Prompt Memoried LLMs with 'What, Why, and How' to Reason Implicit Emotions

Wei Cheng[1,†], Pengyu Li[1,†], Hejin Liu[1,†], Hailin Cheng[1], Zhenglin Cai[1] and Juan Wen[1,*]

[1]China Agricultural University, No. 17 Qinghua East Road, Haidian District, Beijing 100083, China

Abstract

Implicit Sentiment Analysis (ISA) aims to capture emotions that are subtly expressed, often obscured by ambiguity and literary devices. This requires a shift in thinking based on memorized prior knowledge (past cases) and multi-stage logical reasoning to uncover the hidden emotions of the speaker. Inspired by human memory and reasoning, we propose a *What, Why, and How* three-question (3Q) framework that incorporates a memory mechanism for past cases. We design a three-step prompt principle for it: based on domain priors, first identify *what* the most crucial entity is, then infer *why* the speaker mentioned it, and finally uncover *how* the hidden emotions are. During the reasoning process, historical queries and responses are stored in memory as past case pairs. These pairs can be used to retrieve and generate improved prompts for any new queries, thereby enhancing the implicit sentiment analysis capabilities of LLMs. In addition, we compile a more practical and complex benchmark for ISA tasks. It spans multiple domains and includes bilingual corpora in both English and Chinese. Our framework is universal and minimalist, and achieves a new state-of-the-art by significantly outperforming previous methods in both zero-shot and fine-tuning settings. It also significantly reduces the tendency to over-interpret emotions.

Keywords

implicit sentiment analysis, large language models, prompt engineering, case-based reasoning, memory

1. Introduction

Sentiment analysis (SA) aims at detecting the sentiment polarity of a given text. SA can be divided into explicit SA (ESA) and implicit SA (ISA), with the former being the mainstream task where sentiment expressions are explicitly present in the text [1]. In contrast to ESA, ISA is more challenging because sentences in ISA may not contain factual descriptions that directly express clear opinions or sentiments [2]. there may also be problems with ambiguous referential components. For example, given the text '*No one can stop him from rampaging like a bull.*' as shown in Figure 1, '*him*' could refer to a particular athlete or a friend of the speaker. Some sentences may also contain literary devices such as irony, metaphor, quotes, rhetorical questions, etc., which serve semantic expression needs but also pose challenges for sentiment analysis tasks.

Traditional research on implicit sentiment analysis has leaned on paradigms such as attention mechanisms and feature extraction [3, 4, 5, 6]. However, these methods struggle to effectively process implicit sentiment data. As we can see in the example, the phrase '*like a bull*' is a simile used to describe a person's behavior. In contrast, humans can infer the sentence sentiment as positive by referring to prior knowledge (past cases) and engaging in multi-step questioning and reasoning processes, given a certain possible domain, such as '*competition*'. Based on human experience in performing sentiment analysis, we argue that there should be two critical elements for tackling ISA tasks: **memorized prior knowledge** and **multi-step logical reasoning**.

ICCBR CBR-LLM'24: Workshop on Case-Based Reasoning and Large Language Model Synergies at ICCBR2024, July 1, 2024, Mérida, Mexico

*Corresponding author.

†These authors contributed equally.

weicheng@cau.edu.cn (W. Cheng); Muhai5100@163.com (P. Li); hejin_liu@cau.edu.cn (H. Liu); GalaxyNaomi@163.com (H. Cheng); caizh@cau.edu.cn (Z. Cai); wenjuan@cau.edu.cn (J. Wen)

0000-0002-8190-5367 (W. Cheng); 0009-0007-7869-5411 (P. Li); 0009-0001-0433-0678 (H. Liu); 0009-0004-7186-3169 (H. Cheng); 0000-0002-2282-2158 (Z. Cai); 0000-0002-4199-2988 (J. Wen)

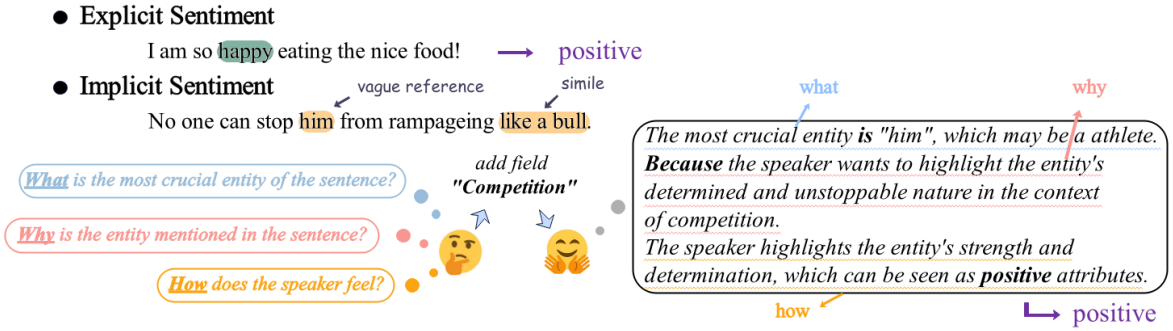

Figure 1: Detecting the explicit and implicit sentiment polarities towards the sentence. Explicit emotions can be directly inferred, while detecting implicit emotions requires acquiring domain priors and simulating human memory reasoning, considering step by step '*What, Why and How*'.

