J&H: Evaluating the Robustness of Large Language Models Under Knowledge-Injection Attacks in Legal Domain

Yiran Hu[1,2*], Huanghai Liu[1*], Qingjing Chen[1], Ning Zheng[1], Chong Wang[1], Yun Liu[1], Charles L.A. Clarke[2†], Weixing Shen[1‡]

[1]School of Law, Tsinghua University
[2]David R. Cheriton School of Computer Science, University of Waterloo

Abstract

As the scale and capabilities of Large Language Models (LLMs) increase, their applications in knowledge-intensive fields such as legal domain have garnered widespread attention. However, it remains doubtful whether these LLMs make judgments based on domain knowledge for reasoning. If LLMs base their judgments solely on specific words or patterns, rather than on the underlying logic of the language, the "LLM-as-judges" paradigm poses substantial risks in the real-world applications. To address this question, we propose a method of legal knowledge injection attacks for robustness testing, thereby inferring whether LLMs have learned legal knowledge and reasoning logic. In this paper, we propose J&H: an evaluation framework for detecting the robustness of LLMs under knowledge injection attacks in the legal domain. The aim of the framework is to explore whether LLMs perform deductive reasoning when accomplishing legal tasks. To further this aim, we have attacked each part of the reasoning logic underlying these tasks (major premise, minor premise, and conclusion generation). We have collected mistakes that legal experts might make in judicial decisions in the real world, such as typos, legal synonyms, inaccurate external legal statutes retrieval. However, in real legal practice, legal experts tend to overlook these mistakes and make judgments based on logic. However, when faced with these errors, LLMs are likely to be misled by typographical errors and may not utilize logic in their judgments. 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.

Code — https://github.com/THUlawtech/LegalAttack

Introduction

Large Language Models (LLMs) are increasingly applied to knowledge-intensive fields, such as law (Cui et al. 2024; Huang et al. 2023) and medicine (Tu et al. 2023; Yang et al. 2024). In these fields, LLM agents often act as domain experts (Mei et al. 2024), which need to rely on comprehensive domain knowledge and logical reasoning to complete domain-specific tasks (Miao, Teh, and Rainforth 2023). However, the reliability and robustness of LLMs in performing these domain-specific tasks have not been fully 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 (Zhu et al. 2023; Wang et al. 2023) 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(Zhou et al. 2024). Unlike in the general domain(Xu et al. 2023; Ni et al. 2023), 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 judgment. Unlike the existing reasoning work in math (Zhou et al. 2024) and chemistry (Ouyang et al. 2024), our work focuses on reasoning which is both natural language based and requires logical reasoning in natural lanuage. Understanding human language and the logic behind it is more complex than merely learning numbers and operation symbols.

Can we trust LLMs in the legal domain? LLMs are fragile and small perturbations in the prompt can have a significant impact on their performance. Especially in the knowledge-intensive domains, domain experts will automatically ignore those small errors and changes, making judgments based on 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?(Chen et al. 2023).

Based on these considerations, this paper proposes a knowledge attack framework **J&H**[1] directed at knowledge-

*These authors contributed equally.
†Corresponding Author. charles.clarke@uwaterloo.ca
‡Corresponding Author. wxshen@mail.tsinghua.edu.cn

[1]J&H: Originally taken from an old novel: Jekyll&Hyde. 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 knowledge attacks.

