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. Wuhan University 3. Sea AI Lab 4. Alibaba DAMO Academy

TL;DR

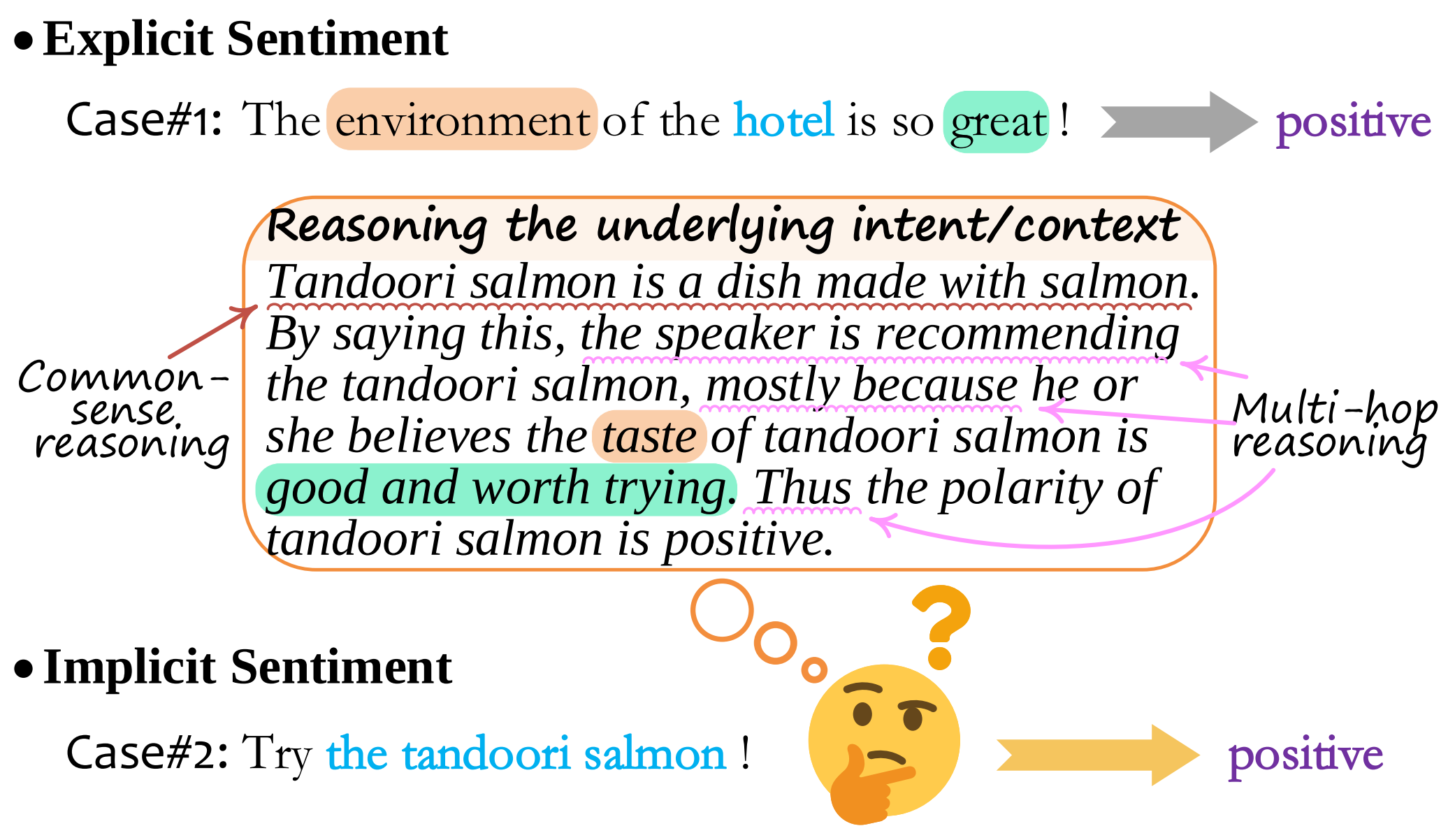
We solve the implicit sentiment analysis with a three-hop reasoning framework based on the chain-of-thought prompting method.

Paper

Code

► 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. Different from ESA, ISA is much more challenging, because in ISA the inputs contain only factual descriptions with no explicit opinion expression directly given. For example, given a text '*Try the tandoori salmon!*', having no salient cue word, almost all existing sentiment classifier predicts a neutral polarity towards '*the tandoori salmon*'.



- **Explicit Sentiment**

Case#1: The environment of the hotel is so great ! positive

Reasoning the underlying intent/context

Tandoori salmon is a dish made with salmon. By saying this, the speaker is recommending the tandoori salmon, mostly because he or she believes the taste of tandoori salmon is good and worth trying. Thus the polarity of tandoori salmon is positive.

