DO LLMs through synchronous verification. Specifically, we construct sentiment elements and develop a three-hop reasoning (TH-RE) prompting approach to guide DO LLMs in explicitly inferring sentiment elements before determining the final sentiment. The ED model, as the backbone, is fine-tuned based on the generated rationales and gold labels in datasets. To ensure the quality of reasoning learning, we introduce an answer-based verification mechanism, which serves as an additional layer for assessing the reliability of the rationale. This framework promotes dialectical learning, facilitating the identification and rectification of potential inaccuracies.

The contributions of this work can be summarized as follows:

- We propose a novel two-stage learning framework, Reasoning and Verification for Implicit Sentiment Analysis (RVISA), to improve the proficiency of ED backbone models as adept reasoners in ISA, leveraging the generative strengths of DO LLMs.
- We introduce a straightforward yet efficacious answer-based verification mechanism that can reliably

supervise reasoning learning, thereby improving overall performance.

- Evaluation on two benchmark datasets demonstrates the efficacy of the proposed method in achieving state-of-the-art performance in ISA tasks.

## II. RELATED WORK

In this work, we train a skilled reasoner using LLMs to conduct ISA and learn fruitful information from rationales generated by reasoning prompts. This section outlines the existing research on ISA and methods aimed at learning from reasoning prompts, leveraging the emergent abilities of LLMs.

### A. Implicit Sentiment Analysis

ISA has gained considerable attention in the field of SA [4], [20]. Initial efforts were dedicated to exploring the implicit sentiment detected at the sentence level [21], [22]. However, with increasing social demand, scholars have recently attempted to develop effective paradigms addressing the unique characteristics of ISA in a more fine-grained manner at the aspect level [5], [6], [23]. To capture implicit sentiment expressions, various methods have been developed, which leverage additional knowledge to improve the learning performance. For example, Li et al. [5] pre-trained a model on large-scale sentiment-annotated corpora with supervised contrastive learning objectives to align the representation of explicit and implicit sentiment expressions. Instead of leveraging external knowledge, Ouyang et al. [24] realized explicit sentiment augmentation using the backbone language model to enhance implicit classification abilities. Considering the difficulties of obtaining comprehensive knowledge through additional means, Wang et al. [23] applied reasoning learning under causal intervention to capture correlations within expressions. In fine-grained SA, relationships can be defined by four key sentiment elements, i.e., *target*, *aspect*, *opinion*, and *sentiment polarity* [25]. These sentiments are crucial for understanding the underlying sentiment. Given the impressive CoT performance and in-context learning abilities of LLMs, Fei et al. [6] introduced the THOR method with CoT fine-tuning to guide the ED backbone model in inferring sentiment elements step-by-step during training. In contrast, our approach aims to address the limited generation capabilities of ED LLMs. Instead of directly inferring the sentiment elements from backbone models, we train ED backbone models to become proficient reasoners by leveraging the informative rationale generated by DO LLMs.

### B. Reasoning Prompting

LLMs have demonstrated impressive complex reasoning abilities with CoT prompting [7], [8]. Reasoning prompting based on the CoT concept has been shown to enhance the zero-shot and few-shot performance of LLMs [26], [27], [28], [29]. CoT decomposes complex problems into smaller, manageable steps. This principle aligns with cognitive load theory [30], which posits that problem decomposition can reduce cognitive overload, allowing models to process information more effectively

**Reasoning (RE):**

Given the sentence [input\_text], **what is the sentiment polarity towards [aspect\_term], why?**

**Rationalization (RA):**

Given the sentence [input\_text], **the sentiment polarity towards [aspect\_term] is [label\_polarity], why?**

**Zero-CoT:**

Given the sentence [input\_text], **what is the sentiment polarity towards [aspect\_term], why? Let's think step by step.**

**Three-hop Reasoning (TH-RE):**

Given the sentence [input\_text], **what is the sentiment polarity towards [aspect\_term], why?**

**Let's think step by step. The mentioned aspect** towards [aspect\_term] is about ... **The underlying opinion** towards [aspect\_term] is about ... **Therefore, the sentiment polarity** towards [aspect\_term] is ...

**Three-hop Rationalization (TH-RA):**

Given the sentence [input\_text], **the sentiment polarity towards [aspect\_term] is [label\_polarity], why?**

**Let's think step by step. The mentioned aspect** towards [aspect\_term] is about ... **The underlying opinion** towards [aspect\_term] is about ... **Therefore, the sentiment polarity** towards [aspect\_term] is ...

Fig. 2. Reasoning prompts applied to sentiment analysis. Left: commonly used prompting modes. Right: three-hop prompting for ISA.

and arrive at more accurate conclusions. Fig. 2 illustrates various reasoning prompts applied to SA, including reasoning and rationalization techniques:

*a) Reasoning (RE):* Hase et al. [31] introduced multi-task learning with reasoning prompting, in which the question-answer pairs and question-explanation pairs were simultaneously learned. This method encourages the model to infer answers through step-by-step reasoning, allowing the explanation to differ from the gold label.

*b) Rationalization (RA):* Camburu et al. [32] proposed the concept of rationalization, which attempts to retrieve the explanation for a question by explicitly giving the correct answer. This method aims to connect the question with the gold label and elucidate the reasoning behind the question-answer relationship.

Kojima et al. [8] revealed that LLMs are capable of incremental reasoning without exemplars. By simply incorporating the prompt “let’s think step by step” (i.e., Zero-CoT in Fig. 2), this approach can be applied universally across various tasks. However, the granularity of the reasoning steps generated by Zero-CoT remains unpredictable, depending on the inherent knowledge of the LLM and varying between models. Expanding on this, Jin et al. [33] examined the influence of reasoning step length in prompts and reported that maintaining a certain step size according to the task complexity is crucial for generating a correct response. Our method builds on these insights by incorporating a TH-RE strategy for constructing sentiment elements, ensuring that critical reasoning steps are maintained.

### C. Learning From Rationale

The concept of learning from explanations to enhance the reasoning abilities of the training model has been explored in various fields [10], [11], [12], [34]. LLMs are capable of validating their responses with reasonable intermediate steps [7], [8]. Moreover, rationales can be used as demonstrations [26] or additional fine-tuning data [10], [35], [36] to improve the learning performance. Considering the high training cost associated with LLMs, rationales can also serve as valuable supervised signals for smaller, task-specific models, which are easier to deploy [11], [12], [34], [37]. However, some approaches [12] directly kept the

answer generated by LLMs as supervision signals, overlooking potential errors. Other methods [34] reorganized rationale sets based on answer-based filtering to reduce error learning, yet misleading information may persist. In contrast, our approach retains informative rationales throughout the learning process. To alleviate the negative effect of inaccurate rationales, we incorporate an answer-based verification mechanism as an additional supervision signal for multi-task learning. This innovative strategy not only preserves the valuable information within the rationales but also leverages them to refine the learning process with both positive and negative signals.

