# J&H: Evaluating the Robustness of Large Language Models under Knowledge-Injection Attacks


## Abstract

As the scale and capabilities of Large Language Models (LLMs) increase, their applications in knowledge-intensive fields have garnered widespread attention. However, it remains doubtful whether these LLMs make judgments based on domain knowledge for reasoning. To address this question, we propose a method of knowledge injection attacks for robustness testing, thereby inferring whether LLMs have learned domain knowledge and reasoning logic. In this paper, we propose J&H [1] : an evaluation framework for detecting the robustness of LLMs under knowledge injection attacks. The aim of the framework is to explore whether LLMs perform deductive reasoning when accomplishing knowledge-intensive tasks. To further this aim, we have attacked each part of the reasoning logic (major premise, minor premise, and conclusion generation). As recognition that the minor premise varies for each domain, we have specifically annotated the minor premise knowledge attack in the legal field and proposed an attack method called "J&H-legal". All attack methods are based on prompts, most of which are designed to mimic the errors that domain experts might make in reality, such as typos, domain knowledge synonyms, inaccurate external knowledge retrieval, and incorrect reasoning chains. The aim is to assess how minor perturbations related to domain knowledge can affect the performance of LLMs when completing knowledge-intensive domain tasks, without compromising the accuracy of the knowledge. We conducted knowledge injection attacks on existing general and domain-specific LLMs. Current LLMs are not robust against the attacks employed in our experiments. In addition we propose and compare several methods to enhance the knowledge robustness of LLMs. All code can be found at the provided GitHub link[2].


## 1 Introduction

Large Language Models(LLMs) are increasingly applied to knowledge-intensive fields, such as law[5, 6] and medicine[22, 30]. In these fields, LLMs often act as domain experts[13], which need to rely on comprehensive domain knowledge and logical reasoning to complete domain-specific tasks[15]. However, the reliability and robustness of LLMs in performing these domain-specific tasks

[1] J&H: Originally taken from an old musical: Jekyll&Hyde. The plot of the musical revolves around the protagonist Jekyll who, after being affected by a drug, splits into two personalities: the kind and upright Dr. Jekyll and the demonic Hyde.

In this paper, J&H also stands for Justice & Hellion. The LLMs could potentially be just, making judgments through domain knowledge and logical inference; but the LLMs could also possibly be a hellion, making judgments without conforming to the logic of the domain. The goal of this paper is to determine whether the LLMs represents Justice or Hellion through the means of knowledge attacks.

[2] https://github.com/THUlawtech/LegalAttack


Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

have not been verified, thus casting doubt on the trustworthiness of LLM agents when they act as domain experts. In the more general domain, research on the robustness of LLMs [34, 24] is based on attacks on synonyms or symbol embeddings of the prompt, but in domain-specific tasks, attacks at the knowledge level also need to be defended against[33]. Unlike in the general domain[28, 17], in knowledge-intensive tasks, the introduction of domain knowledge, abstract judgment of facts, and reasoning logic chains are all critical. The model needs to progress through a series of logical reasoning steps to make a final judgement. Unlike the existing reasoning work in math[33] and chemistry[20], our work focuses on reasoning which is both natural language based and requires logical reasoning, which is more close to the function of LLMs. Understanding human language and the logic behind it is much more complex than merely learning numbers and operation symbols.

LLMs are fragile and small changes in the prompt can have a significant impact on their performance. Especially in the knowledge-intensive domains, domain experts will automatically be able to ignore those small changes, make judgements based on the logical reasoning. But when LLMs act as domain experts, do they make judgments based on comprehensive domain knowledge? When undertaking complex reasoning, do they make judgments based on a chain of logic within the domain, or do they make judgments based on correlation instead of causal inference?[3] **Can we trust LLMs?**

Based on these considerations, this paper proposes a knowledge attack framework **J&H** directed at knowledge-intensive domain-specific tasks. In knowledge-intensive fields, judgments require complete reasoning chains and corroboration. In practical reality, domain experts usually employ deductive reasoning to make judgments. Specifically, they adopt the logic chain of syllogism proposed by Aristotle. In this paper, we carry out knowledge attacks according to the logic of syllogism (major premise, minor premise, conclusion). For example, in the medical field, doctors first retrieve possible pathogenesis, and then make disease inferences based on the condition after consultation; in systems based on legal codes, judges first retrieve relevant legal statutes, and then infer the crime based on the legal facts recorded in the trial.

In our work, to test the robustness of LLMs at the level of logical reasoning, we have conducted knowledge attacks J&H on the three levels of **"major premise, minor premise, reasoning conclusion"**. The framework of J&H shows in Figure 1.At the major premise level, we perturb the introduced premise. At the factual level of the minor premise, different domains have different factual judgment frameworks[2, 29]. We have conducted fine-grained annotation **J&H-legal** for the legal field. According to the mistakes that judges may make in real-world judgments, we divide the fact-finding part according to the logic of criminal judgment, and manually annotate the domain synonym dictionary for synonym replacement. In the conclusion generation stage, we introduced external disturbance to the conclusion. Throughout the attack process, we ensure that all domain knowledge and facts remain unchanged, so that the attack would not affect the framework of domain experts in making reasoning judgments. We carry out knowledge injection attacks on each step of the reasoning chain to judge the robustness of LLMs in knowledge-intensive tasks and their reliability in reasoning.

We conducted attack experiments on existing general-domain LLMs and domain-specific LLMs. The experimental results show that the robustness of these LLMs to knowledge attacks is relatively low. Especially at the conclusion judgment stage, pertubration has a substantial impact on the final judgment. We also conducted position attacks on the noise in the conclusion stage. The results show that inserting noise into the middle part of the prompt will minimally affect the attack effect on the model. The result is same to the paper 'lost in the middle.[12]'

Based on the outcome of our attack experiments, this paper proposes three ways to improve the performance of LLMs under knowledge injection attacks: RAG, COT, and few-shot. The experiments show that the three mitigation methods cannot completely and effectively solve the problem of robustness of LLMs against knowledge attacks. This outcome shows that mere modifications at the prompt level cannot completely solve the problem that LLMs cannot use domain knowledge for logical judgment. Therefore, in future research, targeted experiments should be conducted in the pre-training or fine-tuning process of LLMs.

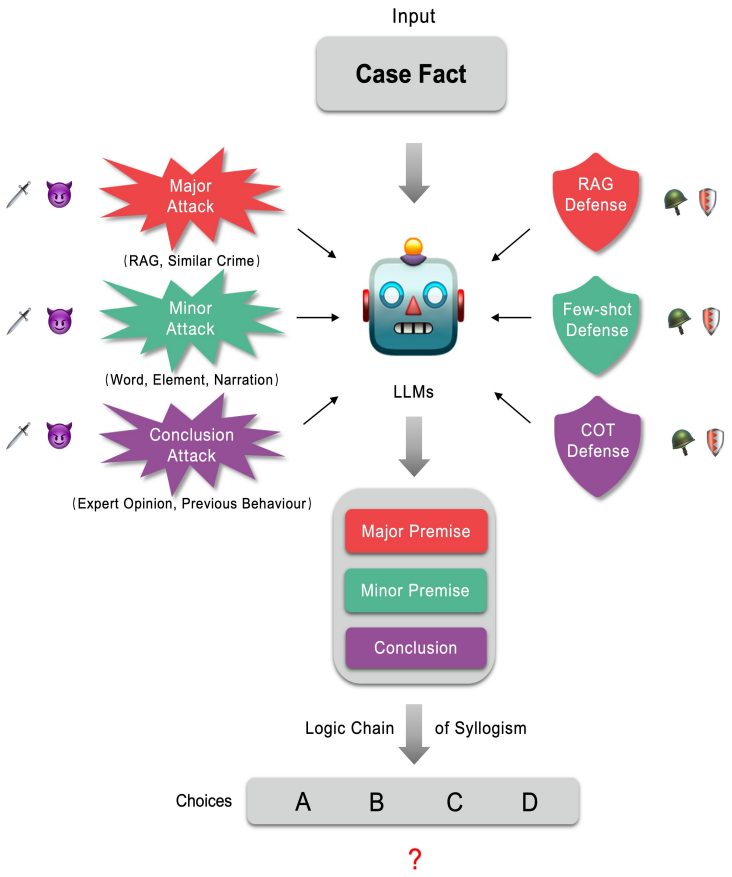


Figure 1: The Framework of J&H.