The advent of large language models (LLMs) such as ChatGPT [7] has made significant progress in realizing the above two key factors. LLMs are recognized not only for their extensive prior knowledge [8], but also for their ability to maintain consistent dialogue. This makes it possible to equip them with the memory of past cases [9]. Some studies [10, 11] have also shown the remarkable common sense understanding and logical reasoning of LLMs. In addition, the Chain-of-Thought (CoT) method has revealed LLMs' potential for complex logical reasoning [12, 13], which can be harnessed through strategic prompting. Inspired by these ideas, THOR [14] designed a least-to-most prompt process to get LLMs to explore aspects, opinions, and sentiment polarity of sentences.

While THOR has made progress in ISA tasks, there are cases where its performance is not effective. This problem motivates us to rethink the ISA task from the perspective of simulating human memory and reasoning. First and foremost, domain knowledge is essential for judging emotional polarity. Consider the sentence '*No one can stop him from rampaging like a bull.*' again, its interpretation and associated emotions vary significantly between contexts. In a competitive setting, this might praise an athlete's impressive performance and convey a positive sentiment. Conversely, in daily life, it might criticize someone's careless behavior, reflecting a negative sentiment. Without the relevant domain or context, humans would also experience cognitive ambiguity in decision making. Furthermore, people will subconsciously recall past cases and use them to identify the most crucial entity that evokes emotion in a sentence (e.g., '*him, which may refer to an athlete*'). They then use empathy to understand why the speaker mentioned that entity by probing for its true intent (e.g., '*The speaker wants to highlight this entity's determined and unstoppable nature in relation to competition.*'). Finally, based on the previous inferences and past cases, we can discern the underlying emotions (e.g., *'which can be seen as positive attributes*').

Based on our findings, we propose a universal *What, Why, and How* three-question (3Q) framework with a memory function for past cases. It is structured in three steps: based on domain priors, 1) identify *what* is the most crucial entity in the sentence, 2) infer *why* the speaker wants to mention it, 3) uncover *how* the hidden emotion is like. During the above process, historical queries and responses are stored in memory as past case pairs. These pairs allow the retrieval and creation of improved prompts for any new query. This minimalist approach, which mirrors human memory and reasoning, effectively captures the essence of sentence emotions and reveals hidden meanings, thereby simplifying the judgment of sentiment polarity.

Moreover, to evaluate the effectiveness of the proposed framework, we enable a Chinese-English bilingual dataset derived from several official datasets. Experimental results show that the 3Q framework sets a new state-of-the-art (SOTA) benchmark in both supervised fine-tuning and zero-shot settings. Furthermore, it significantly outperforms THOR and Direct on F1 and neutral F1 metrics, while mitigating the negative effects of excessive emotion interpretation.

To sum up, our contributions are:

- We propose a *What, Why, and How* three-question (3Q) framework with a memory mechanism

for past cases. It mimics human memory and reasoning and puts LLMs in the shoes of speakers for ISA tasks. It refines ISA tasks into three simple steps, each coupled with a memory function: based on domain priors, 1) identify *what* the most crucial entity is in the sentence, 2) infer *why* the speaker wants to mention it, 3) uncover *how* the implied emotions are like.
- We enable a more practical and complex benchmark for ISA tasks, where it includes both English and Chinese languages, covering 28 domains or scenes.
- Extensive evaluations on the benchmark reporting SOTA results in both zero-shot and fine-tuning settings. Experimental results show that the 3Q framework can effectively handle various complex situations, such as over-interpretation of emotions.

2. Related Work

2.1. Implicit Sentiment Analysis (ISA)

Current research on ISA task predominantly centers on deep learning based methods. For example, the SCAPT model [5] utilizes supervised contrastive pretraining to better capture implicit and explicit sentiment orientations towards specific aspects. The CLEAN model [6], using causal intervention, eliminates confounding causal effects in the corpus, suppresses sentiment polarity judgments influenced by explicit cues, and extracts pure causal effects between sentences and emotions. However, these studies primarily focus on extracting features from the plain meaning of the text using a combination of neural networks and attention mechanisms, overlooking the essential role of human logical reasoning and extensive external knowledge in addressing ISA task.

We observe that many implicit sentiment analysis models, including THOR, are based on traditional fine-grained sentiment analysis, such as Aspect-Based Sentiment Analysis (ABSA) tasks. We find that by referencing these fine-grained sentiment analysis tasks, decomposing the sentiment analysis of the corpus into key emotional elements such as 'aspect words' can improve the accuracy of sentiment analysis. However, ABSA requires manual annotation of aspect words, which is difficult to achieve in large datasets. Inspired by human memory and reasoning, we propose 3Q, a *What, Why, and How* three-question framework that integrates a memory function for past cases. It is minimalist and universal, facilitating efficient implicit sentiment analysis.

2.2. Large Language Models (LLMs) with Case-Based Reasoning (CBR)

There are several common methods of case-based reasoning with LLMs. Few-shot prompting involves incorporating case demonstrations into prompts to guide the model toward better performance [15]. Retrieval-augmented generation (RAG) is used to address complex and knowledge-intensive tasks, where an external knowledge system is established to access case examples from external knowledge sources to aid in inference [16]. Previous research has combined the above techniques to enable LLMs to tackle more complex problems such as mathematical applications and ethical reasoning [17, 18]. For more complex ISA tasks, however, these approaches are difficult to apply. This is because many implicit sentences involve multiple ambiguous entities and even complex literary techniques like metaphor, irony, hyperbole, and quotation. To deal with these situations effectively, we need not only classical paradigms to refer to and learn by transfer, but also multi-step logical reasoning [14] to dive into implicit emotions.

