Reasoning Implicit Sentiment with Chain-of-Thought Prompting[*]

Hao Fei[1], **Bobo Li**[2], **Qian Liu**[3], **Lidong Bing**[4], **Fei Li**[2†], **Tat-Seng Chua**[1]

[1]Sea-NExT Joint Lab, School of Computing, National University of Singapore
[2]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
[3]Sea AI Lab, [4]DAMO Academy, Alibaba Group
haofei37@nus.edu.sg, boboli@whu.edu.cn, liuqian@sea.com, l.bing@alibaba-inc.com, lifei_csnlp@whu.edu.cn, dcscts@nus.edu.sg

Abstract

While sentiment analysis systems try to determine the sentiment polarities of given targets based on the key opinion expressions in input texts, in implicit sentiment analysis (ISA) the opinion cues come in an implicit and obscure manner. Thus detecting implicit sentiment requires the common-sense and multi-hop reasoning ability to infer the latent intent of opinion. Inspired by the recent chain-of-thought (CoT) idea, in this work we introduce a *Three-hop Reasoning* (THOR) CoT framework to mimic the human-like reasoning process for ISA. We design a three-step prompting principle for THOR to step-by-step induce the implicit aspect, opinion, and finally the sentiment polarity. Our THOR+Flan-T5 (11B) pushes the state-of-the-art (SoTA) by over 6% F1 on supervised setup. More strikingly, THOR+GPT3 (175B) boosts the SoTA by over 50% F1 on zero-shot setting. Our code is open at https://github.com/scofield7419/THOR-ISA.

1 Introduction

Sentiment analysis (SA) aims to detect the sentiment polarity towards a given target based on the input text. SA can be classified into explicit SA (ESA) and implicit SA (ISA), where the former type is the current mainstream task, in which the emotional expressions explicitly occur in texts (Pontiki et al., 2014). Different from ESA, ISA is much more challenging, because in ISA the inputs contain only factual descriptions with no explicit opinion expression directly given (Russo et al., 2015). For example, given a text '*Try the tandoori salmon!*', having no salient cue word, almost all existing sentiment classifier[1] predicts a neutral polarity towards '*the tandoori salmon*'. Human can easily determine the sentiment states accurately, because we always grasp the real intent or opinion behind the texts. Thus, without truly understanding *how* the sentiment is aroused, traditional SA methods are ineffective to ISA.

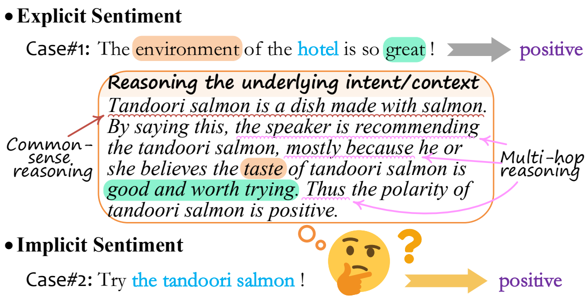

Figure 1: Detecting the explicit and implicit sentiment polarities towards targets. Explicit opinion expression helps direct inference, while detecting implicit sentiment requires common-sense and multi-hop reasoning.

In fact, it is critical to first discover the hidden opinion contexts to achieve accurate ISA. For the explicit case#1 in Fig. 1, it is effortless to capture the overall sentiment picture (e.g., '*environment*' is the aspect, '*great*' is the opinion), and thus can precisely infer the *positive* polarity towards the given target *hotel*. Inspired by such fine-grained sentiment spirit (Xue and Li, 2018; Zhang et al., 2021; Xu et al., 2020), we consider mining the implicit aspect and opinion states. For the implicit case#2 in Fig. 1, if a model can first infer the key sentiment components, e.g., the latent aspect '*taste*', latent opinion '*good and worth trying*', the inference of final polarity can be greatly eased. To reach the goal, the capabilities of **common-sense reasoning** (i.e., infer what is '*tandoori salmon*') and **multi-hop reasoning** (i.e., infer the aspect and then the opinion) are indispensable.