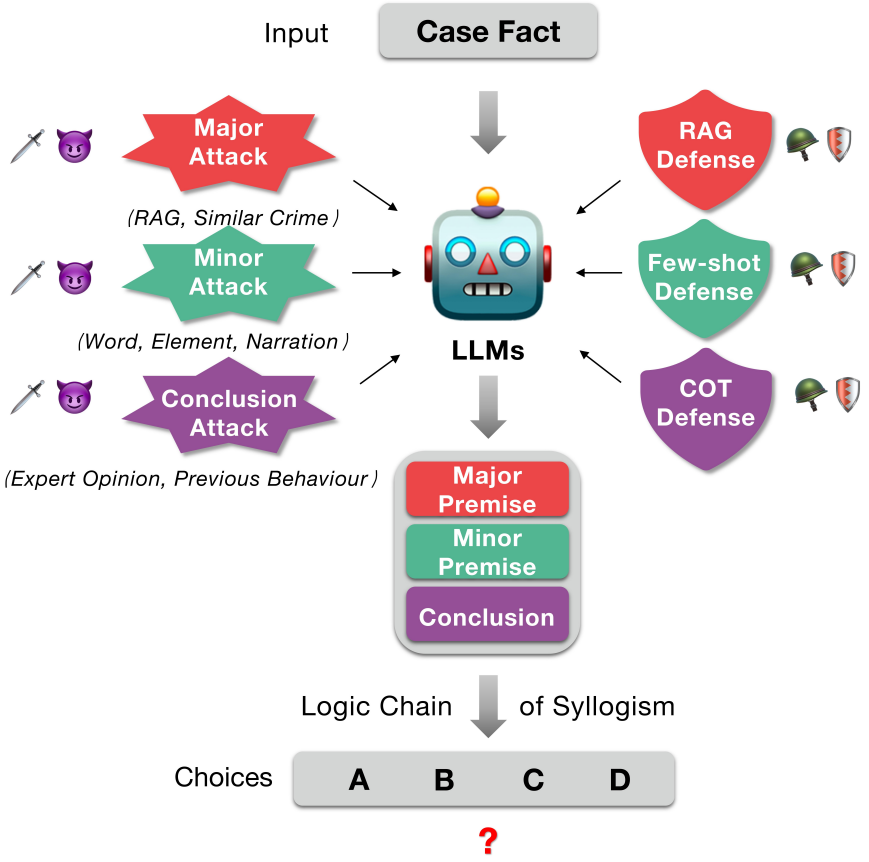

Figure 1: The Framework of J&H.

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", and "conclusion generation"**. The framework of J&H is shown 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(An et al. 2022; Xue et al. 2023). We have conducted fine-grained annotation J&H 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 reasoning and making 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 at tasks that require logical 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, perturbation has a substantial impact on the final judgment. We conducted additional position attacks by inserting noise in the conclusion stage. Results show that inserting noise into the middle part of the prompt will minimally affect the attack effect on the model. In this case, the noise is "lost in the middle" (Liu et al. 2023)'

Based on the outcome of our attack experiments, this paper proposes three methods to improve the performance of LLMs under knowledge injection attacks: RAG, COT,

and few-shot. However, our experiments show that these 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, improvement should be targeted towards the pre-training or fine-tuning process of LLMs.

This paper makes the following contributions:

1. **We propose J&H: an evaluation framework for evaluating the robustness of LLMs under legal knowledge injection attacks.** In the framework, we use syllogism as the theoretical basis and carry out knowledge attacks on each layer separately. We conducted fine-grained annotation in the legal field. 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.
2. **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; LLMs cannot use domain knowledge to make correct judgments under the framework of reasoning logic.
3. **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. This outcome shows that merely through prompting we may not be able to 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.

Methodology

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 equation 1 shows, $\mathcal{A}$ is the major premise, $\mathcal{B}$ is the minor premise and $\mathcal{C}$ is the conclusion.

$$\mathcal{A} \Rightarrow \mathcal{B}, \mathcal{B} \Rightarrow \mathcal{C} \vdash \mathcal{A} \Rightarrow \mathcal{C} \quad (1)$$

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 legal tasks, we attack each level of the syllogistic reasoning process in the legal judgment inference. As shown in Figures 2, 3, and 4, J&H has three level attacks. Different levels have different attack methods based on the facts of the case, along with four choices for the conclusion. The choices are generated based on similar crimes related to the correct crime, where similar crimes are those crimes that are easily confused by legal professionals. To create a list of these similar crimes, we invited ten law school graduate students to annotate the cases. In the attack methods of our framework, each choice represents a similar crime. 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.

Attack at the Major Premise Level The legal provision plays the role of a major premise in the logical deduction of legal tasks. When legal practitioners solve practical legal problems, they first retrieve the most relevant articles from laws and regulations. These articles also serve as the premise and foundation for all reasoning. In practice, judges would discover through comparison that these facts cannot be applied to the major premise; they 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 judgment of the conclusion. In the J&H major premise level attack, we insert incorrect major premises as references into the facts to evaluate whether the LLMs can be affected by the incorrect premise.

As Figure 2 shows, we consider two attacks at the major premise level, through the insertion of legal articles and the names of 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 inferring major premises.