In summary, we introduce a novel two-stage framework, RVISA, that effectively combines the strengths of ED and DO LLMs to enhance reasoning performance. RVISA uniquely integrates TH-RE for rationale generation with an answer-based verification mechanism, resulting in more reliable reasoning learning for ISA.

## III. TWO-STAGE REASONING FRAMEWORK

Fig. 3 illustrates the proposed RVISA framework, which aims to enhance the reasoning ability of ED models while incorporating an answer-based verification mechanism for reasoning refinement during model learning. In the initial stage, DO LLMs are used to generate insightful rationales and predict labels through the TH-RE prompting approach. Subsequently, verification signals are derived based on the correctness of LLM prediction labels. In the second stage, the generated rationales are used for multi-task fine-tuning on an ED backbone model. To ensure the reliability of the generated rationales, we implement a straightforward yet effective answer-based verification mechanism with an additional task supervised by the verification signals to facilitate self-revision during reasoning learning. Different tasks are distinguished by task-specific prefixes. By simultaneously training on reasoning rationales, self-verification signals, and annotated gold labels, the model can learn to understand the underlying logic and relationships among sentiment elements that govern implicit sentiment prediction.



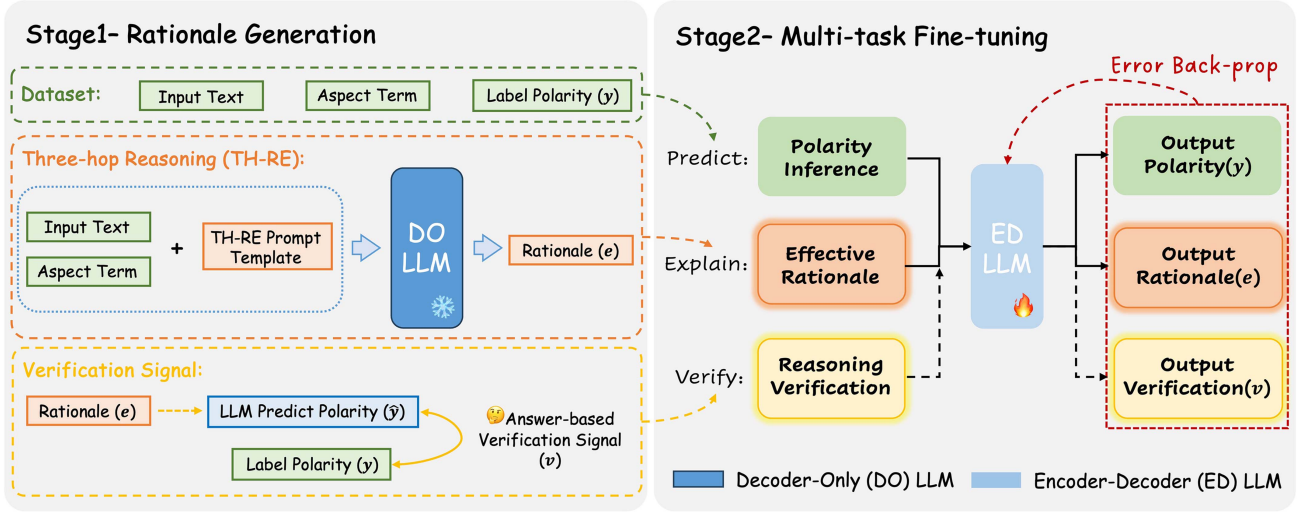


Fig. 3. Overview of proposed two-stage reasoning framework RVISA. Left: rationale generation stage leveraging the DO LLM to generate effective rationales and corresponding answer-based verification signals. Right: multi-task fine-tuning stage for training an ED backbone model as an enhanced reasoner with additional explanation tasks along with verification supervision.

#### A. Problem Definition

Consider an SA task including ESA and ISA, with a dataset  $D = (x_i, y_i)^N$ , where  $1 \leq i \leq N$ ,  $x_i$  represents an input sentence serving as a data example. Within each sentence  $x_i$ , an aspect term  $t_i$  is identified, denoted as  $t_i \subset x_i$ . The relevant sentiment elements consist of aspect  $a_i$ , opinion  $o_i$ , and sentiment polarity  $y_i$ . The objective of the task is to infer the sentiment polarity  $y_i$  toward the aspect term  $t_i$ , given the input sentence  $x_i$  and the specified aspect term  $t_i$ . In the conventional prompting approach for direct fine-tuning, the LLM predicts the sentiment polarity  $\hat{y}_i$  solely through  $\hat{y}_i = \text{argmax}(y_i | x_i, t_i)$ , without incorporating the intermediate sentiment elements. This approach potentially limits the ability of models to capture the sentiment nuances within the text.

#### B. Rationale Generation

To improve the generation of informative rationales, we prompt DO LLMs to generate intermediate steps during the inference of implicit sentiments. Sentiment elements are essential in directing the inference process for understanding how the sentiment is evoked as they contribute to a complete SA picture. Therefore, we design a three-hop (TH) prompting approach, as illustrated in Fig. 2, which deviates from conventional prompting modes. The objective of this design is to emphasize the extraction of strongly correlated sentiment elements in the reasoning process. Moreover, this design helps standardize the generative structure, facilitating improved learning of patterns and interconnections among rationales. The details of the TH prompting are explained as follows.

a) *Three-hop Reasoning (TH-RE)*: Fine-grained SA involves analyzing key sentiment elements, i.e., the target, aspect, opinion, and sentiment polarity [25]. Various approaches exist for solving these individual subtasks or their combinations, collectively contributing to a comprehensive SA framework. To address the complexity of the task holistically, it is essential to consider the components systematically and tackle them

incrementally. THOR [6] designed a prompting method based on CoT to explicitly infer sentiment elements by three-step generation during model fine-tuning. However, the prompting for each step was inferred separately for each sentiment element at a time, with the results concatenated as context information for the subsequent step without further validation.

In the proposed approach, we leverage a structured design that explicitly presents sentiment elements in a natural language sequence to construct a TH-RE prompt. This approach highlights the causal relationships among sentiment elements and the final sentiment polarity prediction in a single iteration.

As shown in the template below, we incorporate sentiment elements as cues at the end of “let’s think step-by-step,” guiding the language model to generate reasoning steps aligned with the understanding of the sentiment elements and then infer the sentiment polarity. DO LLMs are expected to predict the explanation as  $\hat{e}_i = \text{argmax}(e_i | x_i, t_i)$ , where  $\hat{a}_i, \hat{o}_i, \hat{y}_i \subset \hat{e}_i$ .

Given the sentence  $x_i$ , what is the sentiment polarity towards  $t_i$ , why? Let’s think step by step. The *mentioned aspect* towards  $t_i$  is about... The *underlying opinion* towards  $t_i$  is about... Therefore, the *sentiment polarity* towards  $t_i$  is...

b) *Three-hop Rationalization (TH-RA)*: We adapt the TH-RE framework with rationalization to establish three-hop rationalization (TH-RA) prompting. Specifically, the gold label is provided as the reference, which prompts the LLMs to elucidate the annotated sentiment label through a systematic and step-by-step inference process guided by sentiment elements. DO LLMs are expected to predict the explanation as  $\hat{e}_i = \text{argmax}(e_i | y_i, x_i, t_i)$ , where  $\hat{a}_i, \hat{o}_i, \hat{y}_i \subset \hat{e}_i$ .