Fortunately, the recent great triumph of pre-trained large-scale language models (LLMs) offers a promising solution. On the one hand, LLMs have been found to carry very rich world knowledge, showing extraordinary ability on common-sense understanding (Paranjape et al., 2021; Liu et al., 2022). On the other hand, the latest chain-of-

[*]The work is substantially supported by Alibaba Group through the Alibaba Innovative Research (AIR) Program.

[†]Corresponding author: Fei Li.

[1]We pre-experiment with total 20 existing SA models.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1171–1182
July 9-14, 2023

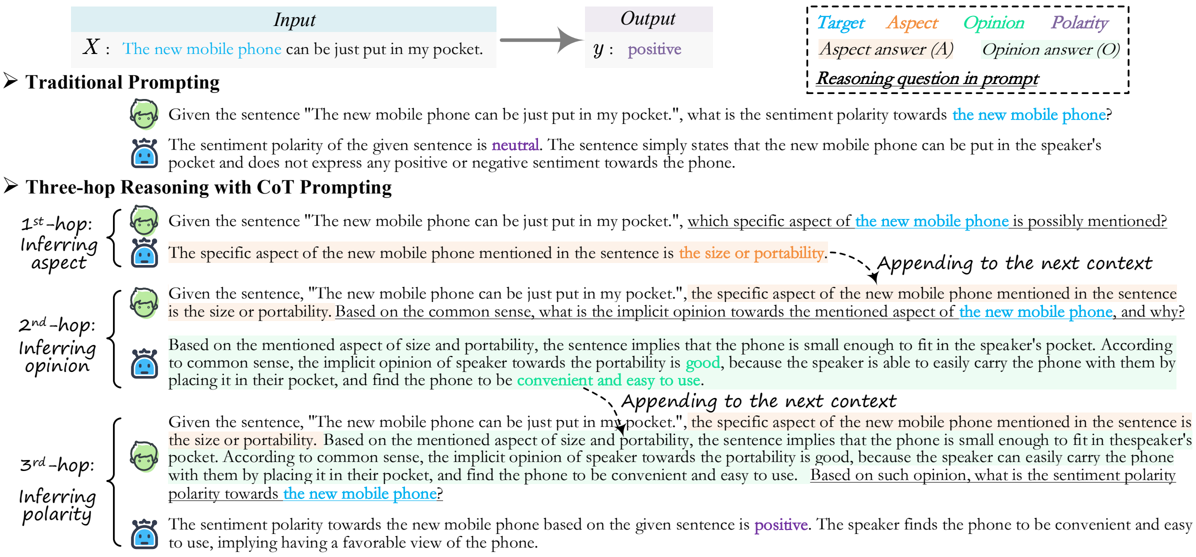

Figure 2: An illustration of our THOR framework for three-hop reasoning of implicit sentiment.

thought (CoT) idea has revealed the great potential of LMs' multi-hop reasoning (Wei et al., 2022; Zhou et al., 2022; Zhang et al., 2023), where an LLM with some prompts can do chain-style reasoning impressively. Built on top of all these successes, in this work we implement a Three-hop Reasoning CoT framework (namely THOR) for ISA. Based on an LLM, we design three prompts for three steps of reasoning, each of which respectively infers 1) the fine-grained aspect of the given target, 2) the underlying opinion towards the aspect, and 3) the final polarity. With such easy-to-hard incremental reasoning, the hidden contexts of the overall sentiment picture can be elicited step by step to achieve an easier prediction of final polarity, which effectively alleviates the difficulties of the task prediction.

To ensure the correctness of each reasoning step, we consider a self-consistency mechanism for CoT inspired by Wang et al. (2022b), which is to select the candidate answers (at each step) with high voting consistency of inferred aspect and opinion. For supervised fine-tuning setup, we further propose a reasoning revising method. We use the intermediate reasoning answers as model inputs to predict the final labels, where the supervision from gold labels will teach LLM to generate more correct reasoning. On supervised fine-tuning setup, our Flan-T5 based THOR improves the current best-performing baseline by more than 6% in F1 score, and such margins are further magnified on zero-shot setup. Most strikingly, our GPT3-based THOR with 175B parameters boosts the baseline to a high-to 51.10% increase of F1 score.