- **Implicit Sentiment**

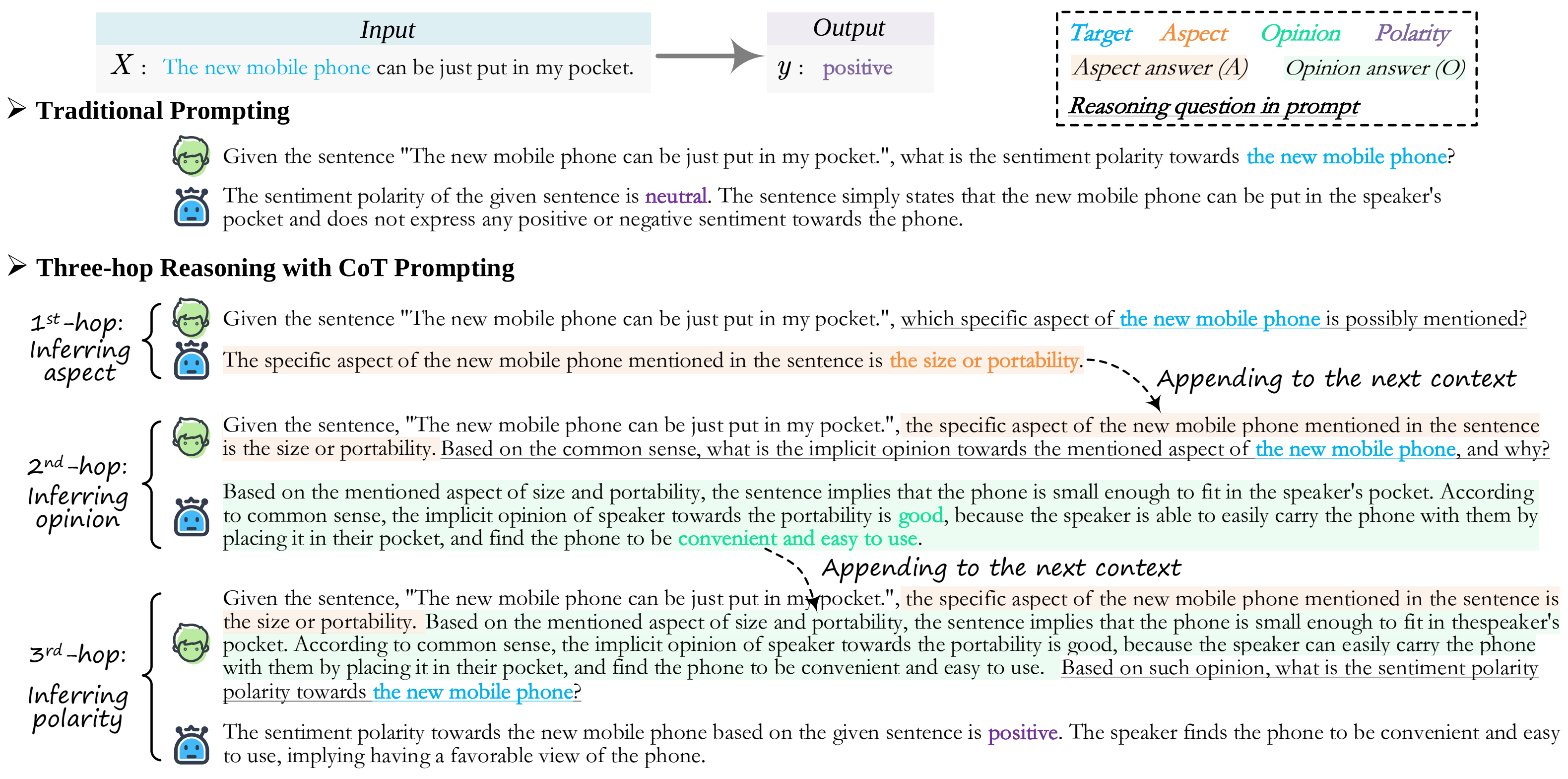
Case#2: Try the tandoori salmon ! positive

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, 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.

► 2. Three-hop Reasoning Framework

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 for common-sense understanding. On the other hand, the latest chain-of-thought (CoT) idea has revealed the great potential of LMs' multi-hop reasoning, 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.



Input

X : The new mobile phone can be just put in my pocket.

Output

y : positive

Target Aspect Opinion Polarity
Aspect answer (A) Opinion answer (O)
Reasoning question in prompt

➢ **Traditional Prompting**

Given the sentence "The new mobile phone can be just put in my pocket.", what is the sentiment polarity towards the new mobile phone?

The sentiment polarity of the given sentence is neutral. The sentence simply states that the new mobile phone can be put in the speaker's pocket and does not express any positive or negative sentiment towards the phone.