Given the sentence  $x_i$ , the sentiment polarity towards  $t_i$  is  $y_i$ , why? Let’s think step by step. The *mentioned aspect*

towards  $t_i$  is about... The *underlying opinion* towards  $t_i$  is about... Therefore, the *sentiment polarity* towards  $t_i$  is...

### C. Multi-Task Fine-Tuning

We apply multi-task fine-tuning to simultaneously learn the rationales generated by the LLM and the annotated labels. Given the dataset  $D = \{(x_i, y_i)\}^N$ , where  $1 \leq i \leq N$ , we generate an explanation  $e_i$  to serve as a rationale for each input  $x_i$ , as outlined in Section III-B. Each explanation  $e_i$  encompasses a generated label  $\hat{y}_i$  from the LLM, denoted as  $\hat{y}_i \subset e_i$ . Subsequently, we construct a new dataset  $D_{exp} = \{(x_i, e_i)\}^N$ , where  $1 \leq i \leq N$ . The objective during the training phase is to effectively leverage the generated content and learn from two distinct tasks: an explanation task that uses data from  $D_{exp}$  and a prediction task harnessing data from the original dataset  $D_{pre} = D = \{(x_i, y_i)\}^N$ , where  $1 \leq i \leq N$ . To enhance the reasoning performance, we incorporate an answer-based verification mechanism within the existing multi-task learning framework. This mechanism enhances the learning process by providing verification signals for additional-task learning. This mechanism is explained in the subsequent Section III-C2.

1) *Learning With Rationale*: To train proficient reasoners, we use multi-task learning and divide the learning task into explanation and prediction subtasks. The explanation task tends to furnish the rationales based on the input sample and task objective, while the prediction task focuses solely on inferring sentiment polarity. Through the implementation of multi-task learning, the training phase incorporates the losses associated with both explanation and prediction tasks. The loss function is delineated as follows, where  $\mathcal{L}_{exp}$  and  $\mathcal{L}_{pre}$  denote the losses for the explanation and prediction tasks, respectively:

$$\mathcal{L}_{loss} = \alpha \mathcal{L}_{exp} + (1 - \alpha) \mathcal{L}_{pre} \quad (1)$$

where the prediction  $\mathcal{L}_{pre}$  aims to minimize the cross-entropy loss for label prediction:

$$\mathcal{L}_{pre} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(\hat{y}_i, y_i) \quad (2)$$

while the explanation loss  $\mathcal{L}_{exp}$  minimizes the generation loss for the rationale, with subtle distinctions between reasoning (RE) and rationalization (RA) scenarios.

$$RE : \mathcal{L}_{exp} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(f(x_i, t_i), \hat{e}_i) \quad (3)$$

$$RA : \mathcal{L}_{exp} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(f(x_i, t_i, y_i), \hat{e}_i) \quad (4)$$

The objective is to equip the model with strong explanation and prediction capabilities, thereby enhancing its reasoning ability. However, during the inference phase, only the prediction task is executed for evaluation to optimize the inference efficiency and mitigate computational costs.

2) *Reasoning With Verification*: If the rationale generated by an LLM is directly used without any post-filtering processes, error patterns may be introduced, which can degrade the performance of multi-task fine-tuning. In prior studies, answer-based filtering has been used to improve the rationale quality. For example, [34] removed the incorrect rationales obtained through RE prompting based on the final prediction and replaced them with the rationales generated through RA prompting to complete the final rationale set for training. [10] demonstrated that answer-based filtering can be compensated by using diverse reasoning paths and retaining only the rationales leading to the correct answer. In the proposed approach, we preserve the sets of rationales generated by the LLM by introducing a verification signal to facilitate further analysis of the rationale quality within the multi-task learning framework. This is achieved by incorporating an additional task for verification.

Specifically, we leverage the rationale set generated by our TH-RE prompting and introduce an answer-based verification mechanism guided by the prediction label  $\hat{y}_i$  from the LLM and the ground truth annotation  $y_i$ . Rationales leading to the correct answer label are deemed to be superior and more useful compared with those resulting in incorrect answer labels. Based on this premise, we execute the prompting using the following template:

Given the rationale  $e_i$ , please verify whether the above-given rationale is reasonable. Return True or False.

To generate the verification signal  $v_i$ , we validate the reasoning rationales that successfully yield the correct answer, while incorrect rationales are labeled as *False*, indicating a need for further refinement. However, LLMs with larger parameter scales, such as GPT-3.5-Turbo, tend to produce ambiguous answers containing dual polarities when faced with uncertainty in making a final judgment. Therefore, we establish the verification signal  $v_i$  based on the following criteria:

$$\begin{cases} \text{if } \hat{y}_i \subset \{\hat{y}_{i(t_1)}, \hat{y}_{i(t_2)}\}: & \hat{y}_i = \hat{y}_{i(t_1)} \\ \text{if } \hat{y}_i = y_i: & v_i \text{ is True} \end{cases} \quad (5)$$

For cases where the rationale presents two polarities,  $\hat{y}_{i(t_1)}$  and  $\hat{y}_{i(t_2)}$ , where  $t_2 > t_1$ , answer-based verification is performed on  $\hat{y}_{i(t_1)}$  based on the first-come-first-served (FCFS) rule. This is because the label generated earlier is considered to have a higher likelihood in the context of the next token generation. Then, the revised loss function incorporating verification signals is formulated as follows:

$$\mathcal{L}_{loss} = \alpha \mathcal{L}_{exp} + \gamma \mathcal{L}_{ver} + (1 - \alpha - \gamma) \mathcal{L}_{pre} \quad (6)$$

where  $\alpha$  and  $\gamma$  are weighting coefficients assigned to the explanation and verification tasks, respectively. These coefficients are determined through grid search on the validation set, as discussed in Section IV. The verification loss is defined considering the answer-based validation outcome under the supervision of