To sum up, this work contributes a multi-hop reasoning solution for implicit sentiment detection, which helps to achieve impressive improvement over the traditional non-reasoning methods. To our knowledge, this is the first attempt to successfully extend the CoT idea to the sentiment analysis community. Our method is simple yet effective, and can be broadly applied to other similar NLP problems without much effort.

2 Three-hop Reasoning Framework

The task of SA (either ESA or ISA) is defined as: given a sentence X with a target term $t \subset X$, a model determines the sentiment polarity y towards t, i.e., *positive*, *neutral* or *negative*. We solve the task using an off-the-shelf LLM with prompt. For the standard prompt-based method, we can construct the following prompt template as LLM's input:

```
Given the sentence X, what is the sentiment polarity towards t?
```

The LLM should return the answer via: $\hat{y}$=argmax$p(y|X,t)$.

2.1 Chain-of-Thought Prompting

Now we consider the CoT-style prompt (Wei et al., 2022; Fu et al., 2022) method for multi-step reasoning. Instead of directly asking LLM the final result of y, in our THOR (cf. Fig. 2) we hope the LLM infer the latent aspect and opinion information before answering the finale y. We here define the intermediate aspect term a and latent opinion expression o. We construct the three-hop prompts as follows.

Step 1. We first ask LLM what aspect a is mentioned with the following template:

```
C_1[Given sentence X], which specific
aspect of t is possibly mentioned?
```

C_1 is the first-hop prompt context. This step can be formulated as A=argmax$p(a|X,t)$, where A is the output text which explicitly mentions the aspect a.

Step 2. Now based on X, t and a, we ask LLM to answer in detail what would be the underlying opinion o towards the mentioned aspect a:

```
C_2[C_1,A]. Based on the common sense,
what is the implicit opinion towards
the mentioned aspect of t, and why?
```

C_2 is the second-hop prompt context which concatenates C_1 and A. This step can be written as O=argmax$p(o|X,t,a)$, where O is the answer text containing the possible opinion expression o.

Step 3. With the complete sentiment skeleton (X, t, a and o) as context, we finally ask LLM to infer the final answer of polarity t:

```
C_3[C_2,O]. Based on the opinion, what
is the sentiment polarity towards t?
```

C_3 is the third-hop prompt context. We note this step as $\hat{y}$=argmax$p(y|X,t,a,o)$.

2.2 Enhancing Reasoning via Self-consistency

We further leverage the self-consistency mechanism (Wang et al., 2022b; Li et al., 2022b) to consolidate the reasoning correctness. Specifically, for each of three reasoning steps, we set the LLM decoder to generate multiple answers, each of which will likely to give varied predictions of aspect a, opinion o as well as the polarity y. At each step, those answers with high voting consistency of inferred a, o or y are kept. We select the one with highest confidence as the context in next step.

2.3 Reasoning Revising with Supervision

We can also fine-tune our THOR when the on-demand training set is available, i.e., supervised fine-tuning setup. We devise a reasoning revising method. Technically, at each step we construct a prompt by concatenating 1) initial context, 2) this step's reasoning answer text and 3) final question, and feed it into LLM to predict the sentiment label instead of going to the next step reasoning. For example, at end of step-1, we can assemble a prompt: [C_1,A, '*what is the sentiment polarity towards t?*']. In the supervision of gold labels, the LLM will be taught to generate more correct intermediate reasoning that is helpful to the final prediction.

