Can Deception Detection Go Deeper? Dataset, Evaluation, and Benchmark for Deception Reasoning

Anonymous ACL submission

Abstract

Deception detection has garnered increasing attention due to its significance in real-world scenarios, with its main goal being to identify lies in an individual's external behaviors. However, these bases are often subjective and linked to personal habits. To this end, we extend deception detection to *deception reasoning*, further providing objective evidence to support subjective judgment. Specifically, we provide potential lies and basic facts and then analyze inconsistencies in the facts and the underlying intentions to explore why a statement may be a lie. Compared with traditional deception detection, this task is more applicable to real-world scenarios. For example, in interrogation, the police should judge whether a person is lying based on solid evidence. This paper presents our initial attempts at this task, including constructing datasets and defining evaluation metrics. Meanwhile, this task can serve as a benchmark for evaluating the reasoning capability of large language models. Our code and data are provided in the supplementary material.

1 Introduction

Deception is defined as an intentional attempt to mislead others (DePaulo et al., 2003). Detecting deceptive behaviors is challenging even for humans, generally requiring specialized knowledge. Despite its difficulties, deception detection is an important research topic due to its widespread applications, such as airport security screening, court trials, and personal credit risk assessment (Masip, 2017).

Deception detection aims to identify deceptive behavior from an individual's external behavior. Current research mainly focuses on laboratory-controlled or in-the-wild scenarios (Karnati et al., 2021; Speth et al., 2021). The former recruits subjects and triggers their deceptive behaviors in well-designed psychological paradigms (Abouelenien et al., 2016). However, some researchers question the practicality of laboratory-controlled datasets because they are different from real deceptive behaviors (Vrij, 2008; Fitzpatrick et al., 2022). Therefore, in recent years, researchers have paid more attention to real-life datasets (Şen et al., 2020).

However, this judgment is relatively subjective and related to personal habits, and real-world applications require evidence to support these judgments. To address this, we extend the field of deception detection and introduce a new task, *Deception Reasoning*, which aims to infer the underlying reasons why a statement might be a lie. This paper makes an initial attempt at this task by establishing datasets that include both real and synthetic data and defining evaluation metrics (including *accuracy*, *completeness*, *logic*, and *depth*) to assess the reasoning results. The main contributions of this work are summarized as follows:

- **Task.** This paper proposes a new task, deceptive reasoning. Unlike traditional deception detection, we further provide objective evidence to support subjective judgments.
- **Groundwork.** To facilitate research, we construct datasets and define evaluation metrics.
- **Benchmark.** This task can serve as a benchmark for assessing the complex reasoning capabilities of Large Language Models (LLMs).

The rest of this paper is organized as follows: Section 2 reviews recent works. In Section 3, we introduce our data generation pipeline. In Section 4, we define evaluation metrics and report the performance of various LLMs on deception reasoning. Finally, we conclude this paper in Section 5.

2 Related Works

In this section, we first review existing works on deception detection and LLMs. Since we focus on deception reasoning, we further review some works on evaluating reasoning capabilities.

2.1 Deception Detection

Deception detection aims to identify deceptive behavior by analyzing individual clues. In this section, we summarize recent works from two perspectives: datasets and solutions.

Datasets. Current datasets are mainly conducted in laboratory-controlled or in-the-wild scenarios.

In laboratory-controlled setups, researchers often use well-designed psychological paradigms to induce deception. For example, Derrick et al. (2010) asked participants to commit mock crimes. They were rewarded if they could convince the professional interviewer of their innocence. Pérez-Rosas et al. (2014) and Abouelenien et al. (2016) collected data using three scenarios: *mock crime*, *best friend*, and *abortion*. In *mock crime*, participants can choose to take or not take the money in the envelope. They were rewarded if they took the money without raising doubts from interviewers. For *best friend* and *abortion*, participants can discuss these topics using true or fake opinions.

Besides laboratory-controlled scenarios, there are many works focusing on in-the-wild scenarios. For example, Şen et al. (2020) collected videos from public court trials and used trial outcomes to indicate whether the subject was deceptive. Bachenko et al. (2008) analyzed criminal narratives, interrogations, and legal testimony and provided a method to assess whether a statement is truthful or deceptive. Fornaciari and Poesio (2013) attempted to identify deceptive statements in hearings collected in Italian courts. Pérez-Rosas et al. (2015) collected videos from TV shows. The participants were considered to be lying if they gave an opinion about a non-existent movie.

Solutions. Regarding solutions, Karnati et al. (2021) proposed a framework leveraging deep neural networks to improve detection accuracy. Ilias et al. (2022) introduced a Transformer-based framework, proving that it outperforms traditional methods in effectiveness. Yang et al. (2021) analyzed emotion-based features, underscoring the pivotal role of emotional clues in identifying deception. Meanwhile, Hazra and Majumder (2024) investigated deceptive behavior in conversations within high-risk environments, indicating that detectors relying solely on linguistic clues can perform on par with humans in discerning truth.

Deception detection mainly uses individual clues to identify deceptive behavior. However, such judgment is related to personal habits. Unlike deception detection, our *deception reasoning* aims to provide objective evidence for subjective judgment, which has greater value in practical applications. For example, during interrogation, these analytical results can provide guidance to the police officer.

2.2 Large Language Model

Recently, LLMs have shown strong text understanding and generation capabilities, which have been widely used in various tasks. For example, Gan et al. (2023) and Qiu et al. (2023) explored the promise of LLMs in education and mental health support. Wang et al. (2023) used LLMs to learn character-specific language patterns and behaviors to enhance role-playing realism and interactive experiences. Park et al. (2023) exploited LLMs to create multiple characters and let them live in a virtual environment. These characters were able to engage in dialogues and spontaneous social activities. Among existing LLMs, GPT-4 shows strong role-playing ability and can generate more human-like behaviors (Guo et al., 2023; Gui and Toubia, 2023). Thus, a portion of our dataset uses GPT-4 to synthesize dialogues for deception reasoning.