TABLE I  
COMPARISON OF THE PROPOSED APPROACH WITH BASELINE METHODS ON RESTAURANT AND LAPTOP DATASETS

	Restaurant			Laptop		
	All <sub>A</sub>	All <sub>F</sub>	ISA <sub>A</sub>	All <sub>A</sub>	All <sub>F</sub>	ISA <sub>A</sub>
<b>- State-of-the-art baselines</b>						
BERT + SPC <sup>†</sup> (110M) [40]	83.57	77.16	65.54	78.22	73.45	69.54
BERT + ADA <sup>†</sup> (110M) [41]	87.14	80.05	65.92	78.96	74.18	70.11
BERT + RGAT <sup>†</sup> (110M) [42]	86.60	81.35	67.79	78.21	74.07	72.99
BERT <sub>Asp</sub> + CEPT <sup>†</sup> (110M) [5]	87.50	82.07	67.79	81.66	78.38	75.86
BERT <sub>Asp</sub> + SCAPT <sup>†</sup> (110M) [5]	89.11	83.79	72.28	82.76	79.15	77.59
T5 <sub>Base</sub> + ABSA-ESA* (220M) [24]	88.29	81.74	70.78	82.44	79.34	80.00
<b>- Prompt-based methods</b>						
Flan-T5 + prompt (250M)	86.88	79.78	65.17	81.98	77.93	73.71
Flan-T5 + prompt (11B)	89.29	83.68	75.28	81.82	77.69	75.43
Flan-T5 + THOR (250M) [6]	87.68	81.10	68.54	81.66	77.51	74.29
Flan-T5 + THOR (11B) [6]	88.57	82.93	73.03	82.29	78.78	76.57
<b>- Our methods</b>						
Flan-T5 + RVISA <sub>v</sub> (250M)	86.43	78.49	65.92	80.72	76.49	73.71
Flan-T5 + RVISA <sub>g</sub> (250M)	86.61	78.92	66.67	81.19	77.13	75.43
Flan-T5 + RVISA <sub>v</sub> (11B)	91.25	86.57	81.65	86.52	83.28	87.43
Flan-T5 + RVISA <sub>g</sub> (11B)	<b>91.52</b>	<b>86.85</b>	<b>82.02</b>	<b>86.68</b>	<b>84.05</b>	<b>88.00</b>

The results with <sup>†</sup> and \* are obtained from [5] and [24], while the other results are derived from self-rerun or self-implementation. For our methods, the subscripts indicate learning from rationales generated by different models, i.e., Vicuna-13B(v) and GPT-3.5-Turbo(g). Subscripts *A* and *F* denote the accuracy and macro-F1 score, respectively.

the verification signal:

$$\mathcal{L}_{ver} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(f(e_i), \hat{v}_i) \quad (7)$$

#### IV. EXPERIMENTS

##### A. Setup

We evaluate the performance of the proposed framework using the Restaurant and Laptop datasets from SemEval-2014 [38]. To test the ISA performance, we follow the prior works utilizing datasets that are labeled with explicit and implicit tags [5]. To generate effective rationales conducive to reasoning learning, we use DO LLMs, Vicuna-13B [18] and GPT-3.5-turbo [14] in the first stage for rationale preparation. Considering the impressive performance of ED models in understanding input information and comprehension among different tasks, Flan-T5 [15] is used as the backbone LLM in the multi-task fine-tuning stage. Different sizes of Flan-T5 are considered, ranging from the base model (250 M) to the XXL model (13B). We compare the proposed approach against seven state-of-the-art baseline methods: BERT+SPC [39], BERT+ADA [40], BERT+RGAT [41], BERT<sub>Asp</sub>+CEPT [5], BERT<sub>Asp</sub>+SCAPT [5], THOR [6], and ABSA-ESA [24]. Among these, THOR [6] used CoT prompting with three-step generation. Unlike this method, we use a multi-task learning framework during training while directly inferring the final prediction during inference. To identify the optimal hyperparameters in the training loss, a grid search is performed using the validation set to determine the final values of  $\alpha$  and  $\gamma$ . Without verification supervision, the best results are obtained with  $\alpha = 0.5$ , using explanation and prediction tasks alone. With verification supervision, we achieve the highest performance when  $\alpha = \gamma = 0.3$ . Thus, all subsequent experiments adopt this hyperparameter setting.

##### B. Main Results

*a) RVISA outperforms baselines:* The main results comparing RVISA and baseline methods are demonstrated in Table I, using accuracy and macro-F1 score as the evaluation metrics. As THOR [6] did not report the accuracy outcomes for implicit sentiment, we rerun the results based on the provided source code. RVISA significantly outperforms the baseline methods, regardless of whether learning is accomplished using Vicuna-13B or GPT-3.5-turbo. This result highlights the efficacy of learning within the proposed two-stage reasoning framework.

*b) Strong teachers enhance learning quality:* Training RVISA<sub>g</sub> with GPT-3.5-turbo leads to enhanced reasoning capabilities in implicit sentiment inference compared with training RVISA<sub>v</sub> using the rationales generated by Vicuna-13B. This disparity is attributable to the superior common sense knowledge and reasoning prowess exhibited by GPT-3.5-turbo, which yields higher quality rationales that help enhance the reasoning abilities of the Flan-T5 backbone model. However, the base-size backbone model (250M) is inferior to some baseline methods due to its limited generation capacity to leverage the rich knowledge derived through in-context learning.

*c) Explicit rationales facilitate implicit reasoning:* We compare our method with THOR, which was built upon a chain-of-thought strategy. Instead of eliciting the reasoning ability of the language model through sequential three-step prompting during training, the proposed method explicitly yields the rationales that are then used as informative resources, along with an answer-based verification mechanism to ensure the learning quality. The comparative results are depicted in Table II. RVISA demonstrates superior performance over THOR in terms of the F1 score for ISA, while maintaining competitive results overall. This observation underscores the effectiveness of RVISA in learning implicit sentiment through reasoning tasks and adeptly capturing implicit relationships among instances.



TABLE II  
COMPARISON OF THE PROPOSED APPROACH WITH THOR [6]

	Restaurant		Laptop	
	All	ISA	All	ISA
Prompt <sup>†</sup> (11B)	83.68	74.48	77.69	72.44
THOR <sup>†</sup> (11B)	82.93	73.08	78.78	72.82
RVISA <sub>v</sub> (11B)	86.57	81.73	83.26	85.36
RVISA <sub>g</sub> (11B)	<b>86.85</b>	<b>82.61</b>	<b>84.05</b>	<b>86.20</b>

The evaluation metric is the F1 score training with Flan-T5. The results with <sup>†</sup> are obtained through self-rerun using the source code from [6].

TABLE III  
ABLATION STUDY OF THREE-HOP PROMPTING (TH) AND VERIFICATION (VE)  
BASED ON F1 SCORE METRIC

	Restaurant		Laptop	
	All	ISA	All	ISA
RVISA <sub>v</sub>	86.57	81.73	83.26	85.36
- w/o VE	85.91	80.40	82.63	83.39
- w/o VE and TH	85.79	79.10	82.57	82.91
RVISA <sub>g</sub>	<b>86.85</b>	<b>82.61</b>	<b>84.05</b>	<b>86.20</b>
- w/o VE	86.16	80.32	82.51	83.83
- w/o VE and TH	85.60	79.68	82.05	83.14

Although THOR claimed that three-step generation during fine-tuning can enhance the reasoning power of the backbone model, the re-run result demonstrates limited improvement in F1 scores. This suggests that the vulnerability of THOR in performing prompt-based inference depends on the capabilities of backbone models (i.e., Flan-T5 [15]) only. In contrast, our method prioritizes effective learning from high-quality sentiment information and closely related tasks, offering a more coherent and justifiable approach to achieving high performance in ISA.