The rise of the Chain-of-thought (CoT) has been effective in alleviating these problems. It enhances the multi-step reasoning ability of LLMs by inducing models that simulate human step-by-step reasoning processes to reach conclusions [19, 20]. Research on ISA based on CoT with LLM is limited. The only THOR model, employing a three-hop reasoning pattern, induces implicit aspects and opinions in the corpus, and elicits sentiment polarity judgments on the language [14]. However, THOR requires excessive prior target information in the corpus. It is idealistic and restrictive because real-world sentences often lack specific targets and require manual annotation. In addition, since THOR emphasizes Least-to-Most prompting, each conversation is independent, which in turn affects reasoning capabilities

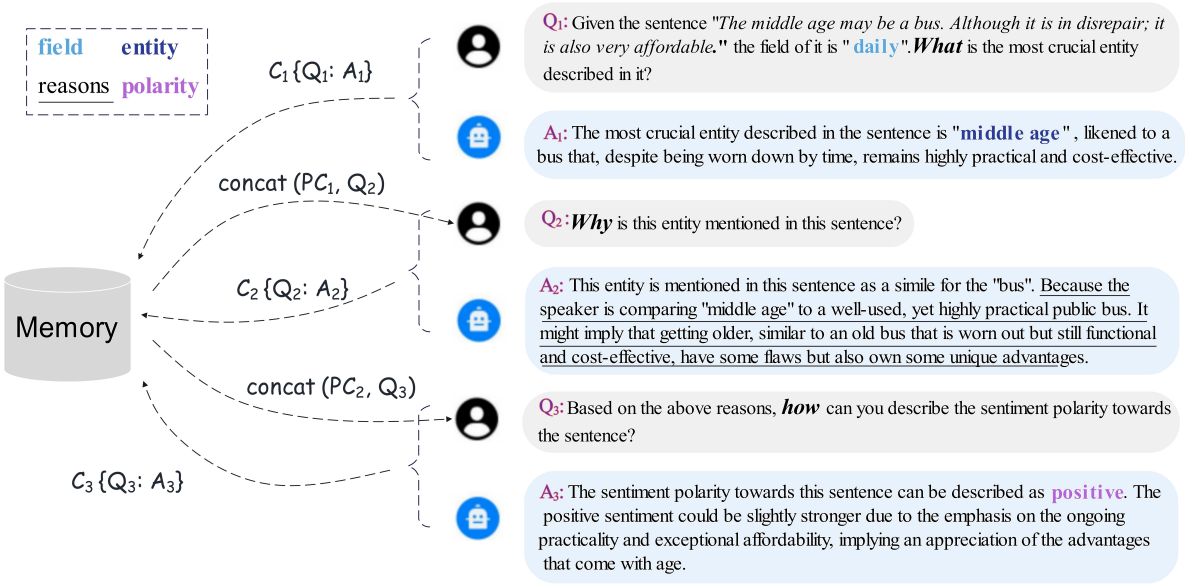

Figure 2: An illustration of 3Q framework for reasoning implicit sentiment.

of LLMs [21]. The above issues prompt us to rethink the ISA task in terms of human memory and reasoning. We propose a *What, Why, and How* three-question (3Q) framework that incorporates a memory function for past cases. Not only does it effectively mimic the way humans think when faced with an ISA task, but it also drastically improves emotional over-interpretation. In addition, the memory mechanism allows LLMs to learn classic paradigms from past cases. This helps to uncover hidden emotions.

3. 3Q Framework with a Memorized Mechanism

3.1. Task Definition

The task of Implicit Sentiment Analysis (ISA) is to determine the sentiment polarity of a given X sentence, categorizing it as positive, neutral, or negative. A standard approach without memory is to use a Direct Prompt template as input for LLMs, typically in the following form:

> Given the sentence X, what is the sentiment polarity towards it?

LLMs should return the answer via: $\hat{y} = \arg\max p(y \mid X)$.

3.2. Implementation of Memory

In the 3Q framework, users sequentially ask three questions Q_i $(i = 1, 2, 3)$ to query the model for sentiment analysis of the input sentence X. At each step, the questions Q_i and the corresponding model-generated answers A_i are stored in memory M as past case pairs. If there are past case pairs in memory, LLMs will automatically retrieve them from memory and integrate them into the current question. The memory implementation consists of the following components:

Memory M: Memory M is a dynamically expanding table managed by the LLM itself that contains past case pairs. The i-th case pair is denoted as $C_i\,(Q_i, A_i)$. The memory M contains all these past case pairs for future reference. It supports the store operation $M.append(x)$ and the fetch operation $f_M(x)$ itself. These operations can be achieved using simple key-value lookups and prompt concatenation. Figure 2 omits the mathematical formulas of these two operations for simplicity.

Past Case Pairs PC_i: All existing cases stored in memory M can be concatenated to construct PC_i, which represents the totality of past information. It can be represented as: $PC_i = f_M(C_1) + f_M(C_2) + ... + f_M(C_i)$. PC_i can also be combined with the next question Q_{i+1} as a new query.

Combiner $concat(x_1, x_2)$**:** It supports combining past case pairs PC_i and questions Q_i as a new query.

The prompt-based memory can be structured as follows:

<C_1> [INST] {Q_1} [/INST] {A_1} </C_1>
<C_2> [INST] {Q_2} [/INST] {A_2} </C_2>
<C_3> [INST] {Q_3} [/INST] {A_3} </C_3>

It is worth noting that the structure of the prompt-based memory will vary slightly to accommodate the different prompt formats required by different LLMs.

