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

Anonymous ACL submission

## Abstract

Deception detection has attracted increasing attention due to its importance in many practical scenarios. Currently, data scarcity harms the development of this field. On the one hand, it is costly to hire participants to simulate deception scenarios. On the other hand, it is difficult to collect videos containing deceptive behaviors on the Internet. To address data scarcity, this paper proposes a new data collection pipeline. Specifically, we use GPT-4 to simulate a role-play between a suspect and a police officer. During interrogation, the suspect lies to the police officer to evade responsibility for the crime, while the police officer uncovers the truth and gathers evidence. Compared with previous datasets, this strategy reduces data collection costs, providing a promising way to increase the dataset size. Meanwhile, we extend the traditional deception detection task to deception reasoning, further providing evidence for deceptive parts. This dataset can also be used to evaluate the complex reasoning capability of current large language models and serve as a reasoning benchmark for further research.

## 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).

Researchers have proposed various algorithms and made great progress (Karnati et al., 2021; Speth et al., 2021). However, deception detection remains understudied due to data scarcity (Soldner et al., 2019). Current datasets can be roughly divided into two categories: laboratory-controlled and in-the-wild datasets. The former hires participants and triggers their deceptive behaviors in well-designed psychological paradigms (Abouelenien et al., 2016). These datasets require plenty of time and money costs, making them difficult to extend to a large scale. The latter collects real-life data and manually annotates deceptive behaviors (Şen et al., 2020). However, it is difficult to collect data containing deceptive behaviors. Meanwhile, the identity information is sensitive, especially when it comes to deception, making it difficult to open source these datasets for subsequent research.

Recently, GPT-4 (OpenAI, 2022) has demonstrated powerful text understanding and generation capabilities. Since deception datasets often employ participants and evoke their deceptive behaviors, can we directly use GPT-4 to simulate this process? To answer this question, we choose one of the most widely used scenarios, *mock crime* (Derrick et al., 2010; Pérez-Rosas et al., 2014). Specifically, we use GPT-4 to simulate the role-play between a suspect and a police officer. We provide complete crime facts to the suspect and incomplete crime facts to the police officer. During interrogation, the suspect needs to deceive the police officer and escape the crime. In contrast, the police officer needs to find out the truth and seize evidence. Compared with previous datasets, this strategy can reduce data collection costs and does not include real people, eliminating potential issues of human identity.

Besides the dataset, we further propose multi-faced metrics to evaluate deception recognition abilities. Traditionally, deception detection mainly uses clues in an utterance (such as blinking, frowning, and stuttering) to identify deceptive behaviors. However, deceptive behaviors may not be conveyed through external clues, and a more accurate judgment should be based on basic facts. Hence, this paper extends deception detection to deception reasoning. Specifically, we pick a potential lie and analyze why this sentence may lie considering factual inconsistencies and intent behind it. To provide a more comprehensive evaluation of reasoning

results, we evaluate them along four dimensions: accuracy, completeness, logic, and depth.

Meanwhile, previous works point out the data leakage problem during LLM evaluation, where the data related or identical to the test data may be used unconsciously when training LLMs, which affects the reliability of evaluation results (Zhou et al., 2023). Differently, our data generation strategy can generate a variety of unseen dialogues, which points out a promising direction for dealing with this problem. The main contributions of this paper are summarized as follows:

- We conduct an initial attempt to use GPT-4 to generate dialogues containing deceptive behaviors. Compared to existing datasets, this strategy reduces collection costs and provides a promising way to increase the dataset size.
- We propose a new task, deception reasoning. Different from traditional deception detection, in this task, we further consider basic facts and provide evidence for potential lies.
- We define multi-faced metrics for deception reasoning and assess the performance of various LLMs. During evaluation, we can alleviate the data leakage problem to some extent and provide more reliable evaluation results.

The rest of this paper is organized as follows: In Section 2, we review some recent works. In Section 3, we provide a detailed description of our data generation pipeline. In Section 4, we illustrate our evaluation metrics and report the reasoning performance of various LLMs. In Section 5 and Section 6, we conclude this paper and discuss limitations.