### C. Ablation Study

We conduct an ablation study on the three-hop (TH) prompting and verification mechanism (VE) components, the results of which are summarized in Table III. The results include the F1 scores in both overall and implicit sentiment scenarios. The findings indicate that removing the verification mechanism leads to performance degradation in both cases, with a more significant decline of over one point observed in the implicit sentiment results. This highlights the critical role of verification signals in reasoning learning from LLMs, as the answer-based verification mechanism aids the backbone model in identifying potential errors or unreasonable attributes during multi-task learning.

The performance further deteriorates when CoT prompting is degraded from TH-RE to RE prompting alone, without using the sentiment elements to guide rationale generation. This performance deterioration is observed in both implicit and general scenarios. Moreover, this observation remains consistent regardless of whether the rationales are generated by Vicuna-13B or GPT-3.5-turbo, indicating that, irrespective of generation quality, the three-hop prompting plays a pivotal role in steering the correct

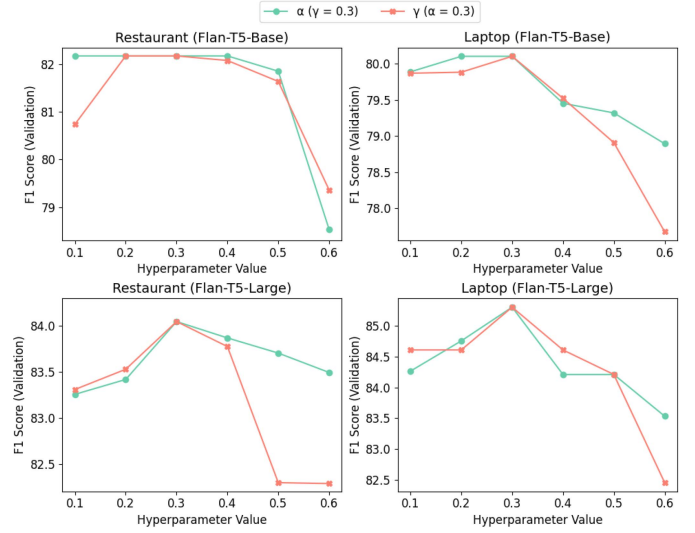


Fig. 4. Overall F1 score for different hyperparameter settings on the two datasets, using rationales generated by GPT-3.5-turbo. The experiments are conducted on the validation set.

direction of reasoning for ISA. Although the performance degradation without TH-RE prompting is less pronounced compared with that in the absence of the verification mechanism, their contributions are mutually reinforcing and non-negligible. This combined effect reflects the essence of multi-task learning, where tasks are strongly related and complement each other.

### D. Influence of Hyperparameters

In our experiments,  $\alpha$  and  $\gamma$  are set as 0.3 based on a grid search conducted on the validation set. Fig. 4 illustrates the influence of various hyperparameter configurations on the two datasets. We set the granularity of weights as 0.1 and examine the changes in the overall F1 score on the validation set. Both the base-size and large-size models achieve optimal performance on the two datasets when  $\alpha = \gamma = 0.3$ . This result underscores the critical interplay between rationale generation and verification, suggesting that a balance between the two aspects enhances the reliability of reasoning learning. When either  $\alpha$  or  $\gamma$  is set to a higher value, the prediction task is assigned a smaller weight and may lose its dominant position, resulting in a more rapid decline in prediction performance. When either  $\alpha$  or  $\gamma$  is set to a lower value, one of the explanation and verification tasks exerts less influence, weakening their joint auxiliary effect on the prediction task. These effects are particularly pronounced in larger models, which exhibit a greater capacity for reasoning and understanding, rendering them more sensitive to variations in hyperparameter settings.

### E. Further Analysis

a) *Influence of prompting method:* To investigate the influence of diverse prompts, we conduct training experiments using rationales generated by Vicuna-13B with various prompting methods, including RE, Zero-CoT, TH-RE, and TH-RA, as

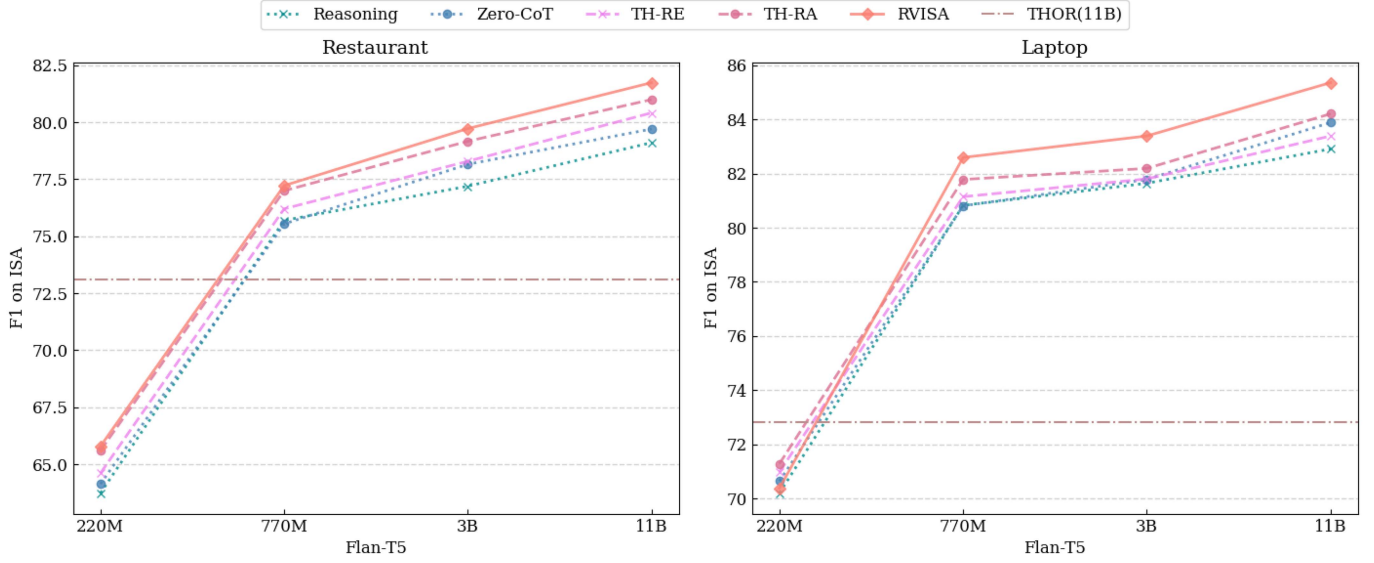


Fig. 5. Influence of rationale and model size on implicit F1 score. The dashed horizontal line represents the best result of THOR re-run with the Flan-T5-XXL (11B) model on the implicit dataset.

depicted in Fig. 2. We compare the results obtained using RVISA and THOR, as shown in Fig. 5, where RVISA is enhanced by TH-RE prompting with the verification mechanism. The model trained with TH-RA demonstrates the second-best performance as rationalization helps leverage the gold labels as context information to elucidate the underlying logic. This approach facilitates the generation of more reasonable rationales that lead to correct answers. Consequently, TH-RA generally outperforms TH-RE without verification, as TH-RE may produce more unreasonable responses, resulting in incorrect answers. However, RVISA consistently outperforms both TH-RA and TH-RE, suggesting that the language model, when trained under verification signals, can leverage erroneous or irrational attributes present in TH-RE-generated rationales. This approach can effectively use LLM-generated labels as an additional verification factor. Furthermore, rationales generated using RE and Zero-CoT methods lag behind three-hop prompting in most scenarios, underscoring the effectiveness of the proposed prompting approach in structuring coherent rationales and extracting highly relevant sentiment elements within the three-hop prompting framework.