Recent advancements in LLMs have significantly enhanced reasoning capabilities, particularly through the use of long chain-of-thought (CoT) prompting and reinforcement learning (RL). Yeo et al. (2025) systematically investigated the mechanics of long CoT reasoning, identifying key factors such as supervised fine-tuning (SFT) and reward shaping that enable models to generate extended reasoning trajectories. Similarly, Kimi (2025) demonstrated state-of-the-art performance in long CoT reasoning by scaling context windows and optimizing policy methods, achieving remarkable results in multi-modal tasks. DeepSeek-AI (2025) highlighted the importance of RL in developing reasoning skills, showing that models can learn complex strategies through reward-driven training. OpenAI's o1 model (OpenAI, 2024) further exemplifies the potential of long CoT reasoning, achieving high performance on various reasoning benchmarks through reinforcement learning. Due to the complexity of deception reasoning, this task can serve as a novel benchmark for evaluating the reasoning capabilities of LLMs.

2.3 Reasoning Performance Evaluation

Reasoning is a necessary ability to solve sophisticated problems. For example, mathematical rea-

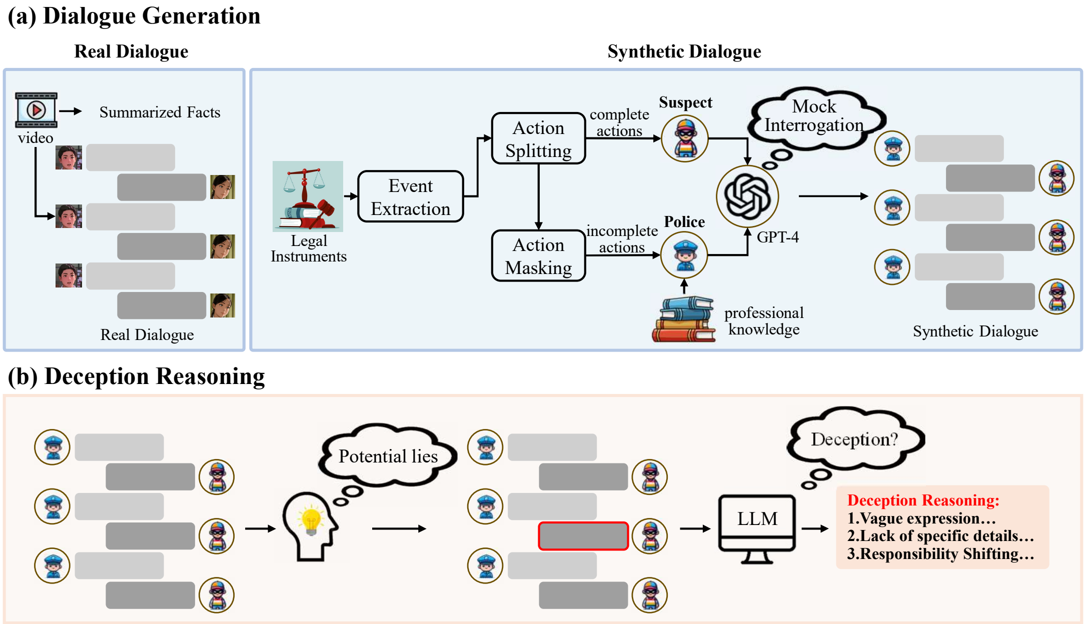


Figure 1: **Overall pipeline. (a) Dialogue Generation:** We construct a dataset that includes both real and synthetic dialogues. For the real data, we manually summarize events and transcribe dialogues. For the synthetic data, we use GPT-4 to simulate interactions between suspects and police. **(b) Deceptive Reasoning:** We manually select a potential lie and exploit the LLM to analyze why this statement might be a lie. We also conduct post-filtering and manual checks to ensure the quality of the deception reasoning.

soning is the ability to reason about math word problems (Mishra et al., 2022a,b). Logical reasoning is a cognitive process of applying general rules or principles to reach specific conclusions (Flach and Hadjiantonis, 2013). In logical reasoning, three elements are usually included: rule, case, and result. These three elements constitute three types of logical reasoning: deductive ($rule + case \Rightarrow result$), inductive ($case + result \Rightarrow rule$), and abductive ($result + Rule \Rightarrow case$). Commonsense reasoning enables computers to understand and apply common knowledge from humans, more effectively simulating human thought processes and decision-making behaviors (Storks et al., 2019).

Existing reasoning datasets mainly use a form of multiple-choice (Geva et al., 2021) or open-ended questions (Weston et al., 2016). For the former, the answer is predefined and the evaluation process is straightforward. For the latter, the model needs to generate the answer, rather than choosing from a given set of options. In our deception reasoning, it is difficult to provide candidate answers and the multiple-choice form may also limit the model's creativity. Therefore, we evaluate this task in the form of open-ended questions.

Previous open-ended questions mainly use the *similarity* between predicted answers and standard answers (Yang et al., 2018). Considering the complexity of deception reasoning, this paper proposes a more comprehensive evaluation strategy covering four dimensions: *accuracy*, *completeness*, *logic*, and *depth*. More details can be found in Section 4.

3 Data Generation

We collect both real data and synthetic data for deception reasoning, and the dataset construction pipeline is summarized in Figure 1.

3.1 Real Dialogue Generation

Our real data is sourced from the documentary "The Guardians of Jiefangxi", which provides some real-world interrogation scenarios. We manually summarized the facts of the cases and recorded the dialogues between suspects and interrogators. These real dialogues offer invaluable insights into the nuances of human interaction during interrogations, including the use of evasive language, vague expressions, and attempts to shift responsibility. However, despite our efforts, the amount of real data we were able to collect is limited. The documentary only covers a specific set of cases, resulting in a lack of diversity in the types of deception behaviors and interrogation contexts. Therefore, we supplement real data with synthetic data to ensure the diversity of our dataset.

3.2 Synthetic Dialogue Generation

We use GPT-4 ("gpt-4-1106-preview") to synthesize dialogues containing deceptive behaviors. Specifically, we choose one of the most widely used scenarios in previous research, *mock crime* (Derrick et al., 2010; Pérez-Rosas et al., 2014). We ask GPT-4 to simulate the role-playing between a suspect and a police officer. During the interrogation, the suspect should deceive the police to evade the crime, while the police should strive to uncover the truth. For clarity, we first define three notations: ***legal instrument***, ***target content***, and ***action***.

