LJPCHECK: Functional Tests for Legal Judgment Prediction

Yuan Zhang[1*], Wanhong Huang[1*], Yi Feng[1†], Chuanyi Li[1]
Zhiwei Fei[1], Jidong Ge[1‡], Bin Luo[1], Vincent Ng[2]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Human Language Technology Research Institute, University of Texas at Dallas, USA
zyjwc@nju.edu.cn, hwh@smail.nju.edu.cn, {fy, lcy}@nju.edu.cn

Abstract

Legal Judgment Prediction (LJP) refers to the task of automatically predicting judgment results (e.g., charges, law articles and term of penalty) given the fact description of cases. While SOTA models have achieved high accuracy and F1 scores on public datasets, existing datasets fail to evaluate specific aspects of these models (e.g., legal fairness, which significantly impact their applications in real scenarios). Inspired by functional testing in software engineering, we introduce LJPCHECK, a suite of functional tests for LJP models, to comprehend LJP models' behaviors and offer diagnostic insights. We illustrate the utility of LJPCHECK on five SOTA LJP models. Extensive experiments reveal vulnerabilities in these models, prompting an in-depth discussion into the underlying reasons of their shortcomings.

1 Introduction

Legal Judgment Prediction (LJP) aims to automatically predict judgment results given the fact description of a case. While English LJP focuses on violation prediction (Chalkidis et al., 2019; Feng et al., 2022a; Chalkidis et al., 2021), Chinese LJP focuses on charge prediction, law article prediction and term of penalty prediction (Feng et al., 2022b; Zhong et al., 2018). There also exists LJP work in other languages, such as French (Sulea et al., 2017) and Greman (Niklaus et al., 2021). Figure 1 illustrates an example of the three subtasks involved in Chinese LJP. Such LJP systems can assist legal professionals in case processing as well as offering low-cost consulting services to nonprofessionals.

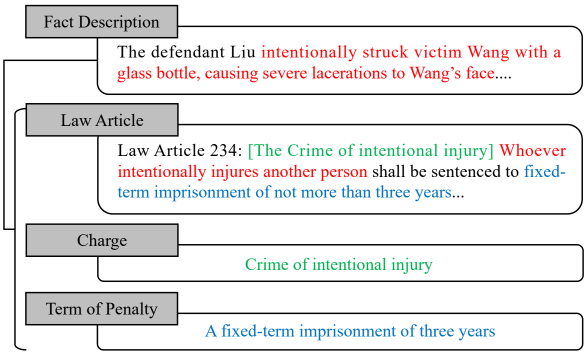
Fact Description
The defendant Liu intentionally struck victim Wang with a glass bottle, causing severe lacerations to Wang's face....
Law Article
Law Article 234: [The Crime of intentional injury] Whoever intentionally injures another person shall be sentenced to fixed-term imprisonment of not more than three years...
Charge
Crime of intentional injury
Term of Penalty
A fixed-term imprisonment of three years

Figure 1: An illustration of the LJP task.

How should the quality of LJP models be evaluated? Currently LJP models are evaluated using generic metrics such as F1 and accuracy on specific LJP subtasks such as charge prediction. However, a simple one-metric evaluation is far from sufficient for LJP since it does not provide an in-depth understanding of how a given model behaves, its weaknesses, and how to improve its performance in general. For instance, judicial fairness, which highlights that the parties involved in a case should be judged based on the case facts rather than demographic information such as gender and age, is one of the crucial considerations in the legal domain that cannot be evaluated using generic metrics. The reason is that a model that achieves a poor accuracy on, for instance, charge prediction can be fair as long as it does not make predictions based on race and gender, whereas a model that performs well on charge prediction can be unfair if it exploits race or gender information during its prediction process. While the use of one-metric evaluation is not uncommon in NLP and the shortcomings of such kind of evaluations are applicable to all NLP tasks, what is characteristic about LJP is its *human-centered* nature: understanding why a particular AI tool decided to charge someone with imprisonment is crucial for any real-world applications.

Motivated in part by the above observation, we propose a new approach to the analysis of LJP models that allows for targeted testing of specific aspects of LJP models, such as fairness. Our approach is inspired by functional testing in soft-

[*] Equal contribution.
[†] Corresponding author.
[‡] Co-corresponding author.