*b) Influence of model size:* Fig. 5 illustrates the influence of backbone model size on reasoning learning. For the Restaurant dataset, smaller models (i.e., base and large) exhibit marginal performance improvements under the verification mechanism, indicating limited capacity to benefit from prompt-based inference within the multi-task framework. As model size increases, the combined benefits of the verification mechanism and TH-RE prompting become evident, widening the performance gap compared with the second-best TH-RA method. Notably, with large-size (770 M) models, RVISA achieves superior performance over the best result of THOR trained with the Flan-T5-XXL (11B) model on both Restaurant and Laptop datasets, showcasing the efficacy of our method in enhancing reasoning abilities for pre-trained models. For the XXL (11B) size, TH-RE, TH-RA, and RVISA outperform THOR in implicit

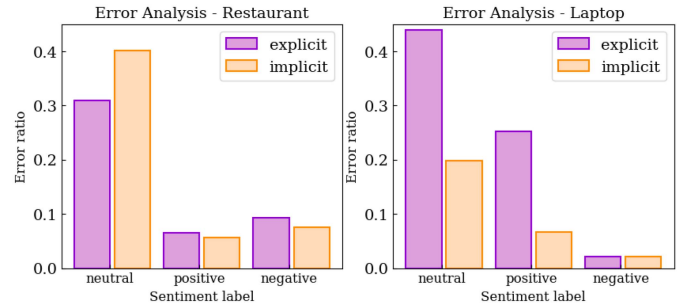


Fig. 6. Error analysis for two datasets with rationales generated by GPT-3.5-turbo. The error ratio refers to the proportion of the number of error types to the total number of error instances.

sentiment prediction, with similar trends observed for the Laptop dataset. All prompting methods with XXL-size models under the multi-task learning framework surpass the best result of THOR, emphasizing the effectiveness of our proposed framework and the scaling effect on the learning capabilities of the trained model.

*c) Error Analysis:* The proposed method demonstrates superior ISA performance compared with the baselines. To further explore the errors, we calculate the error ratio considering sentiment types, including explicit and implicit, and their relationships with the corresponding sentiment labels. The results are shown in Fig. 6, with the rationales generated from GPT-3.5-turbo. For the Laptop dataset, errors in neutral predictions within the explicit dataset exceed those in the implicit dataset, leading to higher F1 scores for the implicit dataset compared with the overall dataset. Conversely, in the Restaurant dataset, the error ratio associated with neutral predictions in the implicit dataset is higher than that for the original dataset. This observation underscores the significant influence of neutral sentiment distribution on error distribution patterns. Moreover, the ratio of



TABLE IV  
REASONING (RE) PROMPTING COMPARED WITH THREE-HOP REASONING (TH-RE) PROMPTING FOR GPT-3.5-TURBO MODEL

<b>Input Text</b>	a cheaper price should not equal a “cheap” product.
<b>Aspect Term</b>	price
<b>Sentiment Polarity</b>	positive
<b>Re Prompt</b>	Given the sentence {input text}, what’s the sentiment polarity towards price, why?
<b>GPT-3.5-turbo</b>	The sentiment polarity towards price in this sentence is <b>negative</b> . The sentence suggests that a lower price does not necessarily mean that the product is of low quality or “cheap.” In other words, the sentence implies that there is a common misconception that cheaper products are of inferior quality, which is not necessarily true. To break it down step by step: 1. The sentence contains the word “cheaper,” which is a synonym for “lower-priced” or “more affordable.” This suggests that the speaker is concerned with the price of the product. 2. However, the sentence also contains the word “cheap,” which is often used to describe products of low quality or value. This suggests that the speaker is concerned with the perception of the product’s quality. 3. The sentence negates the idea that a cheaper price automatically equals a “cheap” product. This suggests that the speaker believes that there is a common misconception that cheaper products are of inferior quality, which is not necessarily true. 4. Therefore, the sentiment towards price is negative because the sentence implies that people often assume that cheaper products are of low quality, which is not always the case.
<b>TH-RE Prompt</b>	Given the sentence {input text}, what’s the sentiment polarity towards price, why? Let’s think step by step. <b>The mentioned aspect</b> towards price is about ... <b>The underlying opinion</b> towards price is about ... <b>Therefore, the sentiment polarity</b> towards price is ...
<b>GPT-3.5-turbo</b>	<b>The mentioned aspect</b> towards price is about the relationship between price and the quality of a product. <b>The underlying opinion</b> towards price is that just because something is offered at a lower price, it should not automatically be assumed to be of low quality or “cheap.” <b>Therefore, the sentiment polarity</b> towards price in this sentence is <b>positive</b> . The sentiment is emphasizing that price should not be the sole determinant of product quality, advocating for a more nuanced understanding of the pricing and value relationship.

incorrect predictions pertaining to *neutral* polarity exceeds 60%. This suggests the nuanced challenges associated with accurately discerning neutral sentiments in SA tasks, highlighting the need for further refinement and optimization in model training and inference processes.

#### F. Case Study

Table IV presents the results of a case study to demonstrate the effectiveness of TH-RE prompting, with the rationales generated using GPT-3.5-turbo with diverse prompting methods. Under the RE prompt, the LLM decomposes the problem into four fine-grained steps, leveraging its own pre-trained knowledge. However, despite additional inference steps, the prediction is the complete opposite of the gold label. In contrast, with the guidance of TH-RE prompting, the rationale follows the generation format specified in the prompting template, which is more concise and structured. This leads to more accurate inference of the sentiment elements, enhancing the final generation.

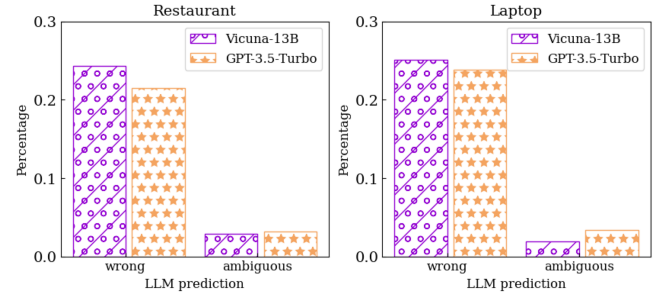


Fig. 7. Analysis of wrong and ambiguous predictions for Vicuna-13B and GPT-3.5-turbo, with ambiguous referring to multiple uncertain predictions in the rationale.