Notation Definition. To obtain crime facts, we turn our attention to ***legal instruments***, which include but are not limited to, details of the prosecution's charges, descriptions of the defendant's criminal behavior, arrests, the evidence presented, explicit charges, and stages of the judicial process. To mimic real interrogation, the suspect should know the complete crime facts while the police officer should miss some details. However, *legal instruments* contain contents that can reduce uncertainty during interrogation, such as explicit charges and convictions. Hence, in *legal instruments*, we only select the ***target content***, which denotes a series of behaviors involving multiple people, places, and times. The *target content* contains multiple ***actions***, where an *action* refers to a continuous and specific behavior performed by subjects within a period of time. Appendix C provides examples to illustrate the differences between the *legal instrument*, *target content*, and *action*.

Legal Instrument Selection. CAIL2018 (Xiao et al., 2018) encompasses 2.68 million criminal law documents, spanning 202 types of charges and 183 legal provisions. In this dataset, legal instruments are written by legal experts, with rigorous wording and standardized forms. These high-quality legal instruments bring great convenience to our work.

Proper legal instruments are important for dialogue generation. On the one hand, short legal instruments contain insufficient content, leading to unclear descriptions of details and generating low-quality dialogues. On the other hand, long legal instruments may contain complex crime facts, increasing the difficulty of dialogue generation. Therefore, we select legal instruments with a length ranging from 400 to 2,000. The length distribution after selection is shown in Figure 2, where the length refers to the number of Chinese characters.

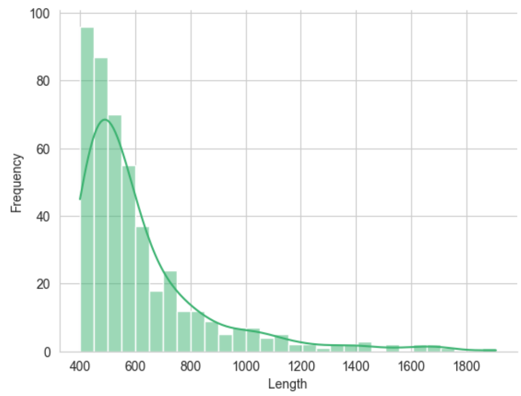


Figure 2: Distribution of lengths after selection (the length refers to the number of Chinese characters).

Sunday afternoon, Chen shot Wu with a gun in the park.
Time / Agent / Patient / Instrument / Location

At unknown time, Chen shot Wu with a gun in the park.
Mask Time / Agent / Patient / Instrument / Location

Figure 3: Example of time masking process.

Target Content and Action Extraction. Then, we attempt to extract the *target content* from *legal instruments* and further disassemble it into multiple *actions*. Specifically, we rely on GPT-4 and adopt a two-stage strategy to achieve this goal. In the first stage, we extract the *target content* from *legal instruments*; in the second stage, we disassemble it into multiple *actions*. To achieve better performance, each stage uses one-shot and chain-of-through prompts (Wei et al., 2022). In Section 4.5, we also analyze the performance of the one-stage strategy, i.e., merging *target content* and *action* extraction into one stage. Experimental results demonstrate that the two-stage strategy is more effective than the one-stage strategy.

Incomplete Action Generation. During the interrogation, the police officer may not have complete crime facts and try to find missing parts from the suspect. To mimic this process, we generate incomplete actions for the police officer.

An action mainly involves the following seven items: (1) *agent* is the entity that performs the action; (2) *patient* is the entity affected by the action; (3) *instrument* is the object used to perform the action; (4) *goal* is the direction or destination of the action; (5) *source* is the place where the action originates; (6) *time* is the time when the action occurs; (7) *location* is the place where the action occurs.