3.3. 3Q Framework

In our novel 3Q framework, we aim to have the LLM identify the most crucial entity $\hat{e}$ that influences sentence sentiment, then ask it to put itself in the speaker's shoes to uncover the reasons $\hat{r}$. Finally, the LLM should determine the sentiment polarity $\hat{y}$ based on these prior inferences. We break this process down into three steps as follows:

What. We first provide the sentence X along with its corresponding field f, and then we ask the LLM to identify the most critical entity e within the sentence:

> Given the sentence X, the field of it is f. What is the most crucial entity described in it?

This step is represented as $\hat{e} = \text{argmax}\ p(e|Q_1(X, f))$, where $\hat{e}$ is the model's inference of the entity e, explicitly stating what the entity is and its possible reasons. Note that the entity needn't be explicitly stated in the sentence; instead, it can be generalized and abstracted based on the semantic content of the sentence. After this step, the query and answer are stored in memory in the form of case pairs: $M = M.\,\text{append}\,(C_1\,(Q_1, \hat{e}))$.

Why. Now, based on C_1, we position the model from the speaker's perspective. In this step, we ask the LLMs to uncover the reasons why the speaker mentioned this particular entity $\hat{e}$:

> PC_2. Why is this entity mentioned in this sentence?

This step can be formulated as: $\hat{r} = \text{argmax}\ p(r|PC_1, Q_2)$, where $\hat{r}$ represents the model's inference regarding the reason for mentioning the entity. After this step, the query and response are stored in the form of case pairs in the memory: $M = M.\,\text{append}\,(C_2\,(Q_2, \hat{r}))$.

How. With the PC_2 as the premise, we prompt LLMs to infer the sentiment polarity y as the final outcome:

> PC_3. Based on the above reasons, how would you describe the sentiment polarity towards the sentence?

This step can be formulated as: $\hat{y} = \text{argmax}\ p(y|PC_2, Q_3)$, where $\hat{y}$ is the sentiment polarity ultimately predicted by the model. After this step, the query and response are stored in the form of case pairs in the memory: $M = M.\,\text{append}\,(C_3\,(Q_3, \hat{y}))$.

The 3Q framework combines the memory mechanism with CoT [22], allowing LLMs to use their conversational consistency capabilities to learn information from past cases. In addition, the prompt itself is simple and universal. In contrast, THOR emphasizes the Least-to-Most prompt [23]. This results in each dialog being independent, which in turn affects the reasoning capabilities of LLMs [21].

Table 1
Statistics on three English datasets and two Chinese datasets. 'Pos', 'Neg' and 'Neu' represent the number of positive sentences, negative sentences and neutral sentences, respectively. 'Implicit Sen' means the percentage of implicit sentences.

Dataset	Pos	Neu	Neg	Total	Implicit Sen(%)
• ***English***					
SemEval-2014 Task 4	995	511	769	2289	0.501
SemEval-2015 Task 9	-	147	120	267	1.000
Twitter US Airline Sentiment	-	325	109	434	0.251
• ***Chinese***					
SMP-ECISA 2021	355	877	921	2153	0.501
CCL2018-Chinese-Metaphor-Analysis	646	122	111	879	0.827
Total	1996	1982	2030	6008	-

4. Experiment

4.1. Datasets

We construct a bilingual dataset in both Chinese and English for ISA tasks. It spans multiple domains and includes high-quality explicit and implicit corpora. We select 6,000 samples from a number of publicly accessible datasets: SemEval-2014 Task 4 [1], SemEval-2015 Task 9 [2], SMP-ECISA 2021 [1], CCL2018-Chinese-Metaphor-Analysis [2], and Twitter US Airline Sentiment [3]. Among them, 4000 samples for training and 2000 for testing. In both the training and testing sets, the ratio of explicit sentences to implicit sentences is 1:1. The ratio of English sentences to Chinese sentences is also 1:1. Each sample is classified into one of three sentiment polarities: positive, negative, or neutral. The ratio of these sentiment polarities in the training set and test set is 1:1:1. The details of the dataset can be found in Table 1.

As we analyzed before, one sentence can convey different meanings or emotions in different fields or contexts. It is unrealistic to perform sentiment analysis without specifying the related field. None of the existing official datasets meet this requirement. To address this issue, we annotate each sample with its corresponding field. Specifically, samples from SemEval-2014 Task 4 are labeled as either 'laptops' or 'restaurants', while those from Twitter US Airline Sentiment are marked as 'aviation'. Chinese datasets represented by SMP-ECISA 2021 have already specified the field or topic in its official description. Therefore, samples from these datasets are annotated accordingly, including fields such as 'daily', 'travel', 'politics', etc. Finally, we obtain a bilingual benchmark dataset [4] with 28 different domains.

4.2. Models

We choose the Llama 2-CHAT model from Hugging Face as our backbone LLM [24]. It is available in three sizes: 7B, 13B, and 70B. We also experiment with a leading closed source model, ChatGPT [25]. Note that ChatGPT does not release its model parameters, and we use it in the querying way via the gpt-3.5-turbo-1106 API. We also compare to the current classic baselines, including BERT-SPC [26] and DistilBERT [27]. Given the model's ability to encode a Chinese and English and to load a bilingual tokenizer, we find only these two pre-trained models in the open source community.

[1]https://github.com/sxu-nlp/ECISA2021
[2]https://github.com/DUTIR-Emotion-Group/CCL2018-Chinese-Metaphor-Analysis
[3]https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
[4]https://github.com/qinfengsama/3Q-framework

Table 2
Results on Zero-shot setting. Best results are marked in bold. 'F1-neu' means neutral F1. 'ZeroCoT' means prompting LLMs with the zero-shot CoT, *'let's think step bu step'* [28].