TABLE V  
EFFECT OF DIFFERENT LLMs ON LEARNING PERFORMANCE

	Restaurant		Laptop	
	All	ISA	All	ISA
RVISA <sub>v</sub> (11B)	86.57	81.73	83.26	85.36
RVISA <sub>g(3.5)</sub> (11B)	86.85	82.61	84.05	86.20
RVISA <sub>g(4o)</sub> (11B)	<b>87.55</b>	<b>82.99</b>	<b>84.48</b>	<b>87.33</b>

The evaluation metric is the F1 score under training with Flan-T5.

#### V. DISCUSSION

We propose a two-stage reasoning framework, RVISA, to learn effectively and reliably from the rationales generated by DO LLMs for ISA. Results of extensive experiments demonstrate that RVISA can enhance the reasoning and learning abilities of ED models under verification supervision. As outlined in this section, we investigate the effects of different LLMs on rationale generation and discuss alternative choices for multi-task learning. Moreover, we identify the limitations of the proposed approach and highlight potential extension directions of the RVISA framework for application in other tasks within this field.

##### A. Rationale Generation

a) *Vicuna-13B versus GPT-3.5-turbo*: To assess the quality of rationales generated from Vicuna-13B and GPT-3.5-turbo, wrong and ambiguous predictions are analyzed, as illustrated in Fig. 7. In both Restaurant and Laptop datasets, Vicuna-13B exhibits a slightly higher count of incorrect predictions compared with GPT-3.5-turbo. This suggests that stronger models such as GPT-3.5-turbo demonstrate a superior capability to generate higher-quality rationales, leading to more accurate final predictions. However, the percentage of ambiguous predictions originating from GPT-3.5-turbo exceeds that of Vicuna-13B, which indicates that the more powerful model exhibits a greater tendency to generate uncertain expressions rather than provide definitive judgments when deciphering the nuanced sentiment. This aspect also highlights the inherent challenge in capturing subtle nuances in sentiment within constrained contextual information.

b) *GPT-4o*: We investigate the potential benefit of using a more advanced LLM. Table V demonstrates the performance evaluation with GPT-4o compared with both Vicuna-13B and

GPT-3.5-turbo. GPT-4o not only surpasses Vicuna-13B but also consistently outperforms GPT-3.5-turbo across both datasets, which supports the observation that *strong teachers enhance the learning quality*. This result underscores the effectiveness and robustness of the RVISA framework in leveraging more advanced LLMs to enhance reasoning learning.

### B. Multi-Task Learning

Multi-task learning has gained significant traction in natural language processing tasks, including SA [42], [43], [44]. Recent multi-task learning methods can be categorized into *joint training* and *multi-step training* [45]. Joint training involves a single model with shared parameters and representations across multiple tasks, while multi-step training includes task-specific components, where the input for some tasks is contingent upon the outputs or hidden representations of preceding tasks. The choice of training methods is largely influenced by *task relatedness* [45]. RVISA utilizes the classical *joint training* technique to perform multi-task fine-tuning on ED LLMs, effectively leveraging both explanation and verification tasks. During multi-task learning optimization, a weighted sum of the losses from each task is typically minimized, which can increase the computational cost, as it requires updating parameters for all tasks, as is the case with RVISA. Given that the parameters in many multi-task learning models can be divided into distinct blocks, it is feasible to alternatively optimize each block while keeping the parameters constant in other blocks [46]. Future work could explore various training and optimization techniques within the multi-task learning framework to promote the application of RVISA.

### C. Limitations

In this study, we propose a straightforward yet effective answer-based verification mechanism to enhance the overall performance in SA. This mechanism plays a vital role in the RVISA framework, underscoring its significance in reasoning learning. While the current answer-based verification mechanism demonstrates effectiveness, there remains potential for further enhancement through the exploration of alternative verification modes or the incorporation of additional pertinent factors. This avenue for future research paves the way for more nuanced and reliable SA. Additionally, the three-hop prompting method proves instrumental in generating effective rationales by deducing sentiment elements. However, this approach is manually designed based on prior research, which may render further optimization challenging. Given the evolving landscape of advanced techniques focused on optimizing prompts for LLMs, it remains unclear whether the prompt can be generated automatically or optimized through the utilization of soft prompts. This presents a promising direction for further exploration.

### D. Extension

In the realm of SA, subjective text analysis (STA) extends beyond sentiment classification and ABSA to recognize and understand a broader range of human emotional states [13],

[47]. Specifically, STA encompasses a wide spectrum of tasks, such as suicide tendency detection, hate speech detection, and sarcasm detection [48], all of which aim to uncover underlying emotional states. However, these tasks address distinct nuances of sentiment expression compared with ISA, delving into the complexities of user intention and attitude conveyed through textual information. For instance, sarcasm detection often entails a deliberate inversion of meaning, where the literal interpretation contradicts the intended sentiment. Understanding this relationship paves the way for extending our approach to encompass tasks within STA. By adapting the TH-RE prompt to incorporate the unique characteristics of sarcastic expressions, we can extend the RVISA framework to effectively reason about sarcasm detection with minimal adjustments. Furthermore, exploring more fine-grained verification mechanisms could enhance the model capacity for complex reasoning. Future research may investigate the integration of STA tasks within our framework, which holds promise for broadening the applicability and effectiveness of RVISA in analyzing intricate sentiment dynamics.

## VI. CONCLUSION

In conclusion, this study sheds light on ISA in the era of LLMs and proposes a novel two-stage reasoning learning framework, RVISA, which incorporates reasoning and verification for ISA tasks. By leveraging the generative prowess of DO LLMs, the proposed framework enhances the reasoning capabilities of ED backbone models. The utilization of TH-RE prompting facilitates the explicit generation of cues guided by sentiment element construction, which is conducive to reasoning learning. Through a straightforward and effective answer-based verification mechanism, we ensure robust and reliable reasoning learning to further improve the proficiency of the ED backbone model in inferring implicit sentiment. Comprehensive experiments on two benchmark datasets demonstrate that the proposed framework achieves state-of-the-art performance in ISA tasks.

## REFERENCES

- [1] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012. [Online]. Available: <http://dx.doi.org/10.1007/978-3-031-02145-9>
- [2] R. Das and T. D. Singh, "Multimodal sentiment analysis: A survey of methods, trends, and challenges," *ACM Comput. Surv.*, vol. 55, 2023, Art. no. 270.
- [3] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, "Survey on sentiment analysis: Evolution of research methods and topics," *Artif. Intell. Rev.*, vol. 56, pp. 8469–8510, 2023.
- [4] I. Russo, T. Caselli, and C. Strapparava, "SemEval-2015 task 9: Cliveval implicit polarity of events," in *Proc. 9th Int. Workshop Semantic Eval.*, 2015, pp. 443–450.
- [5] Z. Li, Y. Zou, C. Zhang, Q. Zhang, and Z. Wei, "Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 246–256.
- [6] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T.-S. Chua, "Reasoning implicit sentiment with chain-of-thought prompting," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 1171–1182.
- [7] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 1800.
- [8] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 1613.