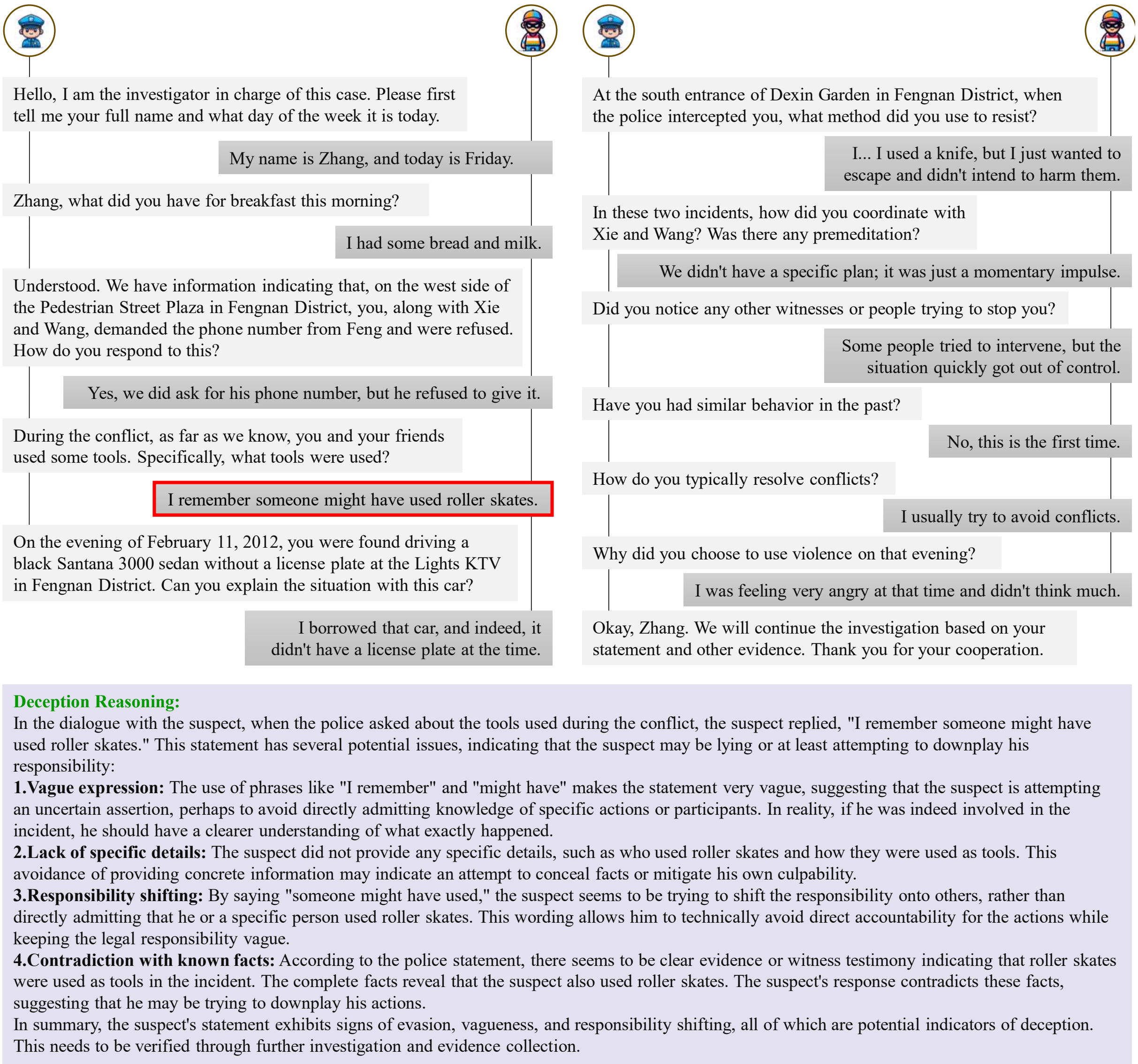


Figure 4: Dialogue, potential lie, and reasoning results. In deception reasoning, we aim to generate multi-faceted analysis results, such as factual inconsistency, ambiguous expressions, intent, etc.

To generate incomplete actions, we randomly mask an item in the action. Specifically, we replace the *agent* and *patient* with unknown people, the *instrument* with unknown tool, the *location* with unknown place, and the specific *time* with unknown time. Figure 3 illustrates the masking process, and Appendix C provides an example of the generated incomplete actions.

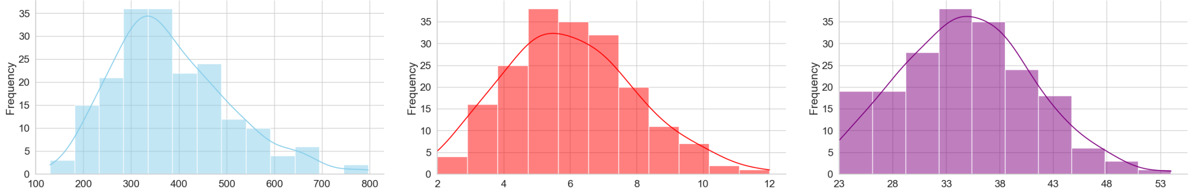
Mock Interrogation. We simulate the interrogation process between the suspect and the police officer. To enhance authenticity, complete and incomplete actions serve as the information held by the suspect and the police officer, respectively. To enhance the professionalism of the police officer, we further provide him with additional interrogation techniques. Specifically, we require the police officer to ask some typical questions (Leo, 1994):

- Control questions: These questions are used to establish a baseline response from the interrogatee. Generally, the interrogatee is honest with these questions. For example, what is your name? What day of the week is it today? Answers to these questions should be truthful so that they can be compared with answers to subsequent questions.

- Relevant questions: They are related to the core of the crime and are often questions to get to the truth. For example, were you involved in an event at a certain time and place? How did you do this? The answers to these

Metric	Synthetic Dialogue			Real Dialogue		
	Max	Min	Avg	Max	Min	Avg
# of turns per dialogue	54	23	34.93	64	6	27.92
# of words per utterance	180	2	19.30	150	1	18.75
# of words per police's utterance	180	7	24.23	126	3	20.24
# of words per suspect's utterance	99	7	20.77	150	1	17.26
police word count divided by suspect word count per turn	9.0	0.17	1.27	63	0.05	2.21

Table 1: Statistics of our deception dataset.



(a) target content length (b) number of actions (c) dialogue turns

Figure 5: Distribution of target content length, number of actions, and dialogue turns.

questions are the focus of the interrogation.

- Comparison questions: These questions are similar to control questions, but they are usually designed to be more challenging to show a distinct physical or psychological response. These questions should be answered in the affirmative. For example, have you ever done anything dishonest? Do you lie often?

- Neutral questions: These questions are often used to relieve tension or provide an opportunity for the interrogatee to relax. They are not related to the subject of the interrogation. For example, what did you have for breakfast this morning? What are your hobbies?

- Randomness and variability: Interrogators usually randomize the order of questions to avoid forming a fixed pattern, thereby reducing the chances that the interrogatee will be able to prepare for or adapt to a particular type of questioning, but neutral and control questions often come first in interrogation.

In this section, we propose two strategies for dialogue generation: (1) we use two GPT-4s playing two roles; (2) we use one GPT-4 to generate a multi-round dialogue between two roles directly. For the first strategy, the output gradually spirals out of control as the dialogue progresses, resulting in a significant drop in quality at the end of the dialogue. Therefore, we turn our attention to the second strategy. We find this strategy can maintain the logic and coherence of the dialogue.

3.3 Deception Reasoning

Figure 1 shows the deception reasoning process. Specifically, we manually select a potential lie and use LLMs to generate multi-faceted analysis results, such as factual inconsistency, ambiguous expressions, intent, etc. Next, we perform manual checks and proofreading to ensure the quality of the reasoning. Figure 4 gives an example to illustrate this process. We also perform post-filtering to remove some dialogues that contain unnatural parts or have no potential lies.

It should be noted that deception reasoning has some similarities with misinformation detection. However, there are also certain differences. As shown in Figure 4, misinformation detection is an aspect of deception reasoning. Differently, deception reasoning is a more comprehensive task that aims to analyze from multiple aspects.

3.4 Data Statistics

Dataset statistics are summarized in Table 1. We observe that the average number of turns per dialogue is around 30, which is sufficient for a short interrogation. In Figure 5, we also provide the distribution of target content length, number of actions, and dialogue turns. Meanwhile, we analyze the cost of data generation. On average, we spend less than $2 per synthetic dialogue, which is less than the cost of the real dialogue.