		ISA		All	
		F1	F1-neu	F1	F1-neu
Llama2-7B-CHAT	Direct	47.91	11.89	51.43	12.72
	THOR	52.84	13.87	54.76	13.83
	ZeroCoT	64.10	52.77	68.09	54.11
	3Q	**72.51**	**67.66**	**77.24**	**70.62**
Llama2-13B-CHAT	Direct	68.05	53.10	71.34	54.39
	THOR	52.38	14.10	55.00	14.13
	ZeroCoT	69.73	59.97	74.72	63.27
	3Q	**74.30**	**70.46**	**79.61**	**73.85**
Llama2-70B-CHAT	Direct	75.00	66.98	78.79	68.46
	THOR	59.66	29.61	62.31	31.01
	ZeroCoT	69.12	55.42	71.59	54.87
	3Q	**79.73**	**76.86**	**84.37**	**81.09**
gpt-3.5-turbo-1106	Direct	73.06	65.56	74.83	65.76
	THOR	60.39	21.88	59.42	17.20
	ZeroCoT	**82.41**	**75.52**	**84.89**	**77.98**
	3Q	79.70	72.06	83.07	75.23
DistilBERT		42.85	14.92	48.31	14.68
BERT-SPC		50.85	29.86	55.22	31.00

4.3. Baselines

In both zero-shot and fine-tuning settings, we compare our methods with two baseline methods: THOR and Direct. THOR is the only method that combines LLMs and CoT. The prompt used for it is the same as the corresponding research [14]. Direct itself has no memory mechanism and no multi-step reasoning. The prompt defined in Section 3.1 is used for comparison. To be fair, we do not provide system messages for all the methods.

In the zero-shot scenario, we also use Zero-Shot CoT, which has no memory mechanism, as the baseline method. We adopt the standard prompt from this study [28]. The only difference is that we concatenate the test question with the prompt '*The sentiment polarity toward the sentence is*' instead of '*The answer is*' as the LLM's input.

4.4. Implementation Details

For the Llama 2-CHAT model, we use a temperature of 1 and top-p of 1. This configuration is consistent across models with 7B, 13B, and 70B parameters. For gpt-3.5-turbo-1106 API, we set temperature to 1, top-p to 1, and leave other parameters (such as frequency-penalty) unchanged.

We use cloud A100x1 GPU (40G) instances, cloud A100x1 GPU (80GB) instances, and cloud A100x4 GPU (40GB) instances to fine tune the Llama2 model with model sizes of 7B, 13B, and 70B. The average tuning time is 1 hour. After fine tuning, it takes an average of 17 hours to perform inference on the 2000-sample test set. For the gpt-3.5-turbo-1106 model, we use the OpenAI API for fine tuning, which takes about 2 hours on average. After tuning, the inference process on the entire test set takes about 5 hours.

To train the BERT baseline, we use the AdamW optimizer with a learning rate of 5e-5. The batch size is set to 64 and the dropout probability is set to p=0.1. We use NVIDIA 3090x1 GPU instances in the cloud.

Table 3
Results on supervised fine-tuning setup.

		ISA		All	
		F1	F1-neu	F1	F1-neu
Llama2-7B-CHAT	Direct	85.87	85.55	90.04	89.42
	THOR	71.96	62.76	76.35	67.17
	3Q	**91.30**	**92.17**	**94.16**	**94.53**
Llama2-13B-CHAT	Direct	86.68	86.76	90.81	90.39
	THOR	77.15	70.42	80.42	72.79
	3Q	**93.47**	**93.71**	**95.39**	**95.26**
Llama2-70B-CHAT	Direct	87.79	86.18	91.24	89.61
	THOR	82.23	78.91	86.11	81.97
	3Q	**94.01**	**93.06**	**95.66**	**94.83**
gpt-3.5-turbo-1106	Direct	92.20	90.69	93.89	92.19
	THOR	86.95	86.95	89.71	85.78
	3Q	**93.61**	**92.74**	**95.60**	**94.67**
DistilBERT		69.13	76.79	74.94	82.50
BERT-SPC		69.97	78.20	76.42	83.09

4.5. Evaluation Metrics

We take the F1 as one of the evaluation metrics. During the experiments, we observe a significant difference in the neutral F1 compared to the positive and negative F1s. Therefore, we also include the neutral F1 in our evaluation metrics, recognizing its importance in assessing the interpretation of excessive emotion.

5. Experimental Results

5.1. Results on Zero-shot Reasoning

Table 2 presents the comparison results under zero-shot settings. It is evident that four methods that combine LLMs with prompt engineering significantly outperform current state-of-the-art (SoTA) baselines. Among these, 3Q stands out for its impressive performance. Specifically, when equipped with the 7B Llama 2-CHAT model, it leads the best performing baseline (BERT-SPC) by 21.66%. As the model scale increases, the gap in F1 between the two also widens, peaking at a 28.88% difference with the 70B parameter model. This is consistent with the study's conclusion [14] that reasoning-based methods can achieve significant advances over traditional non-reasoning methods.