Findings of the Association for Computational Linguistics ACL 2024, pages 5878–5894
August 11-16, 2024 ©2024 Association for Computational Linguistics

ware engineering (Jin, 2009), which is a testing framework that assesses different functionalities of a given model by validating its output on sets of targeted test cases. Specifically, we introduce LJPCHECK, a suite of functional tests for LJP models that are sufficiently complex to allow better comprehension of a model's behavior and offer diagnostic insights.

The contributions of our work are four-fold:

A new approach to the analysis of LJP models. To our knowledge, we are the first to employ functional testing in software engineering to augment the evaluation of LJP models to include specific aspects, such as fairness.

An easily extensible methodology. While our test suite is developed for evaluating Chinese LJP models (i.e., some of the test cases are developed with specific Chinese law articles in mind), the same methodology, which we borrow from Ribeiro et al. (2020), can naturally be applied to the development of test suites for other jurisdictions. More broadly, this methodology is task-agnostic and can therefore be applied to other NLP tasks.

A new test suite. Our test suite[1], which spans over 10,800 examples, provides a valuable resource for future research into the LJP direction.

An analysis of SOTA LJP models. While the vast majority of recent work on LJP has focused on number crunching, where the goal is to improve SOTA results, we focus on understanding the strengths and weaknesses of SOTA LJP models. We believe that our work is also valuable for the broader "analysis-focused research" because it shows how to apply the methodology of Ribeiro et al. (2020) to a new domain.

2 Related Work

Much recent work on LJP has focused on applying deep learning and proposing different kinds of neural networks to solve the subtasks in LJP, with the primary goal of improving performance numbers. For instance, some researchers (Yue et al., 2021; Feng et al., 2022b) extract key elements from facts to improve both accuracy and interpretability, while others employ multi-task learning (Zhong et al., 2018) to simultaneously predict multiple judgments. With the development of Pretrained Language Models (PLMs), legal PLMs like Lawformer (Xiao et al., 2021) pretrained on legal-related data is used to solve different legal tasks (including LJP) in a unified framework.

Nevertheless, given the significant impact of legal decisions on individuals, the evaluation of AI systems in the legal domain is getting attention (Chalkidis et al., 2023). Gumusel et al. (2022) train a Word2Vec model on a US case law dataset and discover that NLP methods make undesirable distinctions between legally equivalent entities that vary only by race, which leads to race bias. Chalkidis et al. (2022) propose a multi-lingual benchmark suite, based on four legal datasets across four jurisdictions, to test model biases in attributes like demographics and court regions. An et al. (2022) design a framework to evaluate whether existing charge prediction models are trustworthy with 3 proposed principles that they believe models should follow. Wang et al. (2021) propose a metric to measure fairness in the legal domain, highlighting that both regional and gender biases exist in the legal systems. In this paper, we provide an evaluation framework tailored to LJP systems. Unlike the above works which only focus on one or a few aspects, we give broader insights with respect to 23 different functionalities. Further, our test framework can be applied to other tasks whereas the above works are tied to their own tasks.

3 LJPCHECK

Functional testing in software engineering refers to the practice of testing specific functionalities of software applications by providing sample inputs and verifying whether the outputs are as expected. While prediction accuracy is a key consideration to LJP models, there are other aspects that are essential but overlooked when evaluating models. We define these considerable aspects as *functionalities* in this paper. For model-agnostic testing, we apply the principle of *decoupling testing from implementation* by treating models as black boxes, which allows different models to be tested without knowing their internal structures. Next, we detail the functionalities to be tested (e.g. *Time Fairness*), the test types used (e.g. *Minimum Functionality Test*), and how we generate tests.

3.1 Selecting Functionalities

To generate the list of 23 functionalities, we review LJP works and interview legal practitioners (judges and lawyers) about their expectations on LJP.

[1] The test suite is available at here (link).