## 2 Related Works

This paper explores the usage of LLMs to construct deception datasets. Hence, we first review the existing deception datasets and LLMs. Since we focus on deception reasoning, we further review some works on evaluating reasoning capabilities.

### 2.1 Deception Corpus

Datasets are the basis for training and evaluating different algorithms. Existing datasets can be roughly divided into two categories: laboratory-controlled datasets and in-the-wild datasets.

For laboratory-controlled datasets, researchers often use well-designed psychological paradigms to induce deceptive behaviors in participants. For example, Derrick et al. (2010) asked participants to commit mock crimes. They received a reward if they can 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 the *mock crime* scenario, the participant can choose to take or not take the money in the envelope. To receive a reward, they should take the money but not raise doubts from interviewers. For *best friend* and *abortion*, participants can discuss these two topics using their true or fake opinions. However, the above data collection strategy is costly.

For in-the-wild datasets, researchers collect videos containing deceptive behaviors from real-life scenarios. Şen et al. (2020) collected videos from public court trials and used trial outcomes to indicate whether the subjects were deceptive. Besides court trials, Pérez-Rosas et al. (2015) collected TV shows and interviews. For example, the participants were considered to be lying if they expressed an opinion about a non-existent movie. However, not all videos contain scenes involving deception, making these datasets difficult to collect.

Different from previous works, this paper utilizes LLMs to construct the deception dataset. Due to the low cost, this strategy provides a promising way to expand the dataset size. Meanwhile, our dataset does not contain real people, so there is no potential identity sensitivity issue.

### 2.2 Large Language Model

LLMs have shown strong text understanding and generation capabilities. Due to their unprecedented performance, current research interests have shifted from traditional tasks such as fill-in-the-blanks and translation to more challenging 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) leveraged 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, showcasing LLMs' strong role-playing abilities.

Among all LLMs, GPT-4 achieves remarkable performance in various tasks and domains (Guo et al., 2023). Therefore, this paper uses GPT-4 to role-play and generate the deception dataset.

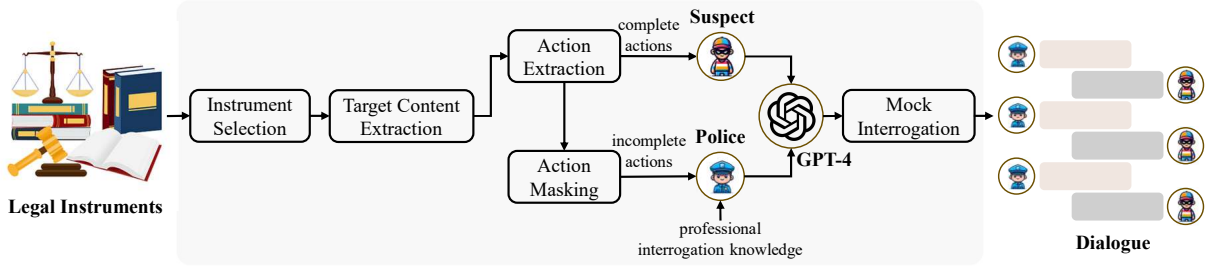



Figure 1: Pipeline of dialogue generation based on legal instruments.

### 2.3 Reasoning Performance Evaluation

Reasoning is a necessary ability to solve real-life sophisticated problems. For example, mathematical reasoning 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. There are three main 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 answers are 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. For the deception reasoning task, it is difficult to provide candidate answers, and the multiple-choice form also limits the model's creativity. Therefore, we evaluate this task in the form of open-ended questions.

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

## 3 Data Generation

In this section, we first clarify the definitions of our deception reasoning task and some important notations: *legal instrument*, *target content*, and *action*. Then, we introduce the data generation process. The overall pipeline is shown in Figure 1.

### 3.1 Task Definition

In deception reasoning, we first select a potential lie from the dialogue and then analyze why this sentence might be a lie considering factual inconsistencies and the intent behind it. Therefore, we should generate dialogues, select potential lies, and provide deception reasoning results.

### 3.2 Notation Definition

In this paper, we ask GPT-4 to conduct a mock interrogation around the crime facts between a suspect and a police officer. 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. Table 1 provides examples of the *legal instrument*, *target content*, and *action*.