➢ **Three-hop Reasoning with CoT Prompting**

1st-hop: Inferring aspect

Given the sentence "The new mobile phone can be just put in my pocket.", which specific aspect of the new mobile phone is possibly mentioned?

The specific aspect of the new mobile phone mentioned in the sentence is the size or portability.

Appending to the next context

2nd-hop: Inferring opinion

Given the sentence, "The new mobile phone can be just put in my pocket.", the specific aspect of the new mobile phone mentioned in the sentence is the size or portability. Based on the common sense, what is the implicit opinion towards the mentioned aspect of the new mobile phone, and why?

Based on the mentioned aspect of size and portability, the sentence implies that the phone is small enough to fit in the speaker's pocket. According to common sense, the implicit opinion of speaker towards the portability is good, because the speaker is able to easily carry the phone with them by placing it in their pocket, and find the phone to be convenient and easy to use.

Appending to the next context

3rd-hop: Inferring polarity

Given the sentence, "The new mobile phone can be just put in my pocket.", the specific aspect of the new mobile phone mentioned in the sentence is the size or portability. Based on the mentioned aspect of size and portability, the sentence implies that the phone is small enough to fit in thespeaker's pocket. According to common sense, the implicit opinion of speaker towards the portability is good, because the speaker can easily carry the phone with them by placing it in their pocket, and find the phone to be convenient and easy to use. Based on such opinion, what is the sentiment polarity polarity towards the new mobile phone?

The sentiment polarity towards the new mobile phone based on the given sentence is positive. The speaker finds the phone to be convenient and easy to use, implying having a favorable view of the phone.

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

Instead of directly asking LLM the final result of y, in our THOR 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:

```
C1[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:

```
C2[C1,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:

```
C3[C2,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$p(y|X,t,a,o)$.

Enhancing Reasoning via Self-consistency

We further leverage the self-consistency mechanism to consolidate the reasoning correctness. Specifically, for each of the 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 the highest confidence as the context in the next step.

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) the initial context, 2) this step's reasoning answer text and 3) the 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.

► 3. Experiments

Main results. 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. GPT3-175B with THOR 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.

	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)	83.98	74.70	81.47	79.52
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).

	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	42.84	52.52	32.40
Flan-T5+THOR (3B)	57.33	50.04	56.36	36.16
Flan-T5+THOR (11B)	<u>61.87</u>	<u>52.76</u>	<u>58.27</u>	<u>40.75</u>
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*'

Further analyses. With the increasing model scale, the efficacy of our multi-hop reasoning prompting is exponentially amplified, cf. Fig. 3. In Fig. 4, both GPT3 and ChatGPT with THOR achieves considerable improvements on ISA. Unsupervised-GPT3 (175B) gives similarity low error rate as with Supervised-T5, while the latter fails much more frequently on the incapability of reasoning. In contrast to Supervised-T5, the majority of failures in Unsupervised-GPT3 comes from problematic data annotation.

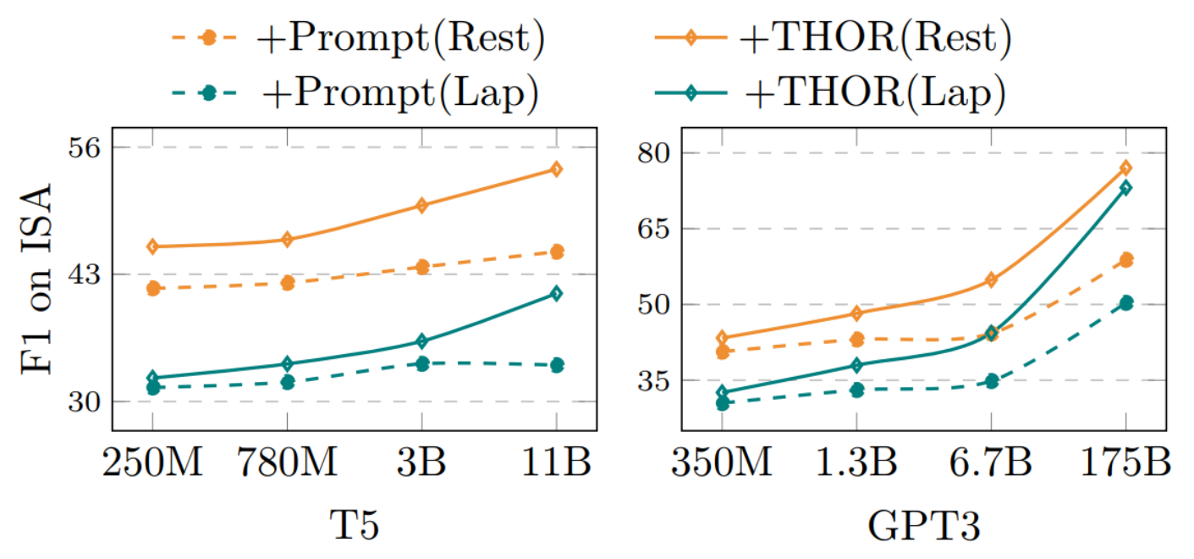


Figure 3: Influences of LLM scales.

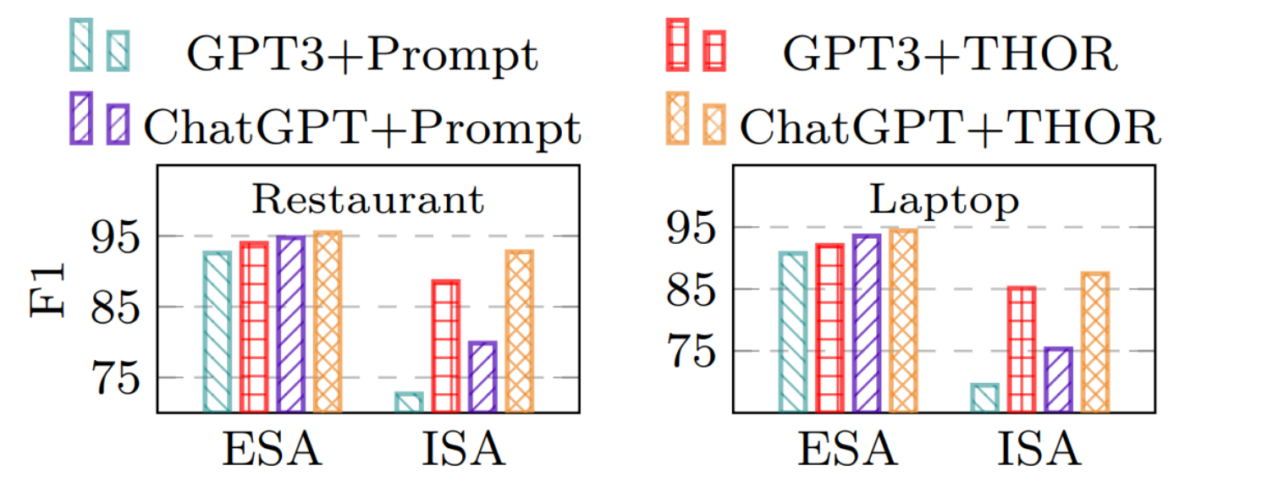


Figure 4: Comparisons between GPT3&ChatGPT on randomly-selected 50 ESA and 50 ISA instances.

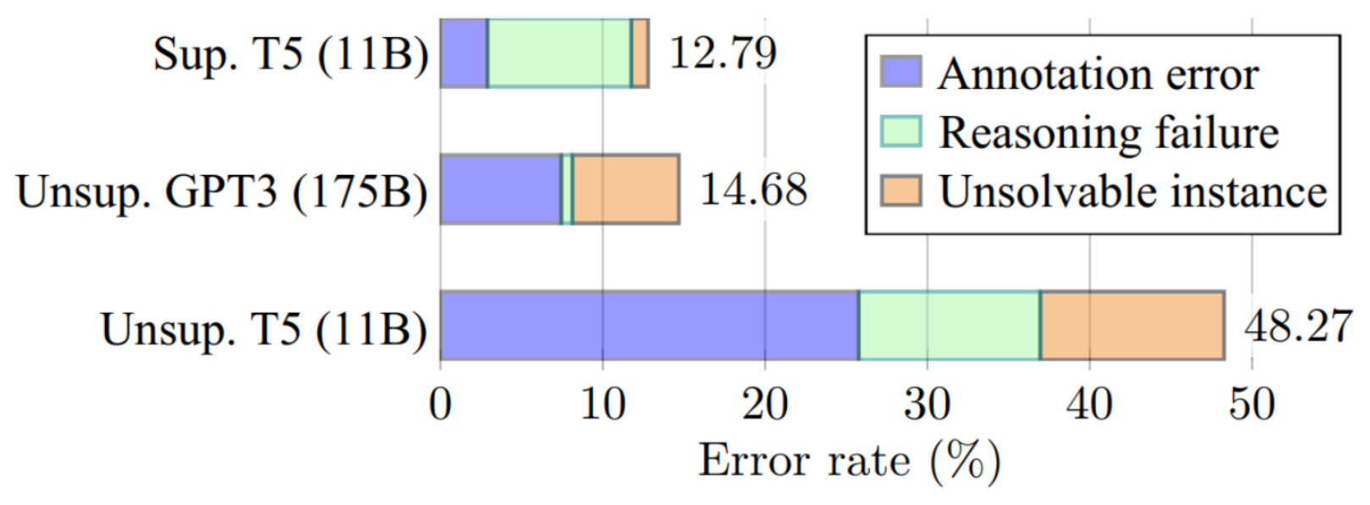


Figure 5: Error analysis.