On all three scales of the Llama2 model, 3Q outperforms the other three prompt methods and achieves a new SOTA performance. Specifically, in the ISA setting, its F1 score is approximately 7.86%, 11.86%, and 20.55% higher than Zero-shot CoT, Direct, and THOR, respectively. This suggests that the combination of both memory mechanisms and multi-step reasoning can provide a greater improvement in implicit sentiment analysis than neither memory mechanisms nor multi-step reasoning. Interestingly, the difference in F1 score between 3Q and THOR is the most significant. This is because a sentence usually contains multiple targets or entities. Sometimes sentence-level sentiment may contradict aspect-level sentiment. We argue that THOR's ability is limited by its own prompt architecture, which is more concerned with aspect-level sentiment. This is likely to lead to significant errors. In contrast, 3Q does not initially specify entities. Instead, it encourages LLMs to infer the entity that best identifies sentence-level sentiment based on the domain. In the subsequent 'why' section, LLMs are prompted to understand and contextualize the reasons behind the most critical entity assertions. In addition, 3Q's memory mechanism helps to model the relationship between the most crucial entity and other entities in the context, thereby improving the understanding of sentiment at the sentence level. Similar benefits

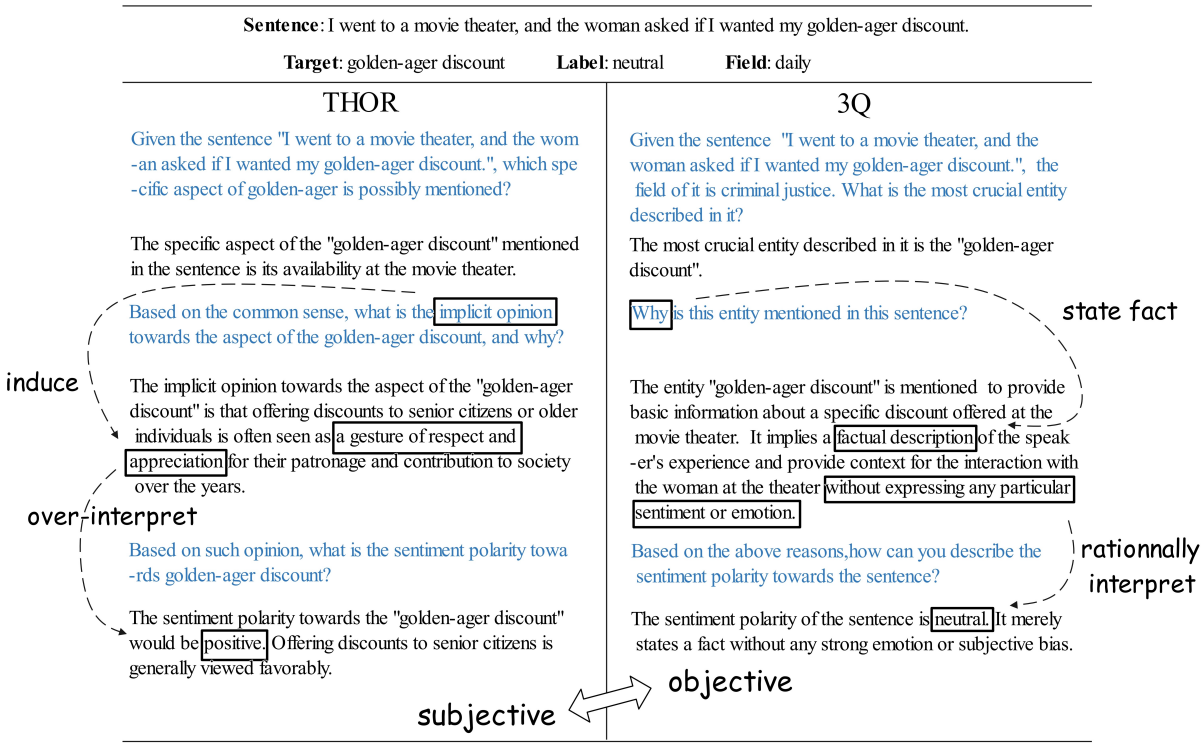

Figure 3: A visual comparison of emotional over-interpretation. It analyzes the effects and results of the THOR and 3Q methods using an implicit sentence as an example.

are seen in both Zero-Shot CoT and Direct.

In addition, we note that as the model scale increases, not only does the gap between 3Q and THOR widen, but Direct has also surpassed THOR. These two phenomena are primarily due to the ever-widening neutral F1 gap, which is related to the 'over-interpretation of neutral emotions' phenomenon that we discuss later.

5.2. Results on Supervised Fine-tuning

The comparisons under fine-tuning are shown in Table 3. Overall, high-quality instructional fine-tuning significantly improves performance. This is consistent with the conclusions of the Flan-T5 study [29], which suggests that model performance improves significantly with more fine-tuning tasks. Taking the 7B model as an example, instruction fine-tuning increases the F1 of 3Q, THOR, and Direct by 18.79%, 19.12%, and 37.96%, respectively. Similar trends are observed in other configurations, and 3Q achieves SOTA performance when equipped with the 70B model. Under this configuration, the F1 score reaches an impressive 94.01%, surpassing that of BERT-SPC by 24.04%. Furthermore, after comparing three prompt methods, we find that 3Q consistently outperforms THOR by at least 10% in F1 score, a stable improvement that is also reflected in the zero-shot settings. In addition, a major benefit of prompt tuning is the increased gap between 3Q and Direct: 3Q leads Direct by 5.43% F1 even on the smallest 7b model.

5.3. Emotional Over-interpretation

Table 2 shows that the F1 score for neutral sample sentiment classification remains low regardless of the configuration. Performance improves slightly when the model parameters reach a large size, as seen in Llama 2-CHAT 70B and gpt-3.5-turbo-1106. This suggests that accurate identification of neutral sentiment continues to be a significant obstacle in ISA tasks.