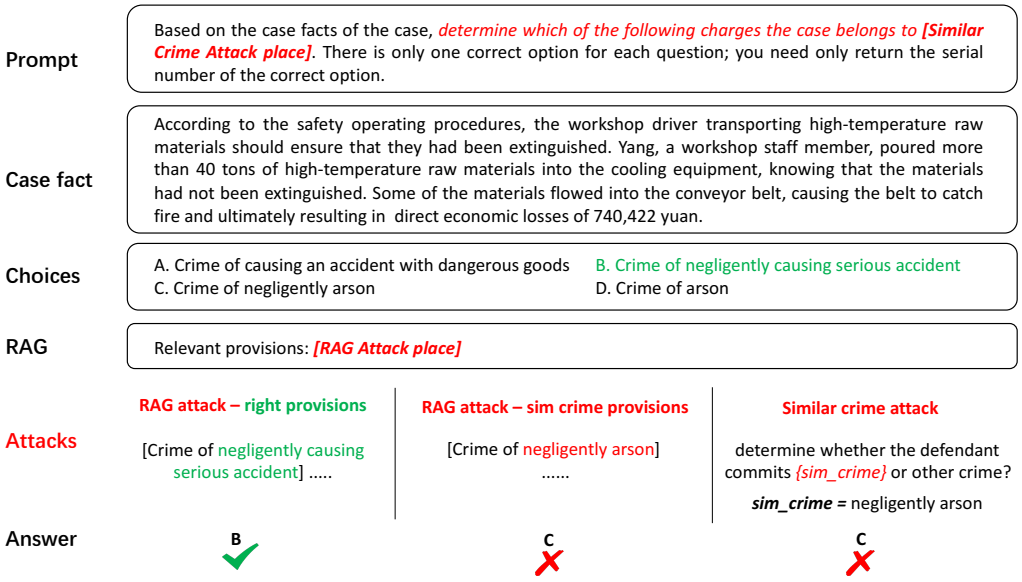

Figure 2: Illustration of Major Premise Attack.

Attack at the Minor Premise Level As shown in the Table 1, we consider three types of attacks at the minor premise level: **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 facts of the case 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 adversarial elements from the similar crime at the end of the case facts. The similar elements were divided into factual elements summarized in the facts of the case and provisional elements summarized in the law provisions.
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".

In criminal trials, judges usually make judgments on crimes based on reasoning about four elements. Analyzing from the constituent elements, every crime has the four elements: 1) **the subject of the crime**, which refers to the person who commits the criminal act; 2) **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; 3) **the objective aspect of the crime**, which refers to the specific manifestation of the criminal act; and 4) **the object of the crime**, which refers to the social relationship that is protected by criminal law and violated by the criminal act.

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 supplementary materials.

Attack at the 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 types: "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 example of Figure 4, 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 criminal judgment. The LLMs should not let

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: "... 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 from the case facts of the corresponding crime	the end of Case fact
		provisional element	Insert four elements summarized from the law of 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

Table 1: Attack methods and details at different levels of J&H.

the logical derivation of the perpetrator's current facts be misled by other logical chains. For example, in the case shown in Figure 4, 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.

Experiments

Datasets

We adopt two legal datasets for our experiments:

- **LEVEN(Yao et al. 2022)** 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(Xiao et al. 2018)** 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 Table 1. Each question consisted of one correct crime and three similar crimes, where similar crimes were selected based on annotations from legal experts. All four choices were randomly shuffled.

Dataset	Size	Charges	Avg length	Max length
CAIL2018	15806	184	419	39586
LEVEN	3323	61	529	2476

Table 2: Dataset distribution.

Setup

We examine the robustness of LLMs in domain-specific tasks on four general LLMs: Azure GPT3.5-turbo(OpenAI 2023) , Baichuan2-7b-chat(Xiao et al. 2024), ChatGLM3-6b(Zeng et al. 2022), LLaMA3(Meta 2024) fine-tuned in Chinese) and one legal-specific LLM: Farui(Aliyun 2024).

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.

Evaluation Metrics

Following the approach of Promptbench(Zhu et al. 2023), 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 the accuracy without attack.

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

Attack Accuracy. Attack Accuracy indicates the accuracy after attack.