	Restaurant		Laptop	
	All	ISA	All	ISA
• ***State-of-the-art baselines***				
BERT+SPC† (110M)	77.16	65.54	73.45	69.54
BERT+ADA† (110M)	80.05	65.92	74.18	70.11
BERT+RGAT† (110M)	81.35	67.79	74.07	72.99
BERT_{Asp}+CEPT† (110M)	82.07	67.79	78.38	75.86
BERT+ISAIV† (110M)	81.40	69.66	77.25	78.29
BERT_{Asp}+SCAPT† (110M)	83.79	72.28	79.15	77.59
• ***Prompt-based methods***				
BERT+Prompt (110M)	81.34	70.12	78.58	75.24
Flan-T5+Prompt (250M)	81.50	70.91	79.02	76.40
Flan-T5+Prompt (11B)	84.72	75.10	82.44	78.91
• ***CoT-based methods***				
Flan-T5+THOR (250M)	82.98	71.70	79.75	67.63
Flan-T5+THOR (11B)	**87.45**	**79.73**	**85.16**	**82.43**
w/o SelfConsistency	86.03	77.68	84.39	80.27
w/o Reason-Revising	86.88	78.42	84.83	81.69

Table 1: F1 results on supervised fine-tuning setup. Best results are marked in bold. Scores by model with † are copied from Li et al. (2021).

3 Experiments

Setups We experiment on the benchmark SemEval14 Laptop and Restaurant datasets (Pontiki et al., 2014), where all the instances are split into explicit and implicit sentiment by Li et al. (2021). Since the encoder-style BERT cannot generate texts to support CoT, we use encoder-decoder style Flan-T5[2] as our backbone LLM. We also test with GPT3 (Brown et al., 2020) and ChatGPT (Ouyang et al., 2022). We used four versions of Flan-T5: 250M (base), 780M (large), 3B (xl) and 11B (xxl), and four versions of GPT3: 350M, 1.3B, 6.7B and 175B. Note that GPT3 does not release the model parameters, and we use it in the prompting manner via the API[3]. This also means that we cannot perform supervised fine-tuning with GPT3. We compare with the current best-performing baselines, including: BERT+SPC (Devlin et al., 2019), BERT+ADA (Rietzler et al., 2020), BERT+RGAT (Wang et al., 2020), BERT_{Asp}+CEPT (Li et al., 2021), BERT+ISAIV (Wang et al., 2022a) and BERT_{Asp}+SCAPT (Li et al., 2021). We consider both the supervised fine-tuning and zero-shot setups. We adopt the F1 as the evaluation metric. On the few-shot setup, we re-implement the baselines via their source codes. Our experiments are conducted with 4 NVIDIA A100 GPUs.

[2] https://huggingface.co/docs/transformers/model_doc/flan-t5

[3] https://beta.openai.com/docs/models/gpt-3

	Restaurant		Laptop	
	All	ISA	All	ISA
• ***State-of-the-art baselines***				
BERT+SPC (110M)	21.76	19.48	25.34	17.71
BERT+RGAT (110M)	27.48	22.04	25.68	18.26
BERT_{Asp}+SCAPT (110M)	30.02	25.49	25.77	13.70
• ***Prompt-based methods***				
BERT+Prompt (110M)	33.62	31.46	35.17	22.86
Flan-T5+Prompt (250M)	54.38	41.57	52.06	31.43
Flan-T5+Prompt (11B)	57.12	45.31	54.14	33.71
• ***CoT-based methods***				
Flan-T5+THOR (250M)	55.86	41.84	52.52	32.40
Flan-T5+THOR (3B)	57.33	42.61	56.36	38.16
Flan-T5+THOR (11B)	61.87	52.76	58.27	40.75
Flan-T5+ZeroCoT (11B)	56.58	47.41	55.53	35.67
GPT3+THOR (175B)	**81.96**	**76.55**	**76.04**	**73.12**

Table 2: Model results on Zero-shot setting. We re-implement the state-of-the-art baselines for the zero-shot performance. 'ZeroCoT' means prompting LLM with the zero-shot CoT, '*let's think step by step*' (Brown et al., 2020).