It is worth noting that THOR maintains an average of 20% F1-neu in most cases. Such poor perfor-

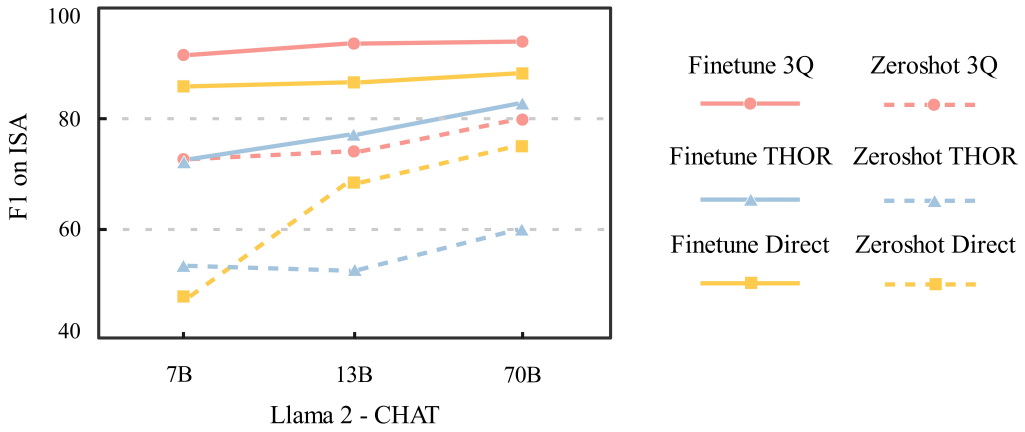

Figure 4: Influence of LLM scales.

mance results from its second step, the induction of the implicit opinion. The original prompt is, '*Based on common sense, what is the implicit opinion towards the mentioned aspect of t, and why?*'. The phrase '*implicit opinion*' causes LLMs to infer inappropriate implicit opinions from neutral sentences, which introduces great interference and error into the next step of judging sentiment polarity.

In contrast, 3Q outperforms THOR by nearly 50% in average neutral F1 and establishes a new SOTA score of 76.86% on the 70B Llama 2-CHAT model. It is also 9.88% higher than Direct under similar configurations. This is because in the 'why' section, 3Q does not introduce subjective interventions like '*implicit opinion*'. Instead, it only leads LLMs to infer the reasons for mentioning entities. In neutral sentences, the most critical entity tends to state a specific fact or serve a specific purpose. It is usually mentioned for objective reasons. Therefore, 3Q can effectively mitigate the problem of over-interpreting emotions. Figure 3 visualizes the entire analysis. While Direct also has similar benefits with a relatively high neutral F1 score, it is still 10% lower on average than 3Q. This suggests that the combination of both memory mechanisms and multi-step reasoning can provide a greater improvement in implicit sentiment analysis than neither memory mechanisms nor multi-step reasoning.

5.4. Influence of Different Model Sizes of LLMs

In Figure 4 we examine the influence of different LLM scales. As the model scale increases, the F1 for all three methods generally show a growth trend. This is consistent with the existing findings of CoT prompting methods [12, 13, 30], suggesting that larger LMs have more extensive prior knowledge and improved logical reasoning abilities. In addition, high-quality instruction fine-tuning leads to a significant improvement in F1 scores. This aligns with the conclusions of the Flan-T5 study[29], which proposes that instructional fine-tuning is a general method for improving the performance and usability of pre-trained language models. Among these prompt methods, 3Q experiences the largest increase, with an average improvement of 18%. Most strikingly, its zero-shot performance approaches that of THOR after instructional fine-tuning, suggesting that 3Q has significant potential and a higher ceiling for performance.

5.5. Error Analysis

We examine the error rates and error cases of the 3Q model in both zero-shot and supervised fine-tuning settings on our test set. Errors are categorized into three types: Type A error represents over-analyzing neutral corpora, where the model mistakenly interprets neutral content as positive or negative sentiment; Type B error occurs when the model fails to capture sentiment clues in the corpora and misjudges negative and positive as neutral; Type C error means the model confusing two non-neutral emotions, labeling positive as negative or vice versa. The distribution of these three types of errors is illustrated in Figure 5.

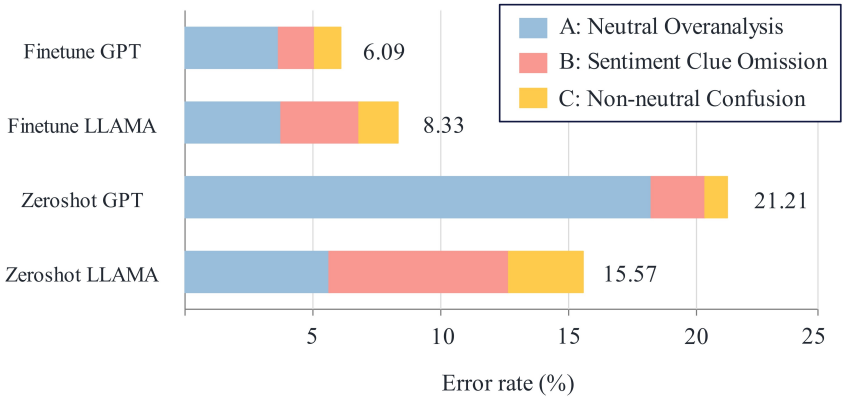

Figure 5: Error analysis.

From the experimental results, it is evident that in the zero-shot setting, the gpt-3.5-turbo-1106 is predominantly associated with Type A errors, i.e, among the total error rate of 21.21%, Type A errors accounts for 18.72%. Supervision Fine-tuning reduces the Type A error rate to a low level (3.64%), i.e., eliminating approximately 80% of Type A errors. In contrast, the proportions of Type B and Type C errors show no significant changes before and after supervised fine-tuning. Moreover, these two error types account for a smaller portion of the overall errors, possibly due to the inherent difficulty of the self-built corpora and the limitation of model capabilities.