### 3.3 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

| 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.<br>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. |
| Action |
| 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.<br>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.<br>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.<br>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. |

Table 1: Examples of the legal instrument, target content, and action.

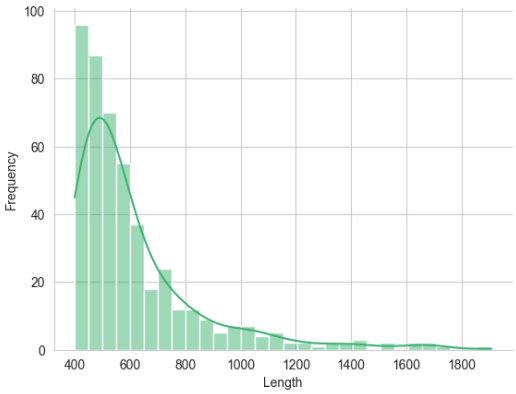



Figure 2: Distribution of lengths after selection.

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

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

Figure 3: Example of the masking time process.

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 only select legal instruments with lengths ranging from 400 to 2,000. The distribution of lengths after selection is shown in Figure 2.

### 3.4 Target Content and Action Extraction

In this section, we aim to extract the *target content* from *legal instruments* and further disassemble it into multiple *actions*. To achieve this goal, we rely on GPT-4, a powerful LLM with excellent performance in understanding semantics and extracting key information. Specifically, we adopt a two-stage strategy. In the first stage, we extract the *target content* from *legal instruments*; in the second stage, we disassemble it into multiple *actions*. For better performance, each stage uses one-shot and chain-of-through prompting (Wei et al., 2022). In Section 4.5, we further analyze the performance of the one-stage strategy, i.e., merging *target content* and *action* extraction into one stage. We observe that the two-stage strategy is more effective than the one-stage strategy. Meanwhile, GPT-4 performs better than GPT-3.5. Therefore, we choose the two-stage strategy and use GPT-4 in this section.

### 3.5 Incomplete Action Generation

In real interrogations, 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 ac-

| Incomplete Action |
|---|
| 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 2: Generated incomplete actions for actions in Table 1.

tion; (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. To generate incomplete actions, we randomly mask an item in the action. Specifically, we replace the *agent* and *patient* with someone, the *instrument* with something, the *location* with someplace, and the specific *time* with some time. We provide an example in Figure 3. In Table 2, we provide generated incomplete actions for actions in Table 1. This masking process is realized by GPT-4.

### 3.6 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 during interrogation (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 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 an 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 directly generate a multi-round dialog between two roles. For the first strategy, as the dialogue progresses, it gradually spirals out of control, 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 dialog. Figure 4 provides the generated dialogue for examples in Tables 1∼2.

In deception reasoning, we pick a potential lie and analyze why this sentence might be a lie considering factual inconsistencies and the intent behind it. Figure 4 provides an example to illustrate this process. Specifically, we choose a potential lie (in the red box) and generate reasoning results.

### 3.7 Dataset Statistics

In this paper, we provide an initial attempt that uses GPT-4 to construct the deception dataset. Specifically, we generate 191 dialogues, and statistics are summarized in Table 3. We observe that our average turn per dialogue is 34.05, which is enough


Figure 4: Generated dialogue, potential lie (in the red box), and reasoning results. To mimic real interrogations, we use complete actions for the suspect (in Table 1) and incomplete actions for the police officer (in Table 2).

for a short interrogation process. Meanwhile, we present the distributions of target content length, number of actions, and dialogue turns in Figure 5.

## 4 Deception Reasoning Evaluation

In this section, we first define the evaluation metrics and evaluator for deception reasoning. Then, we discuss how we alleviate the data leakage problem during LLMs evaluation. Following that, we assess different LLMs on deception reasoning and present the evaluation results. Finally, we conduct the ablation study and reveal the rationality of our target content and action extraction strategy.