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

Performance Drop Ratio(PDR)

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

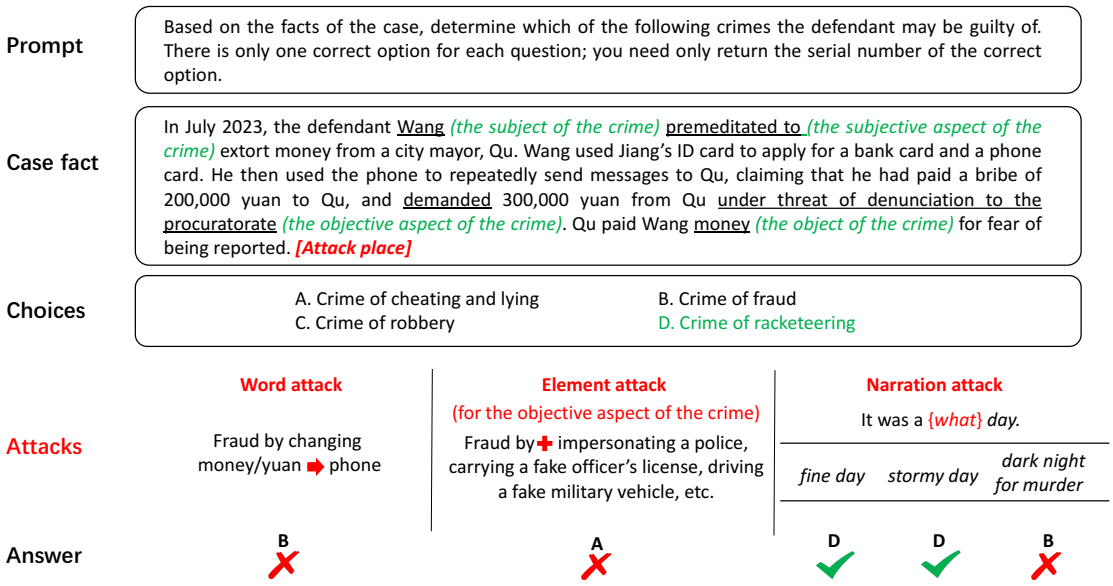

Figure 3: Illustration of Minor Premise Attack.

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	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%

Table 3: Result of attacks at the Major Premise Level.

LEVEN	Ori	Minor Premise Level															
		Word Attack						Element Attack				Narration Attack					
		com2com		ele2com		ele2ele		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	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%	0.670	1.33%	0.613	9.72%	0.560	17.53%	0.430	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%

Table 4: Result of attacks at the Minor Premise Level. 'Ori' means the Original results.

also serves as

$$PDR = 1 - \frac{AttackAcc}{OriginalAcc} \quad (5)$$

Results and Analytics

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 supplementary materials. Experimental results on the Leven dataset can be found in Tables 3, 4 and 5.

Experimental 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 exceed 30%. This result suggests 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 at the Minor

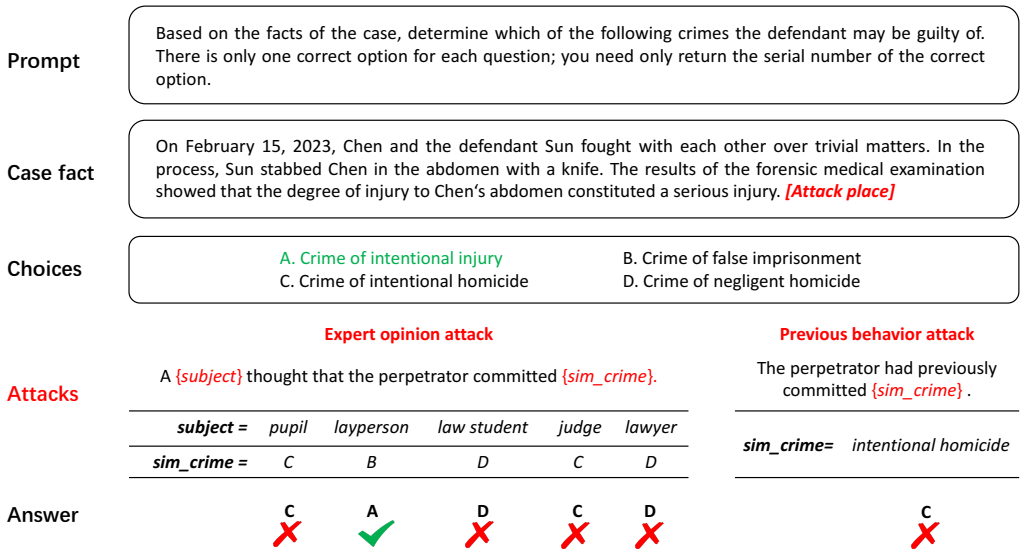

Figure 4: Illustration of Conclusion Attack.