- [9] J. Wei et al., "Finetuned language models are zero-shot learners," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–46.
- [10] N. Ho, L. Schmid, and S. Yun, "Large language models are reasoning teachers," in *Proc. Conf. Assoc. Comput. Linguistics*, 2023, pp. 14852–14882.
- [11] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain yourself! Leveraging language models for commonsense reasoning," in *Proc. Conf. Assoc. Comput. Linguistics*, 2019, pp. 4932–4942.
- [12] C. Hsieh et al., "Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes," in *Proc. Int. Conf. Findings Assoc. Comput. Linguistics*, 2023, pp. 8003–8017.
- [13] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in *Proc. Int. Conf. Findings Assoc. Comput. Linguistics*, 2024, pp. 3881–3906.
- [14] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 2011.
- [15] H. W. Chung et al., "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.
- [16] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, 2019, Art. no. 140.
- [17] Y. Tay et al., "UL2: Unifying language learning paradigms," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–33.
- [18] L. Zheng et al., "Judging LLM-as-a-judge with MT-bench and chatbot arena," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, Art. no. 2020.
- [19] J. Wei et al., "Emergent abilities of large language models," *Trans. Mach. Learn. Res.*, 2022. [Online]. Available: <https://openreview.net/forum?id=yzkSU5zdWd>
- [20] C. Zong, R. Xia, and J. Zhang, *Sentiment Analysis and Opinion Mining*. Singapore: Springer, 2021.
- [21] M. Xu, D. Wang, S. Feng, Z. Yang, and Y. Zhang, "KC-ISA: An implicit sentiment analysis model combining knowledge enhancement and context features," in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 6906–6915.
- [22] D. Zhou, J. Wang, L. Zhang, and Y. He, "Implicit sentiment analysis with event-centered text representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6884–6893.
- [23] S. Wang et al., "Causal intervention improves implicit sentiment analysis," in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 6966–6977.
- [24] J. Ouyang, Z. Yang, S. Liang, B. Wang, Y. Wang, and X. Li, "Aspect-based sentiment analysis with explicit sentiment augmentations," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 18842–18850.
- [25] H. Peng, L. Xu, L. Bing, F. Huang, W. Lu, and L. Si, "Knowing what, how and why: A near complete solution for aspect-based sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8600–8607.
- [26] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 159.
- [27] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–24.
- [28] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, "Complexity-based prompting for multi-step reasoning," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–15.
- [29] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–32.
- [30] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, pp. 257–285, 1988.
- [31] P. Hase, S. Zhang, H. Xie, and M. Bansal, "Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language?," in *Proc. Int. Conf. Findings Assoc. Comput. Linguistics*, 2020, pp. 4351–4367.
- [32] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-SNLI: Natural language inference with natural language explanations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9560–9572.
- [33] M. Jin et al., "The impact of reasoning step length on large language models," in *Proc. Int. Conf. Findings Assoc. Comput. Linguistics*, 2024, pp. 1830–1842.
- [34] S. Li et al., "Explanations from large language models make small reasoners better," in *Proc. Workshop Sustain. AI Annu. AAAI Conf. Artif. Intell.*, 2024, pp. 1–14.
- [35] J. Huang et al., "Large language models can self-improve," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 1051–1068, pp. 1051–1068.
- [36] E. Zelikman, Y. Wu, J. Mu, and N. Goodman, "STaR: Bootstrapping reasoning with reasoning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 15476–15488.
- [37] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Teaching small language models to reason," in *Proc. Conf. Assoc. Comput. Linguistics*, 2023, pp. 1773–1781.
- [38] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androustopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Eval.*, 2014, pp. 27–35.
- [39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
- [40] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 4933–4941.
- [41] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 3229–3238.
- [42] T. Gao et al., "LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis," in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 7002–7012.
- [43] Z. Wang, R. Xia, and J. Yu, "Unified ABSA: A unified ABSA framework based on multi-task instruction tuning," 2022, *arXiv:2211.10986*.
- [44] Z. Gou, Q. Guo, and Y. Yang, "MVP: Multi-view prompting improves aspect sentiment tuple prediction," in *Proc. Conf. Assoc. Comput. Linguistics*, 2023, pp. 4380–4397.
- [45] Z. Zhang, W. Yu, M. Yu, Z. Guo, and M. Jiang, "A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2023, pp. 943–956.
- [46] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022.
- [47] Y. Zhang et al., "Affective computing in the era of large language models: A survey from the NLP perspective," 2024, *arXiv:2408.04638*.
- [48] R. Misra and P. Arora, "Sarcasm detection using news headlines dataset," *AI Open*, vol. 4, pp. 13–18, 2023.



affective computing and NLP for social good.

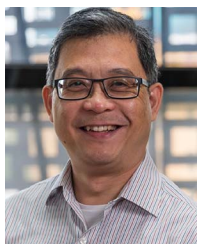


has published 427 research publications, including 251 journal articles. He is the editor-in-chief of the *Natural Language Processing Journal*, *Computers & Education: Artificial Intelligence*, and *Computers & Education: X Reality*. He has been selected as the World's Top 2% Scientists by Stanford University.

**Wenna Lai** (Student Member, IEEE) received the master's degree from the Department of Electrical and Computer Engineering, National University of Singapore. She is currently working toward the PhD degree with the Department of Computing, Hong Kong Polytechnic University, under the supervision of Prof. Qing Li. She has been working closely with Prof. Guandong Xu at the School of Computer Science, University of Technology Sydney, and Prof. Haoran Xie at the School of Data Science, Lingnan University, Hong Kong. Her research interests include

**Haoran Xie** (Senior Member, IEEE) received the PhD degree in computer science from the City University of Hong Kong, and the EdD degree in language learning from the University of Bristol. He is currently a professor and the person-in-charge with the Division of Artificial Intelligence, director with LEO Dr David P. Chan Institute of Data Science, and acting associate dean with the School of Data Science, Lingnan University, Hong Kong. His research interests include natural language processing, large language models, language learning, and AI in education. He





**Guandong Xu** (Member, IEEE) received the PhD degree in computer science from Victoria University, Melbourne, VIC, Australia, in 2009. He is currently a professor and a program leader with the School of Computer Science and Data Science Institute, University of Technology Sydney, Sydney, NSW, Australia. His research interests include data science, data analytics, recommender systems, web mining, user modeling, NLP, social network analysis, and social media mining.



**Qing Li** (Fellow, IEEE) received the BEng degree in computer science from Hunan University, Hunan, China, in 1982, and the MS and PhD degrees in computer science from the University of Southern California, LA, California, USA, in 1985 and 1988, respectively. He is a chair professor and head with the Department of Computing, Hong Kong Polytechnic University. His research focuses on data science, web mining, and artificial intelligence. He is a fellow of the IET, a member of ACM SIGMOD and IEEE Technical Committee on Data Engineering. He is the chairperson of the Hong Kong Web Society, and is a steering committee member of DASFAA, ICWL, and WISE Society.