Results on Supervised Fine-tuning The comparisons are shown in Table 1. It is interesting to see that the BERT with prompt learning underperforms the SoTA baseline BERT_{Asp}+SCAPT. Even the Flan-T5-base (250M) with double-size parameters fails to beat the SoTA. BERT_{Asp}+SCAPT is pre-trained on the large-scale sentiment aspect-aware annotation data, thus showing strong capability on SA. But with our THOR CoT prompting, Flan-T5-base clearly outperforms SoTA. Further, when using the larger LLM, i.e., with 11B parameters, we can find the vanilla prompt-based Flan-T5 surpasses the best baseline. More prominently, Flan-T5-11B with THOR shows significant boosts for ISA, i.e., 7.45%(=79.73-72.28) on Restaurant and 5.84%(=82.43-77.59) on Laptop, with average improvement of 6.65%(7.45+5.84)/2 F1. Also the ablations of the self-consistency and reasoning revising mechanisms indicate their importances in our THOR method.

Results on Zero-shot Reasoning In Table 2 we compare the zero-shot performances. We can find that the improvement of both prompt-based and CoT-based methods over the current SoTA baseline increases dramatically. But overall, the CoT-based methods with our THOR show much more significant improvement on ISA. For example, our Flan-T5-11B THOR system gives over 30% F1 average improvement over the best-performing baseline (BERT_{Asp}+SCAPT) on two datasets. Most strikingly, when THOR is equipped into super-large LLM, i.e., GPT3-175B, we can observe the impressive improvement, near to the level by Flan-T5-11B THOR in supervised setting as in Table 1. Specifically, it boosts the SoTA results by 51.94%(=81.96-30.02) on Restaurant and 50.27%(=76.04-25.77) on Laptop, with an average 51.10%(51.94+50.27)/2 F1 leap.

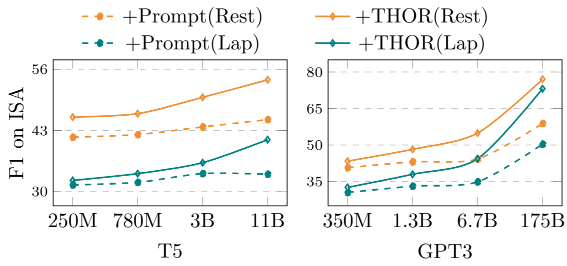

Figure 3: Influences of LLM scales.

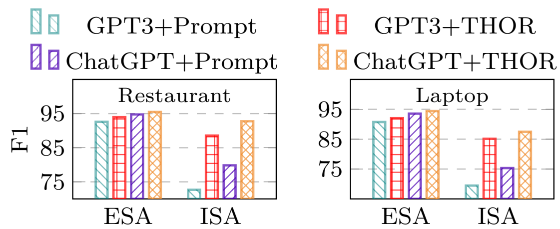

Figure 4: Comparisons between GPT3&ChatGPT on randomly-selected 50 ESA and 50 ISA instances.

Influence of Different Model Sizes of LLMs In Table 1 and 2 we have witnessed the power by using (very) large LLMs. In Fig. 3 we study the influence of different LLM scales. We see that with the increasing model scale, the efficacy of our multi-hop reasoning prompting is exponentially amplified. This coincides much with the existing findings of CoT prompting methods (Wei et al., 2022; Zhou et al., 2022; Fu et al., 2022), i.e., the larger the LMs, the more significant improvement by CoT. Because when the LLM is sufficiently large, the capabilities on common-sense and multi-hop reasoning are greatly developed and strengthened.

Improving ChatGPT with THOR The latest birth of ChatGPT has brought revolutionary advancement in NLP and AI community. Here we compare the improvement of our THOR on GPT3 (175B) and ChatGPT, respectively. In Fig. 4 we show the testing results on 100 testing instances. We can see that both LMs shows very high performances on ESA, and the enhancements by THOR are very limited. But prompting-based GPT3 and ChatGPT still fail much on ISA, where our THOR

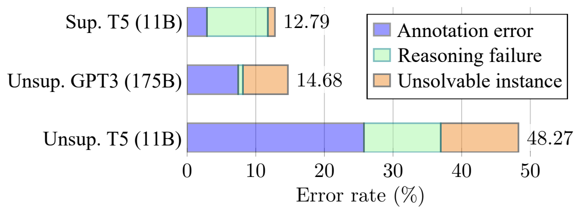

Figure 5: Error analysis.

has improved them on ISA very considerably.