4 Deception Reasoning Evaluation

In this section, we first define evaluation metrics and evaluators. Then, we assess different LLMs

Model	Automatic Evaluation Results					Manual Evaluation Results				
	Acc.	Com.	Log.	Dep.	Sum	Acc.	Com.	Log.	Dep.	Sum
ChatGLM2-6B	4.00	3.56	4.33	3.44	15.33	5.12	5.00	4.90	4.70	19.72
WizardLM-13B	5.20	4.87	6.00	4.38	20.45	5.61	5.30	5.41	5.14	21.46
Baichuan2-13B	5.24	5.00	6.25	4.62	21.11	4.56	4.36	4.49	4.16	17.57
ERINE3.5	5.40	5.00	6.10	5.10	21.60	5.71	5.81	5.71	5.13	22.36
Qwen-14B	6.00	5.60	6.70	5.20	23.50	5.91	5.80	5.42	5.40	22.53
Claude3-Haiku	6.33	5.89	6.89	5.33	24.44	6.36	6.11	5.94	5.70	24.11
GPT-3.5	6.00	5.87	6.87	5.75	24.49	6.80	6.58	6.53	6.18	26.09
ERINE4.0	6.60	6.30	7.30	5.80	26.00	6.95	6.78	6.99	6.81	27.53
GLM-4-9B	6.67	6.44	7.33	6.33	26.77	7.56	7.54	7.59	7.55	30.24
Gemini-1.5-Pro	6.11	6.89	7.67	6.56	27.23	7.56	7.51	7.37	7.23	29.67
Qwen2-7B	6.56	6.72	7.72	6.39	27.39	7.41	7.49	7.48	7.41	29.79
PCC scores	0.81	0.87	0.80	0.89	0.86	0.81	0.87	0.80	0.89	0.86

Table 2: **Main results on synthetic data.** We report the results of automatic and manual evaluations across different metrics. In the last row, we report the Pearson correlation coefficient (PCC) between the automatic and manual evaluation results for the same metric. Here, PCC can measure the correlation between the two sets of data.

and report evaluation results. After that, we prove the naturalness of synthetic dialogues. Finally, we conduct an ablation study and reveal the rationality of our target content and action extraction strategy. This section mainly uses GPT-3.5 (“gpt-3.5-turbo-0613”) and GPT-4 (“gpt-4-1106-preview”).

4.1 Evaluation Metrics

In deception reasoning, we need to figure out why a sentence may be a lie by considering factual inconsistencies and the intent behind it. To provide a more comprehensive evaluation, we propose four metrics for deception reasoning, whose core definitions are provided below:

- Accuracy: It is used to check whether the reasoning is consistent with the basic facts. If the reasoning is based on the facts, the model should receive a high score in this dimension.

- Completeness: It is used to evaluate whether the model takes into account all details. A good model should be comprehensive and not miss any key information.

- Logic: It is used to evaluate whether the reasoning is logically coherent and well organized. The model is required to have common sense and world knowledge, with deductive, inductive, abductive, and other reasoning abilities. If the reasoning is logically confused or contradictory, the model should receive a low score in this dimension.

- Depth: It is used to evaluate whether a model provides an in-depth analysis or only scratches the surface. This metric is different from completeness. Some reasoning merely restates facts and gives a conclusion, which can be complete but not deep. High-quality reasoning should be able to dig deeper into the reasons and motivations behind it.

These metrics can cover different aspects of reasoning. During the evaluation, we use more detailed definitions for each metric, as well as the meaning of each score for each metric. Please refer to Appendix D for more details.

4.2 Evaluator

We conduct both automatic and manual evaluations. Considering that researchers (Zheng et al., 2023; Lian et al., 2023) have proven the consistency between GPT-4 and human assessments, we directly use GPT-4 for the automatic evaluator. Meanwhile, we hire 8 annotators and perform manual evaluation. Each annotator is paid $14 per hour, which is relatively high in China.

4.3 Main Results

This section evaluates the deception reasoning performance of different LLMs. Besides mainstream LLMs such as WizardLM (Xu et al., 2023), we also select LLMs that perform well in Chinese such as Qwen (Bai et al., 2023). During inference, we input basic facts, synthetic dialogue, and potential lies, and ask LLMs to analyze why this sentence might be a lie. In both automatic and manual evaluation, we use the prompts in Appendix D and experimental results are shown in Table 2 and Table 3.

In Table 2, we observe that existing LLMs can deal with deception reasoning to some extent. Meanwhile, we can see the progress of Chinese LLMs in reasoning ability. For example, Qwen2 is better than Qwen and ERINE4.0 is better than

Model	Automatic Evaluation Results				
	Acc.	Com.	Log.	Dep.	Sum
Qwen-14B	6.8	6.2	7.2	5.4	25.6
GPT-3.5	7.3	6.0	7.1	5.7	26.1
Claude3-Haiku	7.1	6.2	7.2	5.9	26.4
GLM-4-9B	7.0	6.2	7.5	6.4	27.1
ERINE4.0	7.5	6.8	7.4	6.6	28.3
Qwen2-7B	7.6	6.8	7.8	6.8	29.0
Gemini-1.5-Pro	8.2	7.9	8.4	7.6	32.1

Table 3: Main results on real data.

ERINE3.5. Furthermore, Table 2 shows the PCC scores between automatic and manual evaluation results. We observe that manual evaluation results have relatively high similarities with automatic evaluation results, proving the reliability of our automatic evaluation strategy. Therefore, we only report the automatic evaluation results in Table 3. We observe that, although there is an absolute difference in results between real and synthetic data, the ranking outcomes for both real and synthetic data show notable similarities. This suggests that the rankings are primarily influenced by the inherent reasoning ability of the LLMs, and our benchmark effectively reflects the complex reasoning capabilities of different LLMs.

4.4 Dialogue Naturalness Evaluation

To test the naturalness of our synthetic dialogue, we use the prompt in Appendix E and conduct both automatic and manual evaluations.

For automatic evaluation, considering that we use GPT-4 to generate dialogues, we choose another Claude3-Haiku for evaluation. Specifically, we randomly select 10 real dialogues from a dialogue dataset IEMOCAP (Busso et al., 2008) and 10 synthetic dialogues from our dataset. The average score of real dialogue can reach 4.00 and the average score of synthetic dialogue can reach 3.88. For manual evaluation, we hire 8 annotators and ask them to score the naturalness. The average score of synthetic dialogue can reach 3.70, close to the automatic evaluation results. These results reflect the naturalness of our synthetic dialogues and the reliability of our automatic evaluation strategy.