Functionality	**Description**	**Test Type**	**# of tests**
F1: Exemption	exemption from penalty	*RCT*	200
F2: Name Fairness	treated equally regardless of name	*INV*: replace name	500
F3: Gender Fairness	treated equally regardless of gender	*INV*: replace gender	500
F4: Ethnicity Fairness	treated equally regardless of ethnicity	*INV*: replace ethnicity	500
F5: Education Fairness	treated equally regardless of education degree	*INV*: replace/add education degree	500
F6: Gender & Ethnicity & Education Fairness	treated equally regardless of gender, ethnicity and education degree	*INV*: replace/add gender & ethnicity & education degree	500
F7: Time Fairness	treated equally regardless of time	*INV*: remove/add time	500
F8: Location Fairness	treated equally regardless of location	*INV*: remove/add location	500
F9: Time & Loc. Fairness	treated equally regardless of time & location	*INV*: remove/add time & location	500
F10: Multi-charge	predicting cases with multiple charges	*RCT*	200
F11: Misdemeanor	reliable performance on misdemeanor cases	*RCT*	550
F12: Felony	reliable performance on felony cases	*RCT*	550
F13-F18: Key Factors	following the legal logic to identify key factors relevant to charges and penalty ranges	*MFT*: templates	6*500
F19: Voluntary Surrender	surrendering leads to lenient penalty	*DIR*: add/remove surrender facts	300
F20: Criminal Attempt	attempted crimes mitigate penalty	*DIR*: add/remove attempt facts	500
F21: Forgiveness	victim's forgiveness mitigate penalty	*DIR*: add/remove forgiveness facts	500
F22: Mental Illness	being mentally ill receive mitigated penalty	*DIR*: add/remove mental illness facts	500
F23: Recidivism	recidivists are subject to severe penalty	*DIR*: add/remove recidivism facts	500

Table 1: Functionality descriptions, test types and the number of test caess (Appendix B shows detailed description).

Literature review We identify challenges and weaknesses that have been mentioned in existing LJP research papers, surveys and empirical studies published in the premier publication venues in NLP, legal-tech and artificial intelligence communities (e.g. ACL, EMNLP, NAACL, COLING, IJCAI, SIGIR, AAAI and ICAIL) from 2017 to 2023. For instance, Wang et al. (2021) introduce the gender bias challenge in their paper specifying that gender differences lead to different prediction results. We thus define the functionality related to gender fairness. Papers on multiple languages (e.g., Chinese, English and French) are included when reviewing. While jurisdiction differs across languages, we aim at obtaining a broader view about the functionalities related to LJP. We also review the papers about legal theories (e.g., Wachter et al. (2021)) to include functionalities from the perspective of legal theory researchers. For instance, we include *Discretionary Sentencing Factor Functionalities* (e.g., surrendering voluntarily results in a reduced penalty) since they are principles established within the Chinese legal system.

Practitioner Interview By interviewing practitioners we aim to (1) verify our understanding from the literature review and (2) obtain practitioners' opinions and expectations on LJP.

We sent invitations to legal professionals who have used LJP systems. In the end, seven agreed to join our interviews, including one judge, two corporate legal affairs, two lawyers and two law school students with experience years varying from 1 to 7. We also extended invitations to 3 non-professionals who have experience in legal disputes or have sought consultation through LJP systems. Interviews are conducted face-to-face or over the phone for 60-80 minutes. Each interview is composed of 3 parts: 1) collecting demographic information and work experience of the interviewees; 2) asking open questions about the challenges and weaknesses they met when using LJP systems (e.g., What errors occurred when you used LJP tools? What challenges do you think exist when applying LJP systems in real-world scenarios?); and 3) discussing their expectations on LJP (e.g., What capabilities should LJP systems possess?).

After the interviews were done, we transformed the interview findings into functionalities. For instance, given that interviewees mentioned the challenge of complex cases involving multiple charges, we define the multi-charge functionality [1].

Selecting functionalities Next, we select the functionalities from our initial list following two criteria as not all functionalities are applicable. First, we set the constraint within the scope of the Chinese legal system as we focus on testing Chinese LJP models. However, certain functionalities such as fairness are also applicable to other jurisdictions. Then, we include functionalities for which we can construct test cases with gold labels. For instance, interpretabilities are significant in the legal domain and are raised by several interviewees.

[1] Due to privacy concerns, we will release the interview records after getting the participants' consent.

In the field of legal AI, model interpretability is typically demonstrated in two mechanisms: a) generating or extracting explicit "explanations" and b) aligning models with human legal reasoning and theories. Regarding the first approach, since most LJP models do not directly generate explanations and a "gold-standard" explanation cannot be determined, we leave out this aspect as it is challenging to design black-box test cases for it.