Failure Analysis In Fig. 5 we show the error rates of failure cases when using THOR, where we summarize three error types. The Flan-T5-11B LLM gives 48.27% error rate on zero-shot setup, while it goes down to 12.79% when fine-tuned with supervision. Unsupervised-GPT3 (175B) gives similarity low error rate as with Supervised-T5, while the latter fails much frequently on incapability of reasoning. In contrast to Supervised-T5, the majority of failures in Unsupervised-GPT3 comes from the problematic data annotation. Since Supervised-T5 is fine-tuned with supervision of 'false' labels, it may actually learn the spurious correlations but with higher testing accuracy.

4 Related Work

Sentiment analysis has long been a hot research topic in NLP community (Pang and Lee, 2007; Dong et al., 2014; Shi et al., 2022). While the explicit SA models can make predictions based on the opinion expressions effortlessly, the implicit SA can be much more tricky due to the hidden opinion characteristics (Li et al., 2021; Wang et al., 2022a). And ISA is often more ubiquitous in realistic scenarios. Although efforts have been made to ISA (Li et al., 2021; Wang et al., 2022a), existing work can still be limited to the traditional paradigm of inference. As aforementioned, ISA should be addressed via reasoning, i.e., common-sense and multi-hop reasoning. Thus, this work follows such intuition, targeting solving ISA with a multi-hop reasoning mechanism.

As a key branch of SA, the fine-grained SA has been well explored (Wang et al., 2017; Li et al., 2018, 2022a). The idea of fine-grained SA is to break down the SA into several key sentiment elements, including *target*, *aspect*, *opinion* and *sentiment polarity*, all of which together form a complete sentiment picture in detail (Peng et al., 2020; Fei et al., 2022). This work draws the same spirit of fine-grained SA. We believe the reasoning of implicit sentiment should be an incremental process, inferring the sentiment elements step by step and finally understand the sentiment polarity in an easy-to-hard manner.

Language model pre-training has received increasing research attention for enhancing the utility of downstream applications (Raffel et al., 2020) Most recently, the large-scale language models (LLMs) have shown great potential to the human-level intelligence, e.g., ChatGPT (Ouyang et al., 2022). LLMs have extensively demonstrated to exhibit extraordinary abilities on common-sense understanding (Paranjape et al., 2021; Liu et al., 2022) and multi-hop reasoning (Wei et al., 2022; Zhou et al., 2022). This work implements the implicit sentiment reasoning built upon LMs, based on the latest proposed chain-of-thought (CoT) idea. CoT prompting is a gradient-free technique that induces large LMs to produce intermediate reasoning steps leading to the final answer. Wei et al. (2022) formally study the CoT prompting in language models, in which they elicit LMs to generate coherent series of intermediate reasoning steps that direct to the final answer to the original question.

5 Conclusion

In this paper, we present a *Three-hop Reasoning* prompting framework to achieve the chain-of-thought reasoning process for implicit sentiment analysis. Based on the existing LLM, we design three prompts for three steps of reasoning, each of which respectively infers the fine-grained aspect, the underlying opinion and the final polarity. On the ISA datasets, different LLMs equipped with our THOR show impressive performances over the existing best-performing baselines on both the supervised and zero-shot setups. We show that the larger the LLMs, the more significant improvement by our THOR method.

Acknowledgments

The work is also partially supported by the National Key Research and Development Program of China (No. 2022YFB3103602) and the Sea-NExT Joint Lab at National University of Singapore.