4.5 Ablation Study

This paper uses a two-stage strategy and GPT-4 for target content and action extraction (see Section 3.2). In this section, we compare the performance between one-stage and two-stage strategies, as well as GPT-3.5 and GPT-4. During evaluation, we randomly select 100 samples. For target content

Strategy	Target (↑)	Action (↓)
one-stage + GPT-3.5	47	36
two-stage + GPT-3.5	83	9
one-stage + GPT-4	69	2
two-stage + GPT-4	98	0

Table 4: Performance comparison of different strategies for target content and action extraction.

extraction, we define a metric called *target accuracy*. If the system extracts non-target content from legal instruments, it will have a low score in this metric. For action extraction, we define a metric called *action complexity*. This metric is related to inappropriate action decomposition. Take the complete actions in Appendix C as an example. These actions are well-decomposed. But if we merge two actions into one action, this decomposition process is inappropriate, leading to an increase in *action complexity*. Therefore, a good model should have high *target accuracy* and low *action complexity*.

Experimental results of different strategies are shown in Table 4. We observe that the two-stage strategy achieves better performance than the one-stage strategy. The reason lies in that if we merge target content and action extraction into one stage, it increases the task difficulty, making it more likely that the output does not meet the requirements.

Meanwhile, GPT-4 can achieve better performance than GPT-3.5. Target content and action extraction require the model to understand not only the literal meaning of the text but also its structure and semantic content. Since GPT-4 can achieve better performance than GPT-3.5 in text understanding, it can also achieve better performance in target content and action extraction.

5 Conclusions

This paper extends deception detection to deception reasoning, further providing objective evidence to support subjective judgment. To facilitate subsequent research, we build datasets, define evaluation metrics, and open-source data and code. Meanwhile, we present the performance of mainstream LLMs and reveal the correlations between automatic and manual evaluation results. Furthermore, we prove the rationality of our synthetic dataset construction strategy and the naturalness of our synthetic dialogues. Our proposed deception reasoning task can also serve as an important reasoning benchmark for current LLMs.

Limitations

Several limitations can be addressed in future research. First, this paper evaluates the performance of mainstream LLMs but does not cover all LLMs. In the future, we will expand the evaluation scope. Secondly, inspired by recent research on CoT, in the future, we will also introduce CoT into our deception reasoning to simulate the human analysis process. Thirdly, video generation has become increasingly popular. We will synthesize multimodal data and expand text-based deception reasoning to multimodal deception reasoning.

Societal Impacts

We use legal instruments for dataset construction. On the one hand, legal instruments may provide guidance to criminals. But on the other hand, legal instruments can also remind people not to commit crimes. This paper has similar potential societal impacts as legal instruments. Although our research revolves around deception, our main goal is to detect deception and provide evidence to support the judgment. This tool is of great significance for the police to improve integration efficiency and strengthen social security.

References

Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2016. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security*, 12(5):1042–1055.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. 2008. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 41–48.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

DeepSeek-AI. 2025. Deepseek-r1: Scaling reinforcement learning with llms. https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf.

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74.

Douglas C Derrick, Aaron C Elkins, Judee K Burgoon, Jay F Nunamaker, and Daniel Dajun Zeng. 2010. Border security credibility assessments via heterogeneous sensor fusion. *IEEE Intelligent Systems*, 25(03):41–49.

Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari. 2022. *Automatic detection of verbal deception*. Springer Nature.

Peter A Flach and Antonis Hadjiantonis. 2013. *Abduction and Induction: Essays on their relation and integration*, volume 18. Springer Science & Business Media.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21:303–340.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4776–4785. IEEE.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *Available at SSRN 4650172*.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8498–8512.

Loukas Ilias, Felix Soldner, and Bennett Kleinberg. 2022. Explainable verbal deception detection using transformers. *arXiv preprint arXiv:2210.03080*.

Mohan Karnati, Ayan Seal, Anis Yazidi, and Ondrej Krejcar. 2021. Lienet: A deep convolution neural network framework for detecting deception. *IEEE transactions on cognitive and developmental systems*, 14(3):971–984.

Kimi. 2025. Kimi k1.5: Scaling reinforcement learning with llms. https://github.com/MoonshotAI/Kimi-k1.5.

Richard A Leo. 1994. Police interrogation and social control. *Social & Legal Studies*, 3(1):93–120.

Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401*.

Jaume Masip. 2017. Deception detection: State of the art and future prospects. *Psicothema*, 29(2):149–159.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. 2022a. Lila: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022b. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523.

OpenAI. 2024. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. 2015. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2336–2346.

Verónica Pérez-Rosas, Rada Mihalcea, Alexis Narvaez, and Mihai Burzo. 2014. A multimodal dataset for deception detection. In *LREC*, pages 3118–3122.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

M Umut Şen, Veronica Perez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2020. Multimodal deception detection using real-life trial data. *IEEE Transactions on Affective Computing*, 13(1):306–319.

Jeremy Speth, Nathan Vance, Adam Czajka, Kevin W Bowyer, Diane Wright, and Patrick Flynn. 2021. Deception detection and remote physiological monitoring: A dataset and baseline experimental results. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, pages 1–60.

Aldert Vrij. 2008. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Jun-Teng Yang, Guei-Ming Liu, and Scott CH Huang. 2021. Multimodal deception detection in videos via analyzing emotional state-based feature. *arXiv preprint arXiv*, 2104:7.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Language of the Dataset

In this paper, the dataset we constructed is in Chinese. For more details, please refer to our supplementary materials.

B Inference Cost Analysis

Table 5 shows the inference cost per sample for each LLM. For closed-source models provided by OpenAI, Google, etc., we calculate the inference cost based on the number of tokens and the price per token. For open-source models such as GLM-4-9B and Qwen2-7B, we calculate the inference cost based on the model inference time and the daily price of the machine usage. Specifically, we use Azure Standard_NC12s_v3 (equipped with 2 V100 GPUs) based on the pay-as-you-go pricing in December 2023. Although these costs are not accurate due to price changes, they provide a rough estimate of the inference cost. We find that for open-source LLMs, large models are often expensive due to their long inference time. For close-source LLMs, Gemini-1.5-Pro is cheaper than GPT-3.5.