### 4.1 Evaluation Metrics

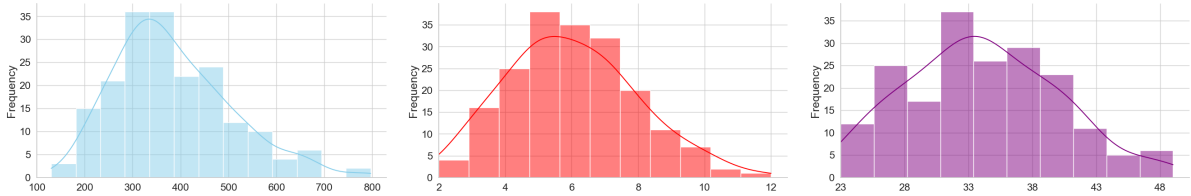
In deception reasoning, we need to figure out why the sentence might be a lie considering factual inconsistencies and the intent behind it. To provide a more comprehensive evaluation, we propose four evaluation metrics for deception reasoning:

- 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.

| Metric | Value |
|---|---|
| # of dialogues | 191 |
| max/min/avg # of turns per dialogue | 49/23/34.05 |
| max/min/avg # of words per utterance | 177/2/19.3 |
| max/min/avg # of words per police's utterance | 177/4/21.71 |
| max/min/avg # of words per suspect's utterance | 101/2/16.89 |
| max/min/avg police word count divided by suspect word count per turn | 18.67/0.14/1.54 |

Table 3: Statistics of our generated 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.

- 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.

### 4.2 Evaluator

Considering that previous works (Zheng et al., 2023; Lian et al., 2023) have demonstrated the consistency between GPT-4 and human assessments, this paper uses GPT-4 as the evaluator. To test its stability, we run GPT-4 multiple times and report the average result along with the standard deviation. Experimental results are shown in Table 4. We observe that the standard deviation is not significant for different runs. These results prove the reliability and stability of the GPT-4 evaluator.

### 4.3 Data Leakage

The data leakage problem occurs in evaluations of LLMs (Zhou et al., 2023), i.e., LLMs are trained and tested using related or identical data. Such a phenomenon may be unconscious, as the content of the future evaluation dataset is not known at the time of preparing the pre-training corpus. However, this problem can invalidate the benchmark and lead to evaluation anomalies (Schaeffer, 2023).

Our deception dataset contains various unseen dialogues and can deal with this problem to some extent. Specifically, we randomly mask one of the action items during data generation (see Section 3.5), which will change the question of the police officer and further affect the suspect's response. For example, in Table 2 (incomplete actions), the instrument is masked. Therefore, in Figure 4 (dialogue), the police officer asks questions about the instrument and the suspect's response revolves around the instrument. This randomness points to a promising way to deal with the problem of data leakage.

### 4.4 Main Results

In this section, we evaluate the deception reasoning performance of different LLMs. Specifically, we select mainstream LLMs, such as GPT-3.5 (OpenAI, 2022), GPT-4 (Achiam et al., 2023), Falcon (Penedo et al., 2023), Llama2-70B (Touvron et al., 2023), and Wizardlm-13B (Xu et al., 2023). Since our deception dataset is in Chinese, we also select some LLMs that perform well in Chinese, including ERNIE3.5, ERNIE4.0, Qwen-14B (Bai

| Model | Accuracy | Completeness | Logic | Depth | Sum |
|---|---|---|---|---|---|
| GPT-4 | $9.00 \pm 0.28$ | $7.50 \pm 0.28$ | $8.50 \pm 0.28$ | $6.50 \pm 0.28$ | 31.50 |
| ERINE4.0 | $6.75 \pm 0.26$ | $6.37 \pm 0.26$ | $7.37 \pm 0.26$ | $5.87 \pm 0.12$ | 26.36 |
| GPT-3.5 | $6.00 \pm 0.10$ | $5.87 \pm 0.41$ | $6.87 \pm 0.12$ | $5.75 \pm 0.21$ | 24.49 |
| Falcon | $7.00 \pm 0.20$ | $5.50 \pm 0.28$ | $6.75 \pm 0.50$ | $5.12 \pm 0.41$ | 24.37 |
| Qwen-14B | $6.00 \pm 0.28$ | $5.75 \pm 0.50$ | $6.75 \pm 0.21$ | $5.25 \pm 0.50$ | 23.75 |
| Llama2-70B | $5.00 \pm 0.28$ | $5.68 \pm 0.63$ | $6.68 \pm 0.63$ | $5.75 \pm 1.64$ | 23.11 |
| ERINE3.5 | $5.25 \pm 0.26$ | $5.12 \pm 0.98$ | $6.25 \pm 1.35$ | $5.12 \pm 0.69$ | 21.74 |
| Baichuan2-13B | $5.24 \pm 0.55$ | $5.00 \pm 0.85$ | $6.25 \pm 0.78$ | $4.62 \pm 0.55$ | 21.11 |
| Wizardlm-13B | $5.00 \pm 0.28$ | $4.87 \pm 0.69$ | $6.00 \pm 0.57$ | $4.70 \pm 0.50$ | 20.57 |
| Chatglm2-6B | $3.50 \pm 0.21$ | $3.25 \pm 0.78$ | $4.00 \pm 1.71$ | $3.25 \pm 0.50$ | 14.00 |