Limitations

THOR helps unleash the full power of LLMs only when being integrated into the large enough models, while on the middle or lower size LLMs, the improvement by THOR will be limited to certain extent, due to the emergence nature of LLMs.

References

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of ACL*, pages 49–54.

Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. 2022. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *CoRR*, abs/2210.00720.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2022a. Diaasq : A benchmark of conversational aspect-based sentiment quadruple analysis. *CoRR*, abs/2211.05705.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of ACL*, pages 946–956.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022b. On the advance of making language models better reasoners. *CoRR*, abs/2206.02336.

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

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8600–8607.

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

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

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. SemEval-2015 task 9: CLIPEval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 443–450.

Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. 2022. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238.

Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022a. Causal intervention improves implicit sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6966–6977.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3316–3322.

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

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

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

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed H. Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *CoRR*, abs/2205.10625.

A Appendix

Here we present several pieces of real testing examples. We compare THOR with the vanilla prompting method, and the zero-shot CoT method (Prompt + '*Lets think step by step*'). We perform the comparisons based on the ChatGPT.[4]

[4] https://chat.openai.com/, Dec. 15, 2022

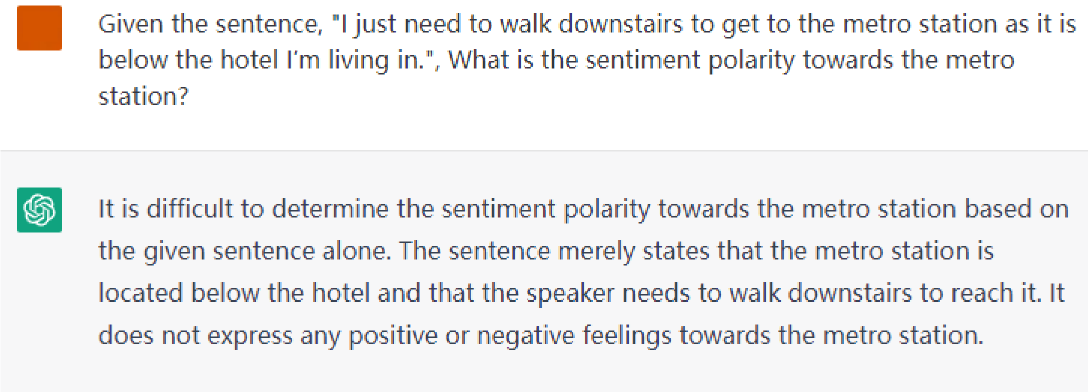

Figure 6: Vanilla prompt-based result for testing case-I.

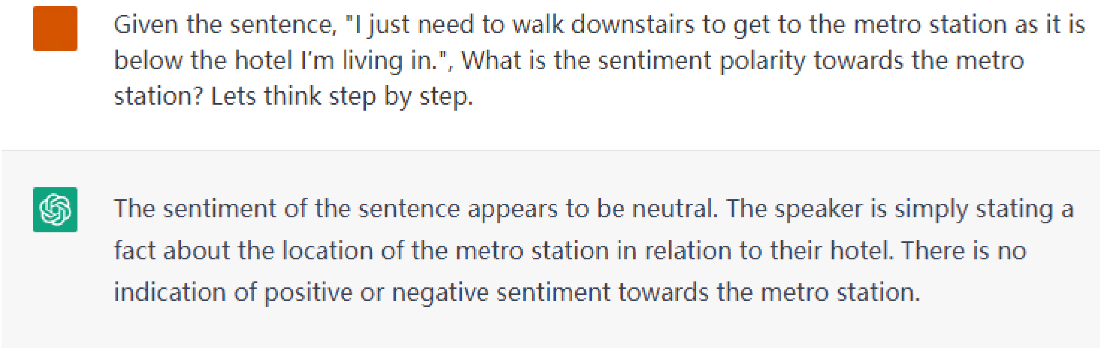

Figure 7: Result by zero-shot CoT method for testing case-I.

- **Case-I** Input text:

  I just need to walk downstairs to get to the metro station as it is below the hotel I'm living in.