In contrast, the Llama2-CHAT-70B model, unlike gpt-3.5-turbo-1106, tends to misclassify neutral sentences. In the zero-shot setting, the Llama 2-CHAT-70B model exhibits a Type B error rate of 7.09%, which surpasses the Type A error rate (5.79%). Even after supervised fine-tuning, the Type B error rate (3.09%) exceeds the zero-shot results of the gpt-3.5-turbo-1106 model (1.60%). This suggests that the Llama 2-CHAT-70B model is prone to misclassifying sentiment-unnoticed samples as neutral.

6. Conclusion and Future Work

In this research, we propose a *What, Why, and How* three-question (3Q) framework that incorporates a memory mechanism for past cases. Specifically, 3Q performs a sequential three-step prompt by simulating the human memory and reasoning process. It first determines *what* the most critical entity in the sentence is, then infers *why* the speaker mentions it. Finally, 3Q uncovers *how* the hidden emotions are based on the past cases. During the reasoning process, historical queries and responses are stored in memory as past case pairs. These pairs enable the retrieval and creation of enhanced prompts for any new query, thereby enhancing the implicit sentiment analysis capabilities of LLMs. To verify the effectiveness of it, we construct a Chinese-English bilingual dataset based on several official datasets. 3Q achieves a new SoTA under both supervised fine-tuning and zero-shot settings. Experimental results show that it significantly outperforms THOR, Direct and Zero-Shot CoT in F1 scores, and effectively reduces the negative effects of over-interpretation of emotions. We release the dataset in an associated repository and hope that the accessibility of the dataset will encourage the community to evaluate novel methods for implicit sentiment analysis. In the future, we will extend this approach to the real-world document understanding. In addition, we plan to further explore the ISA task by simulating the human nonlinear reasoning process from the perspective of Tree of Thoughts (ToT) and Graph of Thoughts (GoT).

References

[1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: P. Nakov, T. Zesch (Eds.), Proceedings

of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35. URL: https://aclanthology.org/S14-2004. doi:10.3115/v1/S14-2004.

[2] I. Russo, T. Caselli, C. Strapparava, Semeval-2015 task 9: Clipeval implicit polarity of events, in: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 2015, pp. 443–450.

[3] R. He, W. S. Lee, H. T. Ng, D. Dahlmeier, Effective attention modeling for aspect-level sentiment classification, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1121–1131. URL: https://aclanthology.org/C18-1096.

[4] H. Tang, D. Ji, C. Li, Q. Zhou, Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6578–6588. URL: https://aclanthology.org/2020.acl-main.588. doi:10.18653/v1/2020.acl-main.588.

[5] Z. Li, Y. Zou, C. Zhang, Q. Zhang, Z. Wei, Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 246–256. URL: https://aclanthology.org/2021.emnlp-main.22. doi:10.18653/v1/2021.emnlp-main.22.

[6] S. Wang, J. Zhou, C. Sun, J. Ye, T. Gui, Q. Zhang, X. Huang, Causal intervention improves implicit sentiment analysis, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6966–6977. URL: https://aclanthology.org/2022.coling-1.607.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[8] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, Large pre-trained language models contain human-like biases of what is right and wrong to do, Nature Machine Intelligence 4 (2022) 258–268.

[9] A. Madaan, N. Tandon, P. Clark, Y. Yang, Memory-assisted prompt editing to improve gpt-3 after deployment, arXiv preprint arXiv:2201.06009 (2022).

[10] B. Paranjape, J. Michael, M. Ghazvininejad, L. Zettlemoyer, H. Hajishirzi, Prompting contrastive explanations for commonsense reasoning tasks, arXiv preprint arXiv:2106.06823 (2021).

[11] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, H. Hajishirzi, Generated knowledge prompting for commonsense reasoning, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3154–3169. URL: https://aclanthology.org/2022.acl-long.225. doi:10.18653/v1/2022.acl-long.225.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems 35 (2022) 24824–24837.

[13] Z. Zhang, A. Zhang, M. Li, A. Smola, Automatic chain of thought prompting in large language models, arXiv preprint arXiv:2210.03493 (2022).

[14] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, T.-S. Chua, Reasoning implicit sentiment with chain-of-thought prompting, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1171–1182. URL: https://aclanthology.org/2023.acl-short.101. doi:10.18653/v1/2023.acl-short.101.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[17] A. Madaan, N. Tandon, P. Clark, Y. Yang, Memory-assisted prompt editing to improve gpt-3 after deployment, 2022. arXiv:2201.06009.

[18] N. Tandon, A. Madaan, P. Clark, Y. Yang, Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback, arXiv preprint arXiv:2112.09737 (2021).

[19] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, Journal of Machine Learning Research 24 (2023) 1–113.

[20] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).

[21] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).

[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.

[23] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, et al., Least-to-most prompting enables complex reasoning in large language models, arXiv preprint arXiv:2205.10625 (2022).

[24] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv e-prints (2018) arXiv:1810.04805. doi:10.48550/arXiv.1810.04805. arXiv:1810.04805.

[27] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv e-prints (2019) arXiv:1910.01108. doi:10.48550/arXiv.1910.01108. arXiv:1910.01108.

[28] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.

[29] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[30] Y. Fu, H. Peng, A. Sabharwal, P. Clark, T. Khot, Complexity-based prompting for multi-step reasoning, arXiv preprint arXiv:2210.00720 (2022).