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 suggests 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 LLMs. 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 is still not fully robust to our attacks, indicating that LLMs may need to incorporate more 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 weak in logical reasoning when handling domain tasks, and their generated conclusions are easily disrupted by conclusion-level adversarial attacks.

Location Attack

Given the success of attacks on the conclusion level, we further explored the impact of the attack location (Li et al. 2023b). We conducted location attacks on the expert opinion part of the Conclusion Level 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 Figure 5, 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.

Discussion

In this section, we propose three methods to enhance the robustness of the LLMs. Results are shown in Table 6.

RAG(Lewis et al. 2020). 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 fully solve this problem.

Chain of thought (COT)(Wei et al. 2022). 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 LLMs do not 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.

Related-Work

General Domain Evaluation

Existing work(Zhu et al. 2023; Wang et al. 2023; Li et al. 2019, 2020; Morris et al. 2020; Nie et al. 2020; Wang et al. 2022) has made substantial progress on the evaluation of

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	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 5: Result of attacks at the Conclusion Level.

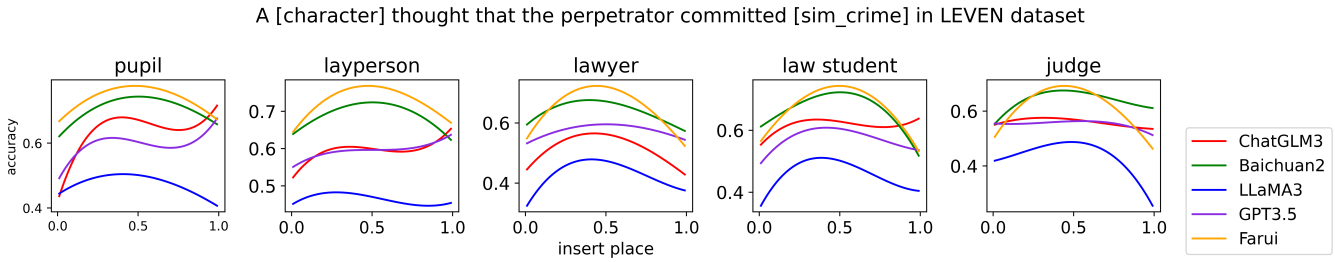

Figure 5: Location Attack on the LEVEN dataset.

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%

Table 6: PDR of Element Attack in RAG, COT and Few-shot. After enhancements, the attack is still effective (PDR $>$ 0), but the model is more robust compared to the Original scenario (PDR $<$ Original PDR).

LLMs. AdvGLUE(Wang et al. 2022), DecodingTrust(Wang et al. 2022), PromptBench(Zhu et al. 2023) undertake comprehensive benchmarks for evaluating the robustness of LLMs. 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.

Domain-Specific Evaluation

Previous work (Li et al. 2023a; Quan and Liu 2024; Su et al. 2024) has demonstrated that LLMs can be used for the domain-specific tasks, but whether they are reliable when making domain judgments remains unclear. Unlike previous work on domain-specific attacks, such as MathAttack(Zhou et al. 2024), ChemistryReasoning(Ouyang et al. 2024), our work depends on the logic underlying language, which can be more complex than numbers and symbols, and it can be widely applied in more knowledge-intensive fields.

Conclusion

In this paper, we propose a framework of legal knowledge injection attacks for robustness testing for LLMs. We use each part of the deductive reasoning logic to evaluate the models. We evaluate the general-domain and legal-domain LLMs based on the framework. The results show the fragility of prompting LLMs. We also explore several methods to alleviate the issue. We use RAG, COT and Few-shot methods, but the problem still cannot be fully solved. Our experiments show that it is not possible to effectively alleviate the success rate of LLMs being attacked by domain knowledge from the perspective of prompts. Especially in legal tasks, existing LLMs are not reliable and are fragile with respect to prompting. These 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 LLMs can be reliable under domain knowledge attacks. J&H is also available for others working on exploring robust reasoning by LLMs. Researchers can utilize this framework to evaluate the robustness of more LLMs, or apply this framework to more fields related to social life, such as the medical and educational domains.