The gold sentiment label is *positive* towards *the metro station*.

In Fig. 6, 7 and 8, we show that our THOR successfully induces the ChatGPT to finally give a correct decision on sentiment polarity, where the other two methods fail.

- **Case-II** Input text:

  Lunch came with pickels and slaw, no extra charge.

The gold sentiment label is *positive* towards *Lunch*.

Fig. 9, 10 and 11 shows the results and the LLM's response, respectively. Our THOR induces the ChatGPT to draw a correct decision on sentiment polarity, but the other two methods still fail.

Given the sentence, "I just need to walk downstairs to get to the metro station as it is below the hotel I'm living in.", which specific aspect of the metro station is possibly mentioned?

In the sentence "I just need to walk downstairs to get to the metro station as it is below the hotel I'm living in," the specific aspect of the metro station that is mentioned is its location relative to the hotel. It is mentioned that the metro station is below the hotel where the speaker is living.

Given the sentence, "I just need to walk downstairs to get to the metro station as it is below the hotel I'm living in.", the location of the metro station is mentioned. Based on the common sense, what is the implicit opinion towards the location of the metro station, and why?

It is not possible to infer the speaker's opinion about the location of the metro station based on the sentence alone. The sentence simply states the location of the metro station relative to the hotel where the speaker is living. It does not express any opinion about the location. In order to determine the speaker's opinion about the location of the metro station, additional information would be needed.

Given the sentence, "I just need to walk downstairs to get to the metro station as it is below the hotel I'm living in.", the location of the metro station is mentioned. Based on the common sense, it could be inferred that the location of the metro station being below the hotel is seen as convenient. Based on such opinion, what is the polarity towards the metro station?

Based on the assumption that the location of the metro station being below the hotel is seen as convenient, it could be inferred that the speaker has a positive attitude towards the metro station. However, this is purely speculative and cannot be definitively determined based on the information provided in the sentence.

Figure 8: Result by our THOR method for testing case-I.

Given the sentence, "Lunch came with pickels and slaw, no extra charge.", what is the sentiment polarity towards the lunch?

The sentiment polarity towards the lunch is neutral.

Figure 9: Vanilla prompt-based result for testing case-II.

Given the sentence, "Lunch came with pickels and slaw, no extra charge.", what is the sentiment polarity towards the lunch? Lets think step by step.

It is difficult to determine the sentiment polarity towards the lunch based on the given sentence alone. The sentence simply states that the lunch came with pickles and slaw, and does not express any positive or negative feelings towards the lunch itself. To determine the sentiment towards the lunch, we would need more information or context about the lunch, such as the quality, taste, or appearance of the food.

Figure 10: Result by zero-shot CoT method for testing case-II.

Given the sentence, "Lunch came with pickels and slaw, no extra charge.", which specific aspect of lunch is possibly mentioned?

The specific aspect of lunch that is mentioned is the cost.

Given the sentence, "Lunch came with pickels and slaw, no extra charge.", the specific aspect of lunch that is mentioned is the cost. Based on the common sense, what is the implicit opinion towards the mentioned aspect of lunch, and why?

The implicit opinion towards the cost of the lunch is that it is included in the price and there is no additional charge. This is based on the phrase "no extra charge," which implies that the cost is not a burden or inconvenience to the customer.

Given the sentence, "Lunch came with pickels and slaw, no extra charge.", the specific aspect of lunch that is mentioned is the cost. The implicit opinion towards the cost of the lunch is that it is included in the price and there is no additional charge. This is based on the phrase "no extra charge," which implies that the cost is not a burden or inconvenience to the customer. Based on such opinion, what is the polarity towards lunch?

Based on the opinion that the cost of lunch is included and there is no additional charge, the polarity towards lunch would likely be positive. This means that the speaker has a favorable or favorable opinion of lunch.

Figure 11: Result by our THOR method for testing case-II.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

☑ A2. Did you discuss any potential risks of your work?
5

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix B

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.