Model	Cost ($\times 10^{-3}$\$)
ChatGLM2-6B	1.3
WizardLM-13B	3.6
Baichuan2-13B	2.1
ERINE3.5	0.1
Qwen-14B	2.2
Claude3-Haiku	0.9
GPT-3.5	4.2
ERINE4.0	3.6
GLM-4-9B	2.8
Gemini-1.5-Pro	0.7
Qwen2-7B	1.8

Table 5: Inference cost per sample for different LLMs.

C Example Visualization

Table 6 provides an example to illustrate the differences between the legal instrument, the target content, and the complete and incomplete actions.

D Metric Calculation

In deception reasoning, we define four evaluation metrics: *accuracy*, *completeness*, *logic*, and *depth*. In Tables 7∼10, we provide detailed definitions for each metric, as well as the meaning of each score.

E Dialogue Naturalness Evaluation

We rank the dialogue's naturalness using five scores. Please refer to Table 11 for more details.

Legal Instrument
The Tangshan Fengnan District People's Procuratorate accuses: On July 16, 2011, at around 21:00, on the west side of the Pedestrian Street Plaza in Fengnan District, the defendant Zhang, along with Xie Mou (already sentenced), Wang Mou (separate case), and others, demanded the phone number from Feng Mou. After being rejected, they continued to verbally harass. Later, the defendant Zhang and Wang Mou used roller skates, while Xie Mou and others used fists and feet to assault Ma Mou, Tao Mou, Xue Mou, and others who tried to intervene. This resulted in Ma Mou sustaining light injuries, Xue Mou minor injuries, and Tao Mou minor injuries. On the evening of February 11, 2012, at around 19:00, the defendant Zhang, driving a black Santana 3000 sedan (without a license plate), was found at the Lights KTV in Fengnan District, suspected of being involved in the January 31, 2012 case at the Fengnan District Billiard Hall. The incident was immediately reported to the Fengnan District Public Security Bureau, notifying police officer Xue Mou. At the south entrance of Dexin Garden in Fengnan District, when police officer Xue Mou and two colleagues intercepted the defendant Zhang in a car, the defendant Zhang stabbed Xue Mou with a knife and fled, causing minor injuries to Xue Mou. In response to the alleged facts, the public prosecution submitted corresponding evidence. The public prosecution authorities believe that the actions of Defendant Zhang constitute the crimes of xxx and xxx and request sentencing according to the provisions of the Criminal Law of the People's Republic of China xxx and xxx.
Target Content
1. On July 16, 2011, around 21:00, on the west side of the Pedestrian Street Plaza in Fengnan District, the defendant Zhang, along with Xie Mou (already sentenced), Wang Mou (separate case), and others, demanded the phone number from Feng Mou. After being rejected, they continued to verbally harass. Later, the defendant Zhang and Wang Mou used roller skates, while Xie Mou and others used fists and feet to assault Ma Mou, Tao Mou, Xue Mou, and others who tried to intervene. This resulted in Ma Mou sustaining light injuries, Xue Mou minor injuries, and Tao Mou minor injuries.
2. On the evening of February 11, 2012, at around 19:00, the defendant Zhang, driving a black Santana 3000 sedan (without a license plate), was found at the Lights KTV in Fengnan District, suspected of being involved in the January 31, 2012 case at the Fengnan District Billiard Hall. The incident was immediately reported. At the south entrance of Dexin Garden in Fengnan District, the defendant Zhang used a knife to injure Xue Mou and fled, causing minor injuries to Xue Mou.
Complete Actions
1. On July 16, 2011, around 21:00, on the west side of the Pedestrian Street Plaza in Fengnan District, the defendant Zhang, along with Xie Mou and Wang Mou, demanded the phone number from Feng Mou but was refused.
2. On July 16, 2011, the defendant Zhang and Wang Mou used roller skates, while Xie Mou and others used fists and feet to assault Ma Mou, Tao Mou, Xue Mou. This resulted in Ma Mou sustaining light injuries, Xue Mou minor injuries, and Tao Mou minor injuries.
3. On the evening of February 11, 2012, at around 19:00, the defendant Zhang, driving a black Santana 3000 sedan (without a license plate), was found at the Lights KTV in Fengnan District. Someone suspected that he was involved in a previous case and immediately reported it to the Fengnan District Public Security Bureau, notifying police officer Xue Mou.
4. On February 11, 2012, at the south entrance of Dexin Garden in Fengnan District, the defendant Zhang used a knife to injure Xue Mou and fled. This attack caused minor injuries to Xue Mou.
Incomplete Actions
1. At an unknown time, on the west side of the Pedestrian Street Plaza in Fengnan District, the defendant Zhang, along with Xie and Wang, demanded Feng's phone number, but was refused.
2. On July 16, 2011, the defendant Zhang and Wang, using unknown tools, along with Xie and others using fists and feet, assaulted Ma, Tao, Xue. This assault resulted in Ma suffering minor injuries, Xue having minor injuries, and Tao having minor injuries.
3. On February 11, 2012, around 7:00 PM, the defendant Zhang drove a black Santana 3000 sedan (without a license plate), and at an unknown location, was found by someone who immediately reported it to Fengnan District Public Security Bureau police officer Xue, suspecting involvement in a previous case.
4. On February 11, 2012, at the south entrance of Dexin Garden in Fengnan District, the defendant Zhang used unknown tools to injure Xue and then fled. This attack caused Xue to suffer minor injuries.

Table 6: Examples of the legal instrument, target content, and complete and incomplete actions.

We provide facts, a dialogue, and a potential lie. Meanwhile, we provide a model's analysis results of whether this sentence might be a lie. Please evaluate whether the model's inference aligns with known facts. If an inference is closer to the real situation or facts, it should receive a higher score in this dimension. **0-2 points (Very low accuracy):** Most of the arguments do not align with known facts, with only a small part possibly slightly related. Displays a severe misunderstanding of the facts or selective ignorance. **3-4 points (Low accuracy):** Some of the arguments align with the facts, but most of the content is still inaccurate or misleading. There is an attempt to use correct facts, but they are handled improperly or misunderstood. **5-6 points (Moderate accuracy):** There is some degree of consistency between the arguments and the facts, but there are still noticeable inaccuracies. Displays a basic understanding of the facts, but lacks thorough or detailed consideration. **7-8 points (High accuracy):** Most of the arguments align with the facts, with only a few details or aspects showing deviations. Shows a good understanding of the facts and correct application, but there is still room for improvement. **9-10 points (Very accurate):** All or almost all of the arguments strictly align with known facts. Accurately and fully understands and applies the facts, with no obvious errors or omissions.