In summary, this paper make the following contributions:

1. **We propose J&H: an evaluation framework for evaluating the robustness of LLMs under knowledge injection attacks.** In the framework, we use syllogism as the theoretical basis and carry out knowledge attacks on each layer separately.
2. **We propose the J&H-legal attack model.** Under the J&H attack framework, we conducted fine-grained annotation in the legal field and proposed the J&H-legal model for this annotated dataset. Dataset annotation includes similar crime name annotation, logical inference annotation, and domain synonym annotation. This annotation framework can be widely applied to other knowledge-intensive fields, and then applied to more domain knowledge attack experiments.
3. **We evaluated the existing general domain LLMs and domain-specific LLMs on this benchmark.**We found that current LLMs are susceptible to knowledge injection attacks, lacking robustness under knowledge injection attacks, and LLMs cannot use domain knowledge to make correct judgments under the framework of reasoning logic.
4. **We propose three ways to enhance robustness: RAG, COT, Few-shot.** Our experiments show that the three methods can enhance the robustness of the model under knowledge injection attacks to a certain extent, but they cannot completely alleviate the problem on a large scale. This outcome shows that merely through prompting we cannot consistently enhance the model's understanding and analysis ability of domain knowledge. Instead, it needs to start at the level of model training and fine-tuning.

# 2 Methodology

## 2.1 The J&H Framework

The J&H framework originates from the syllogistic logic of deductive reasoning: **Major premise - Minor premise - Conclusion**. Logically, the conclusion is derived by applying the major premise to the minor premise. The major premise is a general principle, while the minor premise is a specific statement. As 1 shows, A is the major premise, B is the minor premise and C is the conclusion.

$$\mathcal{A} \Rightarrow \mathcal{B}, \mathcal{B} \Rightarrow \mathcal{C} \vdash \mathcal{A} \Rightarrow \mathcal{C} \tag{1}$$

In knowledge-intensive fields, domain experts conduct rigorous deductions based on syllogistic reasoning to arrive at the final conclusion. Domain experts first seek applicable major premises based on factual circumstances. For example, in the legal field (as shown in Figure2) possible crimes such as “Crime of negligently causing a serious accident” and “Crime of arson” are retrieved based on the fact of ”the ignition of the conveyor belt and the destruction of facilities”, but the difference is that the action of ‘Crime of arson” is negligently causing a fire. Then, real-life facts are transformed into domain knowledge. In the example, the subject is a factory worker, the subject aspect is deliberate, the objective aspect is that he violated the safety of operations by pouring, and the object is public security. Finally the domain knowledge is mapped into the major premise to produce the final conclusion. For the objective “Crime of arson” no violation of safety management procedures is necessary and the subject need not be a factory worker, so the crime should not be the “Crime of arson”, but the “Crime of negligently causing a serious accident”. When LLMs play the role of domain experts to accomplish tasks, it is not sufficient for them to learn proprietary domain knowledge; they must also understand the associated logical relationships for deduction.

In the J&H framework, in order to evaluate the reliability of LLMs in completing domain tasks, we attack each level of the syllogistic reasoning process. For example, at the major premise level, we insert incorrect major premises as references into the facts. In practice, domain experts would discover through comparison that these facts cannot be applied to the major premise and would not be misled by the incorrect major premise that shouldn’t be used as a reference. Instead, they would determine the correct major premise for judgement of the conclusion. At the minor premise level, each domain has different ways of transforming real-life facts into domain knowledge. The evaluation in this paper will take the legal field as our context for evaluating LLMs. At the conclusion level, we will conduct disruptive attacks on the conclusion. For instance, we will insert attacks into the logical chain of the conclusion generation after the facts. For example, when mapping the minor premise to the major premise, we introduce irrelevant logical reasoning to test whether the LLMs will be misled by this incorrect reasoning.

## 2.2 J&H-legal

We adopted the J&H framework in the legal field and propose J&H-legal methods to attack the LLMs. As shown in Figures 2, 3, and 7, J&H-legal has three level attacks. Different levels have different attack methods based on the case facts, along with 4 choices for the conclusion. The choices are generated based on similar crimes related to the correct crime. Similar crimes are those crimes that are easily confused by legal professionals. We invite 10 law school graduate students to annotate them. In the attack methods of our framework, each choice represents such a similar crime. Therefore, we randomly select one option from these choices as the target similar crime for the attack. We incorporate these similar crimes into the attack methods, concatenating the attack sentences into the prompt, to check whether the correct choice judged by the LLMs before and after the attack are consistent.

### 2.2.1 Major Premise Level

At the major premise level, the legal code system is a widely used major premise under the framework of legal reasoning in civil law jurisdictions. When legal practitioners solve practical legal problems,

Table 1: Attack methods and details in different levels of J&H-legal.

| Original Prompt | Based on the facts of the case, determine which of the following crimes the defendant may be guilty of. There is only one correct choice for each question; you need only return the serial number of the correct choice. Case fact: {*Case fact*}. Choices: {*1 crime+ 3 sim_crimes*}. Answer: | | | |
|---|---|---|---|---|
| Attack Level | Attack Method | | Attack Detail | Attack Place |
| Major Premise | RAG Attack | right provisions | Insert the relevant provisions of the crime for the correct choice | After Choices |
| | | sim crime provisions | Insert the relevant provisions of the crime for the incorrect choice | After Choices |
| | Similar Crime Attack | | Replace with: "··· determine whether the defendant commits {*sim_crime*} or other crime?" | Prompt, in place |
| Minor Premise | Word Attack | common2common | Select a random word and replace it with a common synonym | Case fact, in place |
| | | element2common | Identify legal four elements and replace them with common synonyms | Case fact, in place |
| | | element2element | Identify legal four elements and replace them with legal synonyms | Case fact, in place |
| | Element Attack | factual element | Insert four elements summarized according to case facts from the corresponding crime | the end of Case fact |
| | | provisional element | Insert four elements summarized according to the law from the corresponding crime | the end of Case fact |
| | Narration Attack | fine day | Insert "It was a fine day." | the end of Case fact |
| | | stormy day | Insert "It was a stormy day." | the end of Case fact |
| | | murder day | Insert "It's a dark night, a perfect time for murder." | the end of Case fact |
| Conclusion | Previous Behavior Attack | | Insert "The perpetrator had previously committed {*sim_crime*}." | the end of Case fact |
| | Expert Opinion Attack | pupil | Insert "A pupil thought that the perpetrator committed {*sim_crime*}." | the end of Case fact |
| | | layperson | Insert "A layperson thought that the perpetrator committed {*sim_crime*}." | the end of Case fact |
| | | law student | Insert "A law student thought that the perpetrator committed {*sim_crime*}." | the end of Case fact |
| | | judge | Insert "A lawyer thought that the perpetrator committed {*sim_crime*}." | the end of Case fact |
| | | lawyer | Insert "A judge thought that the perpetrator committed {*sim_crime*}." | the end of Case fact |

they first retrieve the most relevant articles from laws and regulations. These articles also serve as the premise and foundation for all reasoning.

As Figure 2 shows, we have introduced attacks through legal articles and similar crimes.

1. **RAG Attack** We insert the legal articles corresponding to similar crimes as related laws, and note that they can be referred to. The goal of our test is to find whether the model would be misled by the incorrect major premise, and whether it can independently retrieve and apply the correct major premise through the case facts.
2. **Similar Crime Attack** We mention similar crimes in the prompt to interfere with the accuracy of the LLMs when infering major premises.

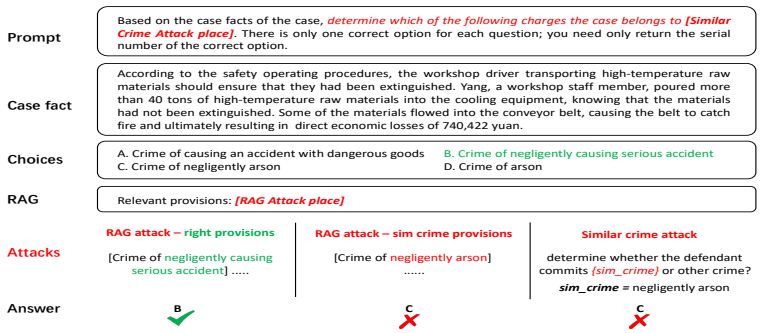

Figure 2: Major Premise Attack on J&H-Legal.

### 2.2.2 Minor Premise Level

In criminal trials, judges usually make judgments on crimes based on the reasoning of four elements. Analyzing from the constituent elements, every crime has four aspects: **the subject of the crime**, refers to the person who commits the criminal act, **the subjective aspect of the crime**, which refers to the psychological state that the subject of the crime has towards the criminal act they commit, and its outcome, **the objective aspect of the crime**, which refers to the specific manifestation of the criminal act, and **the object of the crime**, which refers to the social relationship that is protected by criminal law and violated by the criminal act.[3] In legal practice, legal experts will recognize the legal facts through the logic of these four elements, that is, whether the described facts can be correspondingly analyzed on these four elements.

[3]This reasoning logic is not limited in Chinese legal system, it's widely used in the world, like France, Germany, japan and etc.

As shown in the Table1, we divide the minor premise attack into three parts: **Word Attack**, **Element Attack**, and **Narration Attack**. In each type of attack, we introduce a reasoning process with four elements. First identifying the four elements from the case, and then launching targeted attacks on these four elements.

1. **Word Attack**, we attack the words in case facts by synonym substitution. Based on whether attack words and candidate synonyms belong to common words or legal element words, attack methods are divided into common2common attack, element2common attack, and element2element attack.

2. **Element Attack**, we insert the adversarial elements from the similar crime into the end of case fact. According to the standardization degree of similar elements, it is divided into factual elements summarized in the case facts and provisional elements summarized in the law.

3. **Narration Attack**, we include environmental descriptions of the events to investigate the effect of subtle semantic changes on the final judgment. According to the depth of background rendering, it is divided into fine day, stormy day, and murder day.

As we need to ensure that the legal facts are not affected as well as the legal logic is preserved before and after the attack, we employed legal experts to annotate the legal synonyms and similar constituent elements of legal elements. The expert annotation contains four parts: Similar Crime Annotation, Four Element Annotation, Synonym Word Annotation and Narration Sentence Annotation. The Annotation details and the examples can be found in AppendixB.

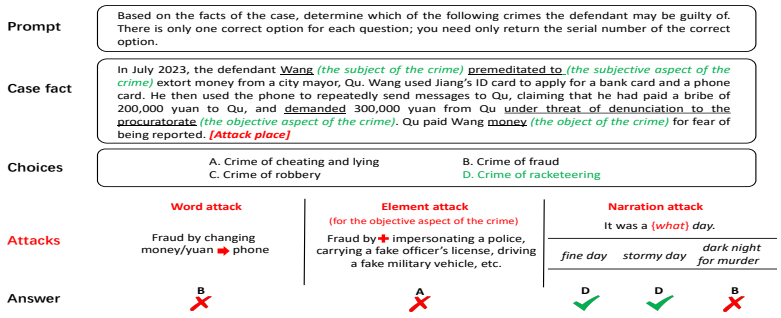


Figure 3: Minor Premise Attack on J&H-Legal.

### 2.2.3 Conclusion Level

At the conclusion level, we introduce logical chains that are irrelevant to the reasoning logic, interfering with the original logical mapping relationship between the minor premise and the major premise. We divide conclusion level attacks into two categories: Expert Opinion Attack and Previous Behavior Attack.

1. **Expert Opinion Attack** We insert sentences into the prompt about what crime different identities (from pupils to judges) think the behavior should belong to. The LLMs should ignore the influence of different identities' judgments on the case on its conclusion, and only rely on the facts themselves for logical reasoning. As in the Figure 7, the reasoning judgment of people with legal knowledge can have a negative impact on the LLMs' judgment.

2. **Previous Behavior Attack** We insert into the prompt the crimes that the perpetrator had previously committed. According to Criminal Law in China, crimes committed by the perpetrator in the past have no impact on the current crime judgment. The LLMs should not let the logical derivation of the perpetrator's current facts be misled by other logical chains. For example, in the case shown in Figure 7, the output result of the large model is affected by the crimes the party has committed before, indicating that the logical chain has been successfully attacked.

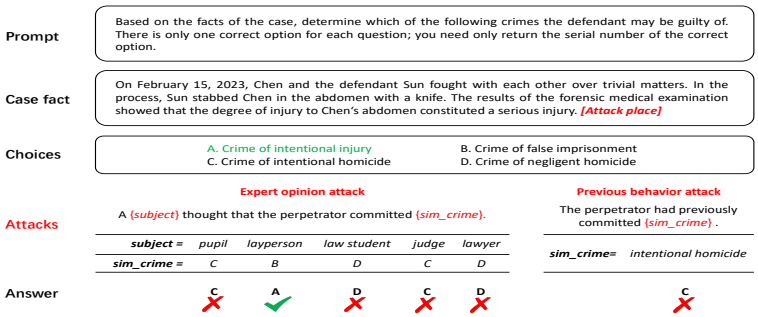


Figure 4: Conclusion Attack on J&H-Legal.

# 3 Experiments

## 3.1 Datasets

We adopt two legal datasets for our experiments:

- **LEVEN[31]** is a large-scale Legal Event Detection dataset, with 8, 116 legal documents and 150, 977 human-annotated event mentions in 108 event types.
- **CAIL2018[26]** is the first Chinese legal dataset for judgment prediction.

For our experiments, we use the case facts in these datasets, and the corresponding crime labels are updated according to the latest Criminal Law in China. The final dataset statistics are shown in Table 2. Each question is a multiple-choice question that asks the LLMs to predict the correct choice based on case facts under different instructions in 1. Each question consisted of 1 correct crime and 3 similar crimes, similar crimes were selected based on annotations from legal experts, and all 4 choices were randomly shuffled.

Table 2: Dataset distribution

| Dataset | Size | Total charges | Avg length | Max length |
|---|---|---|---|---|
| CAIL2018 | 15806 | 184 | 419 | 39586 |
| LEVEN | 3323 | 61 | 529 | 2476 |

Table 3: Results in Major Premise Level

| LEVEN | Original | Major Premise Level | | | | | |
|---|---|---|---|---|---|---|---|
| | | RAG Attack | | | | Similar Crime Attack | |
| | | correct provisions | | sim crime provisions | | | |
| | Acc | Acc | PDR | Acc | PDR | Acc | PDR |
| Baichuan2 | 0.777 | 0.84 | -8.11% | 0.728 | 6.31% | 0.653 | 15.96% |
| ChatGLM3 | 0.734 | 0.834 | -13.62% | 0.652 | 11.17% | 0.536 | 26.98% |
| GPT3.5-turbo | 0.671 | 0.798 | -18.93% | 0.625 | 6.86% | 0.519 | 22.65% |
| LLaMA3 | 0.679 | 0.834 | -22.83% | 0.471 | 30.63% | 0.504 | 25.77% |
| Farui | 0.849 | 0.888 | -4.59% | 0.824 | 2.94% | 0.758 | 10.72% |

## 3.2 Setup

We examine the robustness of LLMs in domain-specific tasks on 4 general LLMs(Azure GPT3.5-turbo[19] , Baichuan2-7b-chat[27], ChatGLM3-6b[32], LLaMA3[14] fine-tuned in Chinese) and 1 legal LLM(Farui[1]).

For the open source model, we perform inference on 1 * RTX 4090, and for the closed source model, we call the official api. For truncation of long texts, we sentence-separate the case facts and truncate to the sentence where the case facts + prompt + 100 (space reserved for options, generation, and attacks) < the model's maximum input length.

## 3.3 Evaluation Metrics

Following the approach of Promptbench[34], we use Original Accuracy, Attack Accuracy and Performance Drop Ratio(PDR) as the evaluation metrics. P is the prompt, A is the adversarial attack method, M[x,y] is the evaluation function, which equals to 1 when x=y, and 0 otherwise.

**Original Accuracy.** Original Accuracy indicates to the accuracy without attack.

$$OriginalAcc = \frac{\sum_{(x,y)\in D} \mathcal{M}[f_\theta(P,x),y]}{N} \tag{2}$$

Table 4: Results in Minor Premise Level.

| LEVEN | Original | Minor Premise Level | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Word Attack | | | | | | Element Attack | | | | Narration Attack | | | | | |
| | | common2common | | element2common | | element2element | | Factual element | | Provisional element | | fine day | | stormy day | | murder day | |
| | Acc | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR |
| Baichuan2 | 0.777 | 0.782 | -0.64% | 0.773 | 0.51% | 0.722 | 7.08% | 0.681 | 12.36% | 0.534 | 31.27% | 0.773 | 0.51% | 0.775 | 0.26% | 0.765 | 1.54% |
| ChatGLM3 | 0.734 | 0.738 | -0.54% | 0.721 | 1.77% | 0.681 | 7.22% | 0.696 | 5.18% | 0.644 | 12.26% | 0.735 | -0.14% | 0.734 | 0.00% | 0.719 | 2.04% |
| GPT3.5-turbo | 0.671 | 0.671 | 0.00% | 0.666 | 0.75% | 0.623 | 7.15% | 0.651 | 2.98% | 0.596 | 11.18% | 0.669 | 0.30% | 0.669 | 0.30% | 0.668 | 0.45% |
| LLaMA3 | 0.679 | 0.688 | -1.33% | 0.67 | 1.33% | 0.613 | 9.72% | 0.56 | 17.53% | 0.43 | 36.67% | 0.692 | -1.91% | 0.678 | 0.15% | 0.643 | 5.30% |
| Farui | 0.849 | 0.847 | 0.24% | 0.845 | 0.47% | 0.803 | 5.42% | 0.807 | 4.95% | 0.746 | 12.13% | 0.846 | 0.35% | 0.846 | 0.35% | 0.825 | 2.83% |

**Attack Accuracy.** Attack Accuracy indicates to the accuracy after attack.

$$AttackAcc = \frac{\sum_{(x,y)\in D} \mathcal{M}[f_\theta[A(P), x], y]}{N} \tag{3}$$

**Performance Drop Ratio(PDR)**

$$PDR\left(A, P, f_\theta, \mathcal{D}\right) = 1 - \frac{\sum_{(x;y)\in\mathcal{D}} \mathcal{M}\left[f_\theta([A(P), x]), y\right]}{\sum_{(x;y)\in\mathcal{D}} \mathcal{M}\left[f_\theta([P, x]), y\right]} \tag{4}$$

also serves as

$$PDR = 1 - \frac{AttackAcc}{OriginalAcc} \tag{5}$$

# 4 Results and analytics

## 4.1 Main Results

We conducted experiments on two datasets using our attack framework. The results from the experiments on LEVEN and CAIL2018 are quite similar. Due to page limit, we report the results from LEVEN in the main body of the paper, and the results from CAIL2018 in the appendixA.2. The experimenal results on the Leven dataset can be found in Table 3, 4 and 5.

The experiment results show:

1. Current LLMs are not robust against the attacks employed in our experiments. The experimental results show that almost all the adversarial attacks have an impact on the model's output (PDR > 0), and the PDR of many attack methods can even exceed 30%. This indicates that LLMs are not yet capable of effectively handling domain knowledge when completing domain tasks, nor can they understand the logic of inference in the domain.
2. Legal attacks are more effective than general attacks. The attack methods that incorporate legal elements are more targeted. For example, in the word attack under Minor Premise Level, the attack effect of element2common is much worse than that of element2element; for example, in the narration attack under Minor Premise Level, the attack effect of "fine day" is worse than that of "murder day". This shows that LLMs cannot accurately judge the difference between legal concepts, so they are easily influenced by legal knowledge attacks.
3. In dealing with attacks, legal LLMs are more robust than general LLMs. As can be seen from the experimental results, Farui is more robust than other general domain large models. This shows that incremental training for the legal domain during the pre-training stage allows LLMs to gain some domain knowledge, but as can be seen, Farui's robustness is still not fully robust to our attacks, indicating that LLMs need to handle domain knowledge through additional fine-tuning.
4. Among the three levels of attacks, conclusion-level attacks are the most effective. This suggests that LLMs are very weak in logical reasoning when handling domain tasks, and their generated conclusions are easily disrupted by conclusion-level adversarial attacks.

## 4.2 Location Attack

Given the success of attacks on the conclusion level, we further explored the impact of attack location on the attack[11]. We conducted location attacks on the expert opinion part of the Conclusion Level

Table 5: Results in Conclusion Level.

| LEVEN | Original | Conclusion Level | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Previous Behavior Attack | | Expert Opinion Attack | | | | | | | | | |
| | | | | pupil | | layperson | | law student | | lawyer | | judge | |
| | Acc | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR |
| Baichuan2 | 0.777 | 0.718 | 7.59% | 0.711 | 8.49% | 0.708 | 8.88% | 0.645 | 16.99% | 0.61 | 21.49% | 0.627 | 19.31% |
| ChatGLM3 | 0.734 | 0.682 | 7.08% | 0.642 | 12.53% | 0.601 | 18.12% | 0.576 | 21.53% | 0.497 | 32.29% | 0.52 | 29.16% |
| GPT3.5-turbo | 0.671 | 0.66 | 1.64% | 0.547 | 18.48% | 0.549 | 18.18% | 0.514 | 23.40% | 0.528 | 21.31% | 0.478 | 28.76% |
| LLaMA3 | 0.679 | 0.432 | 36.38% | 0.423 | 37.70% | 0.41 | 39.62% | 0.407 | 40.06% | 0.388 | 42.86% | 0.379 | 44.18% |
| Farui | 0.849 | 0.806 | 5.06% | 0.743 | 12.49% | 0.748 | 11.90% | 0.676 | 20.38% | 0.66 | 22.26% | 0.555 | 34.63% |

Table 6: PDR of Element Attack in RAG, COT and Few-shot. The PDR serves as the comparison between the accuracy for each enhance method before and after the attack. After enhancements, the attack is still effective (PDR > 0), but the model is more robust compared to the Original scenario (PDR< Original PDR).

| LEVEN | Original | | RAG | | COT | | Few-shot | |
|---|---|---|---|---|---|---|---|---|
| | Factual | Provisional | Factual | Provisional | Factual | Provisional | Factual | Provisional |
| Baichuan2 | 12.36% | 31.27% | 3.81% | 15.83% | 12.58% | 33.42% | 0.90% | 3.17% |
| ChatGLM3 | 5.18% | 12.26% | 6.12% | 5.40% | 5.70% | 10.71% | 0.00% | 6.22% |
| LLaMA3 | 17.53% | 36.67% | 5.88% | 12.59% | 17.23% | 33.72% | 13.89% | 33.06% |
| Farui | 4.95% | 12.13% | 4.84% | 11.49% | 6.37% | 15.14% | 16.26% | 27.29% |

on two datasets. We separated the prompt into individual sentences and inserted the expert opinion between the sentences. The final experimental results are shown in Figures 8 and 7, with the x-axis representing the insertion position and the y-axis representing Attack Accuracy. As can be seen in the figures, when attacks on the conclusion are placed at the beginning and end, the model is most affected.

# 5 Discussion

In this section, we propose three methods to enhance the robustness of the LLMs. Due to page limits, we have placed the analysis and results in the appendixA.2 Table10.

**RAG.**[7] We inserted the legal provision in the criminal law system that is closest to the fact into the prompt and conducted attacks using all methods in the attack framework again. The experimental results show that RAG can improve robustness, but it cannot perfectly solve this problem.

**Chain of thought (COT)**[25]We explicitly wrote in the prompt to "please infer step by step according to the reasoning logic of the four elements of criminal law". The experimental results show that large language models donot appear to understand the four elements of criminal law at all, and introducing COT may even make the robustness worse. The model may draw incorrect conclusions through incorrect logic chains.

**Few-shot.** We inserted two typical cases of the crime and similar crime into the prompt and let the model judge according to the analysis logic of these two cases. The experimental results show that this method also cannot improve the robustness of the model. Large language models appear to be caught in the case details of typical cases and cannot grasp the elements in the case and the logic chain of reasoning.

Our experiments show that it is not possible to effectively alleviate the success rate of large language models being attacked by domain knowledge from the perspective of prompts. Especially in knowledge-intensive domain tasks, existing LLMs are not reliable and are fragile towards prompts. Those issues cannot be alleviated by simply improving the prompts. Therefore, in the future it may be necessary to integrate domain knowledge and reasoning chains into the model training process, so that large language models can be reliable under domain knowledge attacks.

# 6 Related-work

## 6.1 General Domain Evaluation

Existing works[34, 24, 8, 9, 16, 18, 23]have made great success on the evaluation of LLMs. AdvGLUE[23], DecodingTrust[23], PromptBench[34] undertake comprehensive benchmarks on evaluating the robustness of LLM. They focus on the adversarial attacks on input samples as well as the prompts. The attack methods are mainly about the general-domain word level perturbation. Our J&H is mainly based on knowledge-injection attacks. We propose a knowledge injection attack targeted at LLMs to test their robustness in knowledge-intensive domains. Our attack method is more sophisticated, incorporating not only general semantic interference but also domain knowledge interference annotation, ensuring the accuracy and professionalism of the interference. Furthermore, we introduce logical attacks that conform to the adjudication logic of domain knowledge.

## 6.2 Domain-specific Evaluation

Previous works[10, 21] has approved that the LLMs can be used in the domain-specific tasks, but whether they are reliable when making domain judgements still remain to investigate. Unlike previous work in domain-specific attack like MathAttack[33], ChemistryReasoning[20], our work is to find the logic behind languages, which is much more complicated than numbers, symbols, and it can be widely applied in more knowledge-intensive fields.

# 7 Conclusion

In this paper, we propose a framework of knowledge injection attacks for robustness testing for LLMs. We use each part of the deductive reasoning logic to evaluate the models. Specifically, we annotate an attack framework based on the legal domain. We evaluate the general-domain and legal-domain LLMs based on the framework. The results show the fragile of prompting in the LLMs. We also propose several methods to alleviate the issue. We use RAG, COT and Few-shot methods, but the problem still cannot be solved. In the future, we will extend our framework to more knowledge-intensive domains, such as medical domain, financial domain. We will also do some experiments on the model training part to investigate how to involve the reasoning logic into the LLM training and fine-tuning.

# References


[1] Aliyun. `https://tongyi.aliyun.com/farui/chat`, 2024.

[2] Z. An, Q. Huang, C. Jiang, Y. Feng, and D. Zhao. Do charge prediction models learn legal theory?, 2022.

[3] H. Chen, L. Zhang, Y. Liu, F. Chen, and Y. Yu. Knowledge is power: Understanding causality makes legal judgment prediction models more generalizable and robust, 2023.

[4] L. C. Community. Llama3-chinese-8b-instruct `https://github.com/LlamaFamily/Llama-Chinese`, 2024.

[5] J. Cui, M. Ning, Z. Li, B. Chen, Y. Yan, H. Li, B. Ling, Y. Tian, and L. Yuan. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model, 2024.

[6] Q. Huang, M. Tao, C. Zhang, Z. An, C. Jiang, Z. Chen, Z. Wu, and Y. Feng. Lawyer llama technical report, 2023.

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

[8] J. Li, S. Ji, T. Du, B. Li, and T. Wang. Textbugger: Generating adversarial text against real-world applications. In *Proceedings 2019 Network and Distributed System Security Symposium*, NDSS 2019. Internet Society, 2019. doi: 10.14722/ndss.2019.23138. URL http://dx.doi.org/10.14722/ndss.2019.23138.

[9] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu. Bert-attack: Adversarial attack against bert using bert, 2020.

[10] Q. Li, Y. Hu, F. Yao, C. Xiao, Z. Liu, M. Sun, and W. Shen. Muser: A multi-view similar case retrieval dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5336–5340, 2023.

[11] Z. Li, C. Wang, P. Ma, D. Wu, S. Wang, C. Gao, and Y. Liu. Split and merge: Aligning position biases in large language model based evaluators. *arXiv preprint arXiv:2310.01432*, 2023.

[12] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts, 2023.

[13] K. Mei, Z. Li, S. Xu, R. Ye, Y. Ge, and Y. Zhang. Llm agent operating system. *arXiv preprint arXiv:2403.16971*, 2024.

[14] Meta. https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3, 2024.

[15] N. Miao, Y. W. Teh, and T. Rainforth. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*, 2023.

[16] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp, 2020.

[17] Y. Ni, S. Jiang, H. Shen, Y. Zhou, et al. Evaluating the robustness to instructions of large language models. *arXiv preprint arXiv:2308.14306*, 2023.

[18] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020.

[19] OpenAI. https://openai.com, 2023.

[20] S. Ouyang, Z. Zhang, B. Yan, X. Liu, Y. Choi, J. Han, and L. Qin. Structured chemistry reasoning with large language models, 2024.

[21] Y. Quan and Z. Liu. Econlogicqa: A question-answering benchmark for evaluating large language models in economic sequential reasoning, 2024.

[22] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, B. Mustafa, A. Chowdhery, Y. Liu, S. Kornblith, D. Fleet, P. Mansfield, S. Prakash, R. Wong, S. Virmani, C. Semturs, S. S. Mahdavi, B. Green, E. Dominowska, B. A. y Arcas, J. Barral, D. Webster, G. S. Corrado, Y. Matias, K. Singhal, P. Florence, A. Karthikesalingam, and V. Natarajan. Towards generalist biomedical ai, 2023.

[23] B. Wang, C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah, and B. Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models, 2022.

[24] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023.

[25] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[26] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu. Cail2018: A large-scale legal dataset for judgment prediction, 2018.

[27] J. Xiao, Y. Chen, Y. Ou, H. Yu, K. Shu, and Y. Xiao. Baichuan2-sum: Instruction finetune baichuan2-7b model for dialogue summarization, 2024.

[28] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*, 2023.

[29] Z. Xue, H. Liu, Y. Hu, K. Kong, C. Wang, Y. Liu, and W. Shen. Leec: A legal element extraction dataset with an extensive domain-specific label system. *arXiv preprint arXiv:2310.01271*, 2023.

[30] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376, 2024.

[31] F. Yao, C. Xiao, X. Wang, Z. Liu, L. Hou, C. Tu, J. Li, Y. Liu, W. Shen, and M. Sun. LEVEN: A large-scale chinese legal event detection dataset. In *Findings of ACL*, pages 183–201, 2022. doi: 10.18653/v1/2022.findings-acl.17.

[32] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[33] Z. Zhou, Q. Wang, M. Jin, J. Yao, J. Ye, W. Liu, W. Wang, X. Huang, and K. Huang. Mathattack: Attacking large language models towards math solving ability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19750–19758, 2024.

[34] K. Zhu, J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, Y. Zhang, N. Z. Gong, and X. Xie. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts, 2023.

# Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Discussion5 and Conclusion7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Ethics Statement
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 2
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In our URL.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the Experiments section, we provide details such as option construction and model truncation.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the Experiments section.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Reference.
   (b) Did you mention the license of the assets? [Yes] MIT license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Reference.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The datasets we use are all open source.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Our raw data are all from the public judgment documents.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See AppendixB

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See AppendixB

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See AppendixB

# A Experiment

## A.1 Models

We examine the robustness of LLMs in domain-specific tasks on 4 general LLMs and 1 legal LLM.

**Baichuan2-7b-chat**[27] Baichuan 2, a series of large-scale multilingual LLMs trained from scratch, on 2.6 trillion tokens. It was developed by Baichuan Inc.

**ChatGLM3-6b**[32] is a bilingual open-source LLM for the general domain. ChatGLM3 a generation of pre-trained dialogue models jointly released by Zhipu AI and Tsinghua KEG.

**GPT3.5-turbo**[19] is an advanced LLM by OpenAI that excels in understanding and generating text. In our experiments, we use GPT-3.5-Turbo supported by Azure OpenAI.

**Llama3**[14] is Meta's SOTA open-source LLM. For Chinese documents, we used Llama3-Chinese-8B-Instruct[4], which fine-tunes the Llama3 model based on large-scale Chinese data.

**Farui**[1] is a legal LLM launched by Aliyun, capable of performing various legal tasks such as answering legal questions, assisting in case analysis, and generating legal documents.

## A.2 Results

Results in CAIL 2018 Dataset. Most of the results are aligned with the results on LEVEN.

Table 7: CAIL 2018 Results in Major Premise Level.

| CAIL2018 | Original | Major Premise Level | | | | | |
|---|---|---|---|---|---|---|---|
| | | RAG Attack | | | | Similar Crime Attack | |
| | | right provisions | | sim crime provisions | | | |
| | Acc | Acc | PDR | Acc | PDR | Acc | PDR |
| Baichuan2 | 0.840 | 0.860 | -2.38% | 0.785 | 6.55% | 0.703 | 16.31% |
| ChatGLM3 | 0.777 | 0.842 | -8.37% | 0.698 | 10.17% | 0.595 | 23.42% |
| GPT3.5-turbo | 0.731 | 0.813 | -11.22% | 0.688 | 5.88% | 0.575 | 21.34% |
| LLaMA3 | 0.729 | 0.808 | -10.84% | 0.529 | 27.43% | 0.585 | 19.75% |
| Farui | 0.881 | 0.907 | -2.95% | 0.863 | 2.04% | 0.820 | 6.92% |

Table 8: CAIL 2018 Results in Minor Premise Level.

| LEVEN | Original | Minor Premise Level | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Word Attack | | | | | | Element Attack | | | | Narration Attack | | | | | |
| | | common2common | | element2common | | element2element | | Factual element | | Provisional element | | fine day | | stormy day | | murder day | |
| | Acc | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR |
| Baichuan2 | 0.840 | 0.838 | 0.24% | 0.832 | 0.95% | 0.835 | 0.60% | 0.697 | 17.02% | 0.529 | 37.02% | 0.828 | 1.43% | 0.823 | 2.02% | 0.818 | 2.62% |
| ChatGLM3 | 0.777 | 0.77 | 0.90% | 0.764 | 1.67% | 0.764 | 1.67% | 0.733 | 5.66% | 0.656 | 15.57% | 0.750 | 3.47% | 0.752 | 3.22% | 0.741 | 4.63% |
| GPT3.5-turbo | 0.731 | 0.734 | -0.41% | 0.726 | 0.68% | 0.730 | 0.14% | 0.699 | 4.38% | 0.643 | 12.04% | 0.730 | 0.14% | 0.727 | 0.55% | 0.728 | 0.41% |
| LLaMA3 | 0.729 | 0.713 | 2.19% | 0.713 | 2.19% | 0.712 | 2.33% | 0.621 | 14.81% | 0.474 | 34.98% | 0.716 | 1.78% | 0.718 | 1.51% | 0.692 | 5.08% |
| Farui | 0.881 | 0.881 | 0.00% | 0.878 | 0.34% | 0.880 | 0.11% | 0.842 | 4.43% | 0.785 | 10.90% | 0.876 | 0.57% | 0.873 | 0.91% | 0.870 | 1.25% |

Table 9: CAIL 2018 Results in Conclusion Level.

| CAIL2018 | Original | Conclusion Level | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Previous Behavior Attack | | Expert Opinion Attack | | | | | | | | | |
| | | | | pupil | | layperson | | law student | | lawyer | | judge | |
| | Acc | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR | Acc | PDR |
| Baichuan2 | 0.840 | 0.760 | 9.52% | 0.721 | 14.17% | 0.735 | 12.50% | 0.672 | 20.00% | 0.617 | 26.55% | 0.618 | 26.43% |
| ChatGLM3 | 0.777 | 0.691 | 11.07% | 0.657 | 15.44% | 0.642 | 17.37% | 0.598 | 23.04% | 0.582 | 25.10% | 0.585 | 24.71% |
| GPT3.5-turbo | 0.731 | 0.697 | 4.65% | 0.631 | 13.68% | 0.626 | 14.36% | 0.600 | 17.92% | 0.623 | 14.77% | 0.566 | 22.57% |
| LLaMA3 | 0.729 | 0.448 | 38.55% | 0.461 | 36.76% | 0.469 | 35.67% | 0.456 | 37.45% | 0.447 | 38.68% | 0.458 | 37.17% |
| Farui | 0.881 | 0.836 | 5.11% | 0.792 | 10.10% | 0.793 | 9.99% | 0.753 | 14.53% | 0.698 | 20.77% | 0.681 | 22.70% |

We did multiple trials with open-source models like GLM, and the predictions were the same. For the closed-source model, multiple calls would fail to respond, but the results are similar on the smaller dataset.

Table 10: Detailed results in RAG, COT and Few-shot.

| LEVEN | Original element attack | | | | | fewshot | | | | | rag | | | | | cot | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OriAcc | Factual element | | Provisional element | | OriAcc | Factual element | | Provisional element | | OriAcc | Factual element | | Provisional element | | OriAcc | Factual element | | Provisional element | |
| | | Acc | PDR | Acc | PDR | | Acc | PDR | Acc | PDR | | Acc | PDR | Acc | PDR | | Acc | PDR | Acc | PDR |
| Baichuan2 | 0.777 | 0.681 | 12.36% | 0.534 | 31.27% | 0.221 | 0.219 | 0.90% | 0.214 | 3.17% | 0.84 | 0.808 | 3.81% | 0.707 | 15.83% | 0.763 | 0.667 | 12.58% | 0.508 | 33.42% |
| ChatGLM3 | 0.734 | 0.696 | 5.18% | 0.644 | 12.26% | 0.225 | 0.225 | 0.00% | 0.211 | 6.22% | 0.834 | 0.783 | 6.12% | 0.789 | 5.40% | 0.719 | 0.678 | 5.70% | 0.642 | 10.71% |
| LLaMA3 | 0.679 | 0.560 | 17.53% | 0.430 | 36.67% | 0.720 | 0.620 | 13.89% | 0.482 | 33.06% | 0.834 | 0.785 | 5.88% | 0.729 | 12.59% | 0.685 | 0.567 | 17.23% | 0.454 | 33.72% |
| Farui | 0.849 | 0.807 | 4.95% | 0.746 | 12.13% | 0.861 | 0.721 | 16.26% | 0.626 | 27.29% | 0.888 | 0.845 | 4.84% | 0.786 | 11.49% | 0.832 | 0.779 | 6.37% | 0.706 | 15.14% |

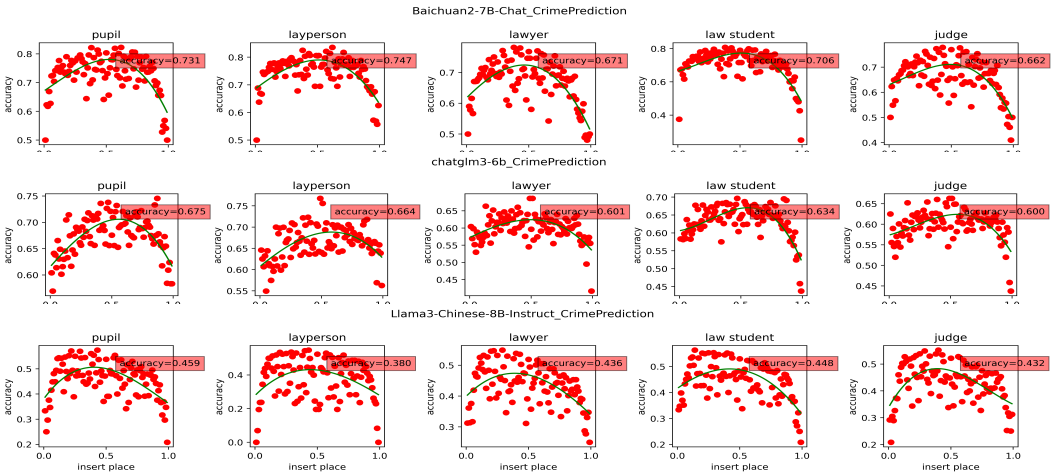


Figure 5: Model Experiments on CAIL2018 on Expert Opinion Attack

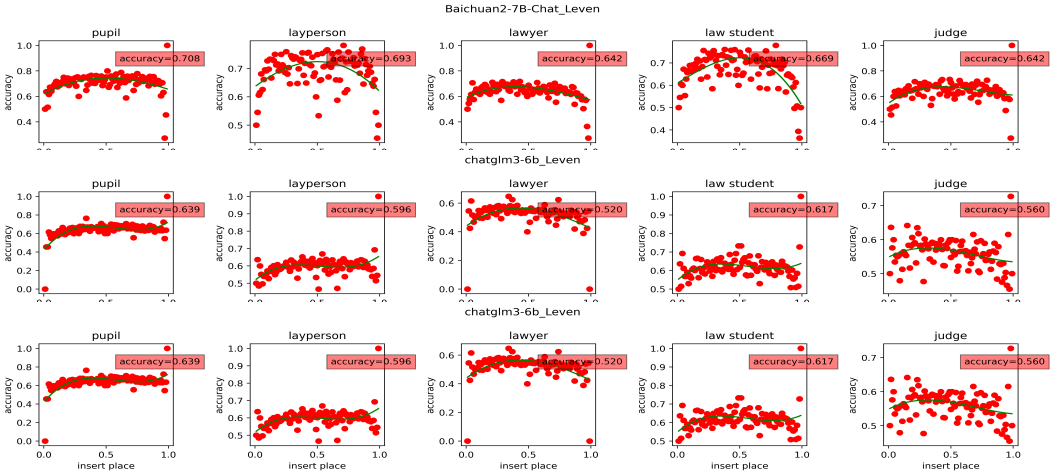


Figure 6: Model Experiments on Leven on Expert Opinion Attack

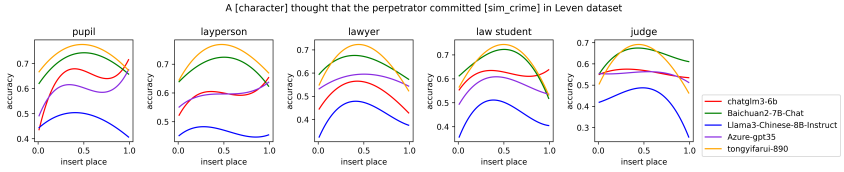


Figure 7: Location Attack on Leven.

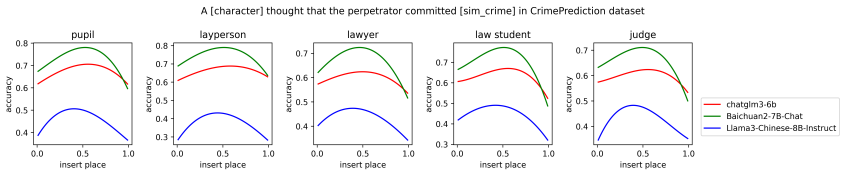


Figure 8: Location Attack on CAIL2018.

# B Annotation

## B.1 Annotation Category

We categorize the annotation into four parts: Crime level, Element level, Word level and Narration level. For each legal documents, the legal experts first annotate the similar crime (as shown in Table 12 shows), then identify the four elements based on the case fact, for each elements, they annotate the factual elements and the provisional elements from the text(as shown in Table 14) Besides, the legal experts annotate the legal synonyms from the case judgements, and generate the common synonyms from GPT-4o. All the legal experts worked together on the annotation of the Narration.

Table 11: Annotation Category.

| **Annotation Framework** | **Data** | **Data details** | **Annotation methods** |
|---|---|---|---|
| Crime | Similar crime provision | Provisions of similar crimes of one crime that may lead to confusion. | Summarize crimes from bar exam questions |
| Element | Subject of the crime | The person who commits the criminal act, including natural persons or units | Annotate crime by crime/case by case |
| | Subjective aspect of the crime | The psychological state of the criminal subject towards its criminal act and its consequences, including intent or negligence. | Annotate crime by crime/case by case |
| | Objective aspect of the crime | The specific manifestations of the criminal act, including harmful act, harmful result and causal relationship between them | Annotate crime by crime/case by case |
| | Object of the crime | The social relations protected by criminal law and violated by crimes | Annotate crime by crime/case by case |
| Word | Legal synonyms | Words with legal synonyms | Extract from Element annotations and LEVEN dataset |
| | Common synonyms | Words with common synonyms | Annotate with GPT4-o |
| Narration | Narrative sentence | It is a fine /stormy/murder day | Annotate with common sense |

### B.2 Annotation Guideline

#### B.2.1 Annotation Method

1. **Similar Crime**: summarizing from the options in judicial examination questions to identify similar crimes related to a particular crime and place them in the same row. If there is no similar crime listed in the same row of the table, then the crimes before and after in the same chapter are considered similar crimes. This dataset is used for generating options for test questions.
2. **Word/Synonyms**: need to think about which legal typographical errors may occur, such as "gun" written as "gum", which has legal ambiguity both before and after, this kind of typographical error needs to be thought out with an open mind, if there is no, do not write.
3. **Element**: Each crime is divided into two rows. The first row is the abstract elements organized according to the legal provisions (e.g., personal injury), and the second row is the concrete expressions of the elements organized according to the cases and other materials (e.g., beating).
4. **Narration**: if a crime has a special environment, time, place, means, it should be marked, or if a certain time, place, means frequently appear in the cases of the crime, such as murder at night, theft always on the bus, fighting always with sticks, robbery always with knives and guns, it is necessary to summarize this high-frequency special vocabulary, if there is no, do not summarize.

#### B.2.2 Annotation Team

We have assembled a team of 10 graduate students to form our annotation team. There are a total of 483 crimes in the Chinese Criminal Law. Three people are responsible for annotating Similar Crime Four person is responsible for annotating four Elements. In additional, Other team members tasked with listing 3-10 items for Narration, Prompt, and Word/Synonyms. Afterwards, team members have been engaged in testing the annotations, organizing revisions, and supplementing additional annotations. Each team member is expected to commit at least 15 hours per week, with a compensation rate of 45 RMB per hour, which highly exceeds the local average wage standard. In total, The duration of the annotation process has been approximately four months.

### B.3 Case Study

**Prompt**: Based on the facts of the case, determine which of the following crimes the defendant may be guilty of. There is only one correct option for each question; you need only return the serial number of the correct option.

**Fact**: According to the safety operating procedures, the workshop driver transporting high-temperature raw materials should ensure that they had been extinguished. Yang, a workshop staff member, poured more than 40 tons of high-temperature raw materials into the cooling equipment, knowing that the materials had not been extinguished. Some of the materials flowed into the conveyor belt, resulting in direct economic losses of 740,422 yuan.

**Choices**: A.Crime of causing an accident with dangerous goods B.Crime of negligently causing serious accident C.Crime of negligently arson D.Crime of arson

**Analysis**: The model affected by Crime of negligently arson retrieved through RAG or by the embedded questions about Crime of negligently arson in the prompt, wrongly predicted actions that should be judged as Crime of negligently causing serious accident as Crime of negligently arson. It failed to follow the logic of using the four elements framework of criminal law to analyze the differences between the crimes and make legal judgement Prediction. The elements for establishing crime are divided into the object of crime, the objective aspect of crime, the subject of crime and the subjective aspect of crime. Although Crime of negligently causing serious accident and Crime of negligently arson are the same in terms of object and subjective aspect, both jeopardizing public security and being negligent, there is a significant difference in terms of objective aspect of crime and object. The subject element of Crime of negligently arson is the general subject, while Crime of negligently causing serious accident is the special subject, i.e., the employees of factories, mines,

forests, construction enterprises or other enterprises or institutions. On the objective aspect element, the Crime of negligently causing serious accident refers to a serious accident that occurs in the course of production or operation in violation of specific safety management regulations, while Crime of negligently arson occurs in daily life. In the this case, since the act took place in the course of production and operation, violated specific safety management regulations and caused major economic losses, and the subject status was that of a workshop worker, the Crime of negligently causing serious accident should be applied instead of Crime of negligently arson.

### B.4 Annotation Examples

The similar crime annotation shows as below, due to issues with existing LLMs, such as hallucinating charges and a limited input window, we presented the prediction of each text's charge in the form of multiple-choice questions. To attack LLMs more accurately, we annotated similar crimes. For each crime in the criminal law code, legal experts identified similar crimes, which are crimes that have very subtle differences in the constitutive elements and logical deduction process from the original crime and are easily mistaken by legal practitioners in practice. For multiple choice questions option making, if the number of manually annotated similar crimes $k < 3$, then $3 - k$ other non-repeating crimes are randomly selected from the same chapter in the legal code as choices. If the number of available crimes in the same chapter is insufficient, non-repeating crimes are randomly selected from all 184 possible crimes as options. We shuffled one correct crimes and three similar crimes into four options and let the LLM answer in the form of a multiple-choice question:

## C Ethics Statement

Our work aims to evaluate whether Large Language Models (LLMs) have truly learned knowledge and whether they can apply logic for causal inference within a domain. This work holds significant progressive implications for society. Before such evaluation standards were established, domain experts could potentially misuse large language models, using them incorrectly for decision-making, which could decrease trust in these LLMs and lead to numerous societal issues. For instance, a doctor might misdiagnose a patient due to flawed decisions influenced by an LLM, exacerbating the patient's condition and heightening tensions between doctors and patients. Similarly, a judge might make an erroneous decision based on an LLM's advice, falsely convicting an innocent person, which would add to societal burdens. As such, the necessity and societal impact of our work are considerable.

However, we must also recognize that our work cannot cover all legal issues. Therefore, when this standard becomes the sole measure of an LLM's robustness in terms of knowledge, it could potentially be exploited. People might defend against this evaluation standard with the aim of getting their 'malicious' LLMs approved and into the market, which could cause potential harm to society.

Yet, I believe this is not a fatal issue. The evaluation methods are comprehensive, and the content of the annotations will become increasingly detailed. Encouraging more similar works to emerge will help standardize the use of LLMs and enhance their credibility.

Table 12: Examples of annotated crimes and similar crimes

| **Damage to Transportation Vehicles** | **Money Laundering** | **Kidnapping** | **Racketeering** | **Damage to Production and Business** | **Assaulting Police** | **Harboring and Sheltering** |
|---|---|---|---|---|---|---|
| Intentional Destruction of Property | [Organizing, Leading, Participating in] Criminal Organizations | Illegal Detention | Fraud | Damage to Electrical and Flammable Equipment | Obstruction of Official Duties | False Testimony |
| Explosion | Entering the Country to Develop Criminal Organizations | Robbery | Cheating and Lying | Damage to Transportation Vehicles | Endangering Public Security by Dangerous Means | Sheltering Drug Offenders |
| Arson | Sheltering, Condoning Criminal Organizations | Fraud | Robbery | Damage to Transportation Facilities | Intentional Injury | Concealing, Hiding Criminal Proceeds |
| Theft | Assisting in Terrorist Activities | Racketeering | Kidnapping | Arson | Intentional Homicide | Assisting in [Destroying, Fabricating] Evidence |
| | Preparing to Implement Terrorist Activities | | | Water Pollution | | |
| | Advocating Terrorism, Extremism, Inciting Terrorist Activities | | | Explosion | | |
| | Harboring, Sheltering Crime | | | Release of Hazardous Substances Crime | | |
| | Concealing, Hiding Criminal Proceeds, Criminal Gains | | | | | |

Table 13: Four elements annotation example.

| Crime No. | Crime Name | Object of Crime (1) | Objective Aspect - Action (2) | Objective Aspect - Result (3) | Subject (4) | Subjective Aspect (5) |
|---|---|---|---|---|---|---|
| **179** | Intentional Injury Crime | Personal Body | Assault | Injury above minor | Person | Intentional |
| | | Body, face | Beating | None | People | premeditate |
| **268** | Group Fight Crime | Public Order | Fighting, Gathering Crowd | Injury below serious | Crowd, Key instigators and active participants | Intentional |
| | | Social Peace, Safety of Citizens | Beating | None | Several People, Group | premeditate |
| **269** | Provoking Trouble Crime | Public Order | Seeking trouble, Making a scene | Injury below serious | general entity | Intentional |
| | | Social Peace, Safety of Citizens | Insulting, Beating, Forcing to take, Causing a disturbance | None | People | disregard for the law |
| **226** | Job Position Embezzlement Crime | Property of the company, enterprise where the perpetrator works | Using the convenience of the position, taking for oneself | None | People other than state functionaries | Intentional |
| | | Money, Company funds | Occupying | None | company employee, staff | Concealment of Truth |
| **402** | Corruption Crime | Public property | Using the convenience of the position, embezzling, stealing, defrauding or by other means | None | state functionary | Intentional |
| | | Subsidies, Confiscated money | Embezzling, Stealing, Defrauding | None | Director, Leader | Greed |

Table 14: Annotation of provisional elements and factual elements.

| **Provisional Element** | **Explosion Incident** | **Accident** | **Poisoning** | **Verbal Conflict** | **Employment** |
|---|---|---|---|---|---|
| **Factual Element** | Accident | Accident | Drug Poisoning | Altercation | Appointment |
| | Blast | Casualty Accident | Food Poisoning | Argument | Commission |
| | Blow Up | Cave-in | Poisoning | Conflict | Employee |
| | Bomb | Collapse | Poisoning Incident | Debate | Employees |
| | Burst | Collision | Toxic | Discord | Employer |
| | Crater | Crash | Toxicity | Dispute | Employing Workers |
| | Deflagration | Drop | Vomiting | Dispute Occurred | Employment |
| | Detonate | Electrocution | | Fight | Hiring |
| | Dynamite/Explosive | Fall | | Mutual Abusive Language | Job Application |
| | Explosion | Landslide | | Quarrel | Labor |