3.2 Functionalities of LJP Models

We ultimately identify 23 fine-grained functionalities (See Table 1), which can be categorized into four classes, as described below. All functionalities stem from the standpoint of law, embodying legal principles that must be adhered to.

Exemption from Criminal Penalty According to Article 37 of the Criminal Law of the People's Republic of China, *If the severity of crime is considered minor and does not meet the threshold for penalty, criminals may be exempted from penalty*. In this case, a criminal may be convicted of a crime without being sentenced to any of the following criminal penalties: *surveillance*, *detention*, *fixed-term imprisonment*, *life imprisonment* and *death penalty*. For instance, criminals who are accomplices may be exempted from penalty as accomplices are subject to a reduced severity of crime. Unlike the presumption of innocence discussed in An et al. (2022), exemption from criminal penalties indicates a guilty verdict without penalty, whereas the presumption of innocence indicates that no legal crimes were committed. Thus, LJP models should possess the capability to predict non-penalties for cases with charge conviction. (**F1**).

Legal Fairness Legal fairness necessitates that *Everyone is treated equally before the law*. Thus, we propose a set of fairness functionalities, which means that individuals, regardless of their race, gender or other features unrelated to the law, should receive equal treatment. Specifically, we test the following features: (1) demographic features, including name, gender, ethnicity, education degree and their combination (**F2-F6**). and (2) spatio-temporal features, including the crime time, the crime location and their combination (**F7-F9**).

Complex Case Handling According to legal practitioners, the practicality of an LJP model is significantly impacted by its ability to handle complex cases. One kind of complexity concerns multi-charge cases, i.e., the defendant may violate multiple charges simultaneously. LJP systems should accurately predict multiple charges. Thus, we include the multi-charge functionality (**F10**). Another kind of complexity concerns the differences between felony cases and misdemeanor cases as felony cases typically involve more complex details than misdemeanor cases. Furthermore, felony cases involve more severe penalties (maybe over 10 years of imprisonment). We are more concerned with the performance of LJP systems in felony cases than in misdemeanor cases as wrong judgment predictions for felony cases can lead to more significant losses for parties involved. Thus, we define the misdemeanor case functionality (**F11**) and the felony case functionality (**F12**).

Key Element Recognition One significant judgment criterion in the Chinese legal system is to follow the legal logic that defined in the law instead of individuals' experience. Thus, LJP systems should follow the legal logic instead of unverified experience learned from datasets. Further, it would offer some interpretability in a round-way if we can verify that LJP systems follow the legal logic, as discussed in 3.1.

The legal logic in China involves identifying key elements. Each law article stipulates specific key elements. So, when a case contains these key elements, it triggers the corresponding law articles and charges, determining the applicable penalties. LJP systems should follow this paradigm, i.e., first identifying the key elements and then predicting the corresponding judgments. Key elements can be categorized into two classes. *Key Factors* refer to the explicitly defined elements of a crime in the law. For instance, Article 264 of the Criminal Law of the People's Republic of China stipulates that "*Stealing public or private property constitutes the crime of theft. If the value of the stolen property is fairly large, the offender may be sentenced to a fixed-term imprisonment of not more than three years. If the value of the stolen property is large, the offender may be sentenced to a fixed-term imprisonment of not less than three years but not more than ten years*". Here, *stealing* triggers the crime of theft, and *the value of the stole property* measures the final penalty. Both of them are *Key Factors*, which typically establish charges and the range of penalties. In contrast, *Discretionary Sentencing Factors* refer to flexible considerations on penalty besides *Key Factors*. *Discretionary Sentencing Fac-*

Test case	Expected	Predicted	Pass?
Example 1 Testing **Key Factor of Theft** with *Min Func Test*	Charge | Term of penalty (mon.)		
Template: "Defendant {Name} sneaked into a house and {Trigger} {Item}. The stolen items were worth {Value}"			
Test Case 1: Defendant Wang sneaked into a house and stole wallets. The stolen items were worth 2,600 RMB.	Theft | 0-36	Theft | 39	×
Example 2 Testing **Voluntary Surrender** with ***DIR**ectional*	Term of penalty decreasing (↓) after additions		
Test Case 2: Defendant Lu was driving a vehicle without a license and hit