Table 7: Prompt for evaluating the accuracy in deception reasoning.

We provide facts, a dialogue, and a potential lie. Meanwhile, we provide a model's analysis results of whether this sentence might be a lie. Please evaluate whether the model has considered all relevant information and details. A good inference should be comprehensive, without omitting any key points. I will provide a "possible answer" which can be considered a reliable answer scoring 8 or above, as a reference for completeness scoring. **0-2 points (Very low completeness):** The inference is extremely one-sided or missing significant content, with almost no resemblance to the "possible answer." Ignores key aspects of the problem, possibly only scratching the surface. Lacks basic understanding of the problem, with results far from the "possible answer." **3-5 points (Moderate completeness):** The inference includes some key aspects of the problem but still falls far short of the "possible answer." There is an attempt to address the problem comprehensively, but some important aspects or details are overlooked. The inference has some resemblance to the "possible answer," but there are still clear omissions or misunderstandings. **6-7 points (High completeness):** The inference is fairly comprehensive, covering most key aspects and closely aligning with the "possible answer." Able to understand and respond well to the core requirements of the problem, though there may still be some omissions in details. The inference has a high degree of similarity to the "possible answer," but there is still room for improvement. **8 points (Very high completeness):** The inference is very comprehensive, covering all key aspects of the problem and highly consistent with the "possible answer." Demonstrates a deep understanding of the problem, responding accurately and thoroughly to all aspects. The result shows depth and detail, with almost no omissions or misunderstandings. **9-10 points (Exceeds completeness):** Not only highly consistent with the "possible answer," but also offers innovation or further depth. Provides additional insights.

Table 8: Prompt for evaluating the completeness in deception reasoning.

We provide facts, a dialogue, and a potential lie. Meanwhile, we provide a model's analysis results of whether this sentence might be a lie. Please evaluate whether the inference is logically coherent and well-structured. If the model provides an inference that is disorganized or self-contradictory, it should score lower in this dimension. **0-2 points (Low level):** The inference has almost no logical coherence, possibly entirely based on guesswork or conjecture. There is no clear connection between evidence and conclusions, or no evidence is used at all. The reasoning process is chaotic and lacks organization, possibly deviating completely from the core of the problem. **3-5 points (Moderate level):** The inference has some logical coherence, but there may be noticeable logical gaps or errors. Some relevant evidence is used, but the application of evidence is either inappropriate or insufficient. The reasoning process, while somewhat structured, may lack in-depth analysis in key areas. **6-8 points (Good level):** The inference has strong logical coherence, with few logical gaps. Evidence is used appropriately and sufficiently to support the conclusion. The reasoning process is clear and well-organized, with in-depth analysis of key parts of the problem. **9-10 points (Excellent level):** The inference is highly logical, with almost no logical gaps. Evidence is used extremely appropriately and sufficiently, strongly supporting the conclusion. The reasoning process is very clear and well-organized, with in-depth exploration of various aspects of the problem, possibly providing new insights or solutions.

Table 9: Prompt for evaluating the logic in deception reasoning.

We provide facts, a dialogue, and a potential lie. Meanwhile, we provide a model's analysis results of whether this sentence might be a lie. Please evaluate whether there has been deep thought and analysis or if it remains superficial. A high-quality inference should be able to deeply explore underlying reasons and motivations. **0-2 points (Superficial thinking):** The inference is superficial, staying only at the surface level. Lacks exploration of underlying reasons and motivations, ignoring the complexity and deeper factors of the problem. The conclusion may be overly simple and direct, not showing multi-angle or in-depth consideration of the problem. **3-5 points (Basic depth):** The inference shows some depth, but still not comprehensive or deep enough. While there is an attempt to explore underlying reasons and motivations, the analysis still appears shallow or incomplete. The reasoning process has some logical coherence, but lacks depth and complexity. **6-8 points (Good depth):** The inference shows good depth, exploring various aspects of the problem fairly comprehensively. There is in-depth analysis of underlying reasons and motivations, providing deep insights. Although the depth is high, there may still be room for further exploration in some areas. **9-10 points (Extremely deep):** The inference is extremely deep and comprehensive, uncovering the core and deeper factors of the problem. The analysis of underlying reasons and motivations is profound, offering unique and deep insights. The reasoning process displays a high level of logical coherence and depth, exploring the problem from multiple angles, showing high-quality thinking.

Table 10: Prompt for evaluating the depth in deception reasoning.

Now you need to rate a conversation. Please ignore its format and focus on the content. The more the conversation resembles a real dialogue, the higher the score. The maximum score is five points. The rating criteria are as follows: **1 point:** very unnatural. The conversation appears very stiff and unnatural, possibly containing numerous grammar errors, incoherent sentences, or content that is completely unrelated to the context. This type of conversation is difficult to understand and gives off a mechanical or robotic feel, lacking the natural fluency of human communication. **2 points:** somewhat unnatural. Although the conversation conveys basic information, it still seems somewhat unnatural. There may be some linguistic or logical inconsistencies that make the conversation lack the smoothness of natural communication. The conversation may occasionally contain content that is unrelated to the context, requiring further improvement to enhance its naturalness. **3 points:** moderately natural. The conversation is somewhat fluent but still has some issues. There may be some lack of coherence in some places, or occasional unnatural expressions. The conversation can generally stay on topic but still has room for improvement to better simulate natural language communication. **4 points:** fairly natural. The conversation is generally fluent and can convey meaning and emotions well. Although there may be occasional minor unnatural aspects, overall, it closely resembles real human dialogue. The conversation is coherent, able to closely follow the topic, and demonstrates good adaptability and understanding. **5 points:** very natural. The conversation is extremely fluent and natural as if it were a real interaction with a person. There are no language or logical inconsistencies throughout the conversation, maintaining consistency and relevance to the topic. The expression is precise, and adaptable, closely simulating human communication habits and emotional responses, giving a very authentic and comfortable feeling.

Table 11: Prompt for evaluating the dialogue naturalness.