Table 4: Deception reasoning performance of different LLMs.

et al., 2023), Chatglm2-6B (Du et al., 2021), and Baichuan2-13B (Yang et al., 2023).

In Table 4, GPT-4 outperforms other LLMs on all evaluation metrics, demonstrating its strong text understanding capabilities. Although lagging behind GPT-4, other LLMs can also deal with deception reasoning to some extent, with appropriate use of evidence to support the conclusions. Meanwhile, Qwen-14B is comparable to Llama2-70B in most metrics, although far behind in the number of parameters. The reason lies in that Llama2-70B is mainly trained with English corpora, thereby performing relatively weakly in Chinese.

### 4.5 Ablation Study

This paper leverages a two-stage strategy and GPT-4 for target content and action extraction (see Section 3.4). In this section, we compare the performance between one-stage and two-stage strategies, as well as GPT-3.5 and GPT-4. For target content extraction, we use the *target accuracy* as the evaluation metric. If the system extracts non-target content from legal instruments, it will have a low score in this metric. For action extraction, we use the *action complexity* as the metric. If the system cannot accurately realize the action decomposition process, it will have high *action complexity*. Therefore, a good model should have high *target accuracy* and low *action complexity*. Table 5 shows the experimental results for different strategies.

In Table 5, we observe that the two-stage strategy can always achieve better performance than the one-stage strategy. The reason lies in that if we merge target content and action extraction into one stage, it will increase the difficulty of the task, making it more likely that the output does not meet the predefined requirements.

| 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 5: Performance comparison of different strategies for target content and action extraction.

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 gives an initial attempt to construct the deception dataset using GPT-4. Specifically, we ask GPT-4 to mock the real interrogation between a police officer and a suspect. During interrogation, the suspect needs to deceive the police officer, while the police officer needs to find out the truth and seize evidence. Besides this dataset, we extend the traditional deception detection task into deception reasoning and further define the evaluation metrics and evaluators for this task. We also evaluate the deception reasoning performance of mainstream LLMs. Experimental results show that GPT-4 can achieve the best performance among all LLMs, demonstrating its strong text-understanding abilities. This dataset can also serve as a new reasoning benchmark for further research on LLMs.

## 6 Limitations

We acknowledge several limitations that can be addressed in future research. First, the construction of our deception dataset relies on GPT-4, which requires a lot of API call costs. Therefore, we only sample 191 legal instruments from CAIL2018, instead of using the entire dataset for dialogue generation. Future research will consider using the entire dataset to generate more samples. Secondly, we assess deception reasoning in mainstream LLMs but do not cover all existing LLMs. In the future, we will expand the evaluation scope. Thirdly, this paper focuses on text-based dialogues. Recently, video generation has become increasingly popular. In the future, we will explore generating video-based conversations for deception reasoning.

## 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.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*.

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.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

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

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.

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*.

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.

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. 2022. Chatgpt.

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.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with

web data, and web data only. *arXiv preprint arXiv:2306.01116*.

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*.

Rylan Schaeffer. 2023. Pretraining on the test set is all you need. *arXiv preprint arXiv:2309.08632*.

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.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. 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 1768–1777.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

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*.

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.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.