References

Aliyun. 2024. https://tongyi.aliyun.com/farui/chat.

An, Z.; Huang, Q.; Jiang, C.; Feng, Y.; and Zhao, D. 2022. Do Charge Prediction Models Learn Legal Theory? arXiv:2210.17108.

Chen, H.; Zhang, L.; Liu, Y.; Chen, F.; and Yu, Y. 2023. Knowledge is Power: Understanding Causality Makes Legal judgment Prediction Models More Generalizable and Robust. arXiv:2211.03046.

Cui, J.; Ning, M.; Li, Z.; Chen, B.; Yan, Y.; Li, H.; Ling, B.; Tian, Y.; and Yuan, L. 2024. Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model. arXiv:2306.16092.

Huang, Q.; Tao, M.; Zhang, C.; An, Z.; Jiang, C.; Chen, Z.; Wu, Z.; and Feng, Y. 2023. Lawyer LLaMA Technical Report. arXiv:2305.15062.

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

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. In *Proceedings 2019 Network and Distributed System Security Symposium*, NDSS 2019. Internet Society.

Li, L.; Ma, R.; Guo, Q.; Xue, X.; and Qiu, X. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. arXiv:2004.09984.

Li, Q.; Hu, Y.; Yao, F.; Xiao, C.; Liu, Z.; Sun, M.; and Shen, W. 2023a. MUSER: A Multi-View Similar Case Retrieval Dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 5336–5340.

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

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.

Mei, K.; Li, Z.; Xu, S.; Ye, R.; Ge, Y.; and Zhang, Y. 2024. LLM Agent Operating System. *arXiv preprint arXiv:2403.16971*.

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

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

Morris, J. X.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. arXiv:2005.05909.

Ni, Y.; Jiang, S.; Shen, H.; Zhou, Y.; et al. 2023. Evaluating the Robustness to Instructions of Large Language Models. *arXiv preprint arXiv:2308.14306*.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. arXiv:1910.14599.

OpenAI. 2023. https://openai.com.

Ouyang, S.; Zhang, Z.; Yan, B.; Liu, X.; Choi, Y.; Han, J.; and Qin, L. 2024. Structured Chemistry Reasoning with Large Language Models. arXiv:2311.09656.

Quan, Y.; and Liu, Z. 2024. EconLogicQA: A Question-Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning. arXiv:2405.07938.

Su, W.; Hu, Y.; Xie, A.; Ai, Q.; Que, Z.; Zheng, N.; Liu, Y.; Shen, W.; and Liu, Y. 2024. STARD: A Chinese Statute Retrieval Dataset with Real Queries Issued by Non-professionals. arXiv:2406.15313.

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

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

Wang, B.; Xu, C.; Wang, S.; Gan, Z.; Cheng, Y.; Gao, J.; Awadallah, A. H.; and Li, B. 2022. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. arXiv:2111.02840.

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

Xiao, C.; Zhong, H.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; and Xu, J. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. arXiv:1807.02478.

Xiao, J.; Chen, Y.; Ou, Y.; Yu, H.; Shu, K.; and Xiao, Y. 2024. Baichuan2-Sum: Instruction Finetune Baichuan2-7B Model for Dialogue Summarization. arXiv:2401.15496.

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

Xue, Z.; Liu, H.; Hu, Y.; Kong, K.; Wang, C.; Liu, Y.; and Shen, W. 2023. LEEC: A Legal Element Extraction Dataset with an Extensive Domain-Specific Label System. *arXiv preprint arXiv:2310.01271*.

Yang, S.; Zhao, H.; Zhu, S.; Zhou, G.; Xu, H.; Jia, Y.; and Zan, H. 2024. 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, 19368–19376.

Yao, F.; Xiao, C.; Wang, X.; Liu, Z.; Hou, L.; Tu, C.; Li, J.; Liu, Y.; Shen, W.; and Sun, M. 2022. LEVEN: A Large-Scale Chinese Legal Event Detection Dataset. In *Findings of ACL*, 183–201.

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

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

Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; Zhang, Y.; Gong, N. Z.; and Xie, X. 2023. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. arXiv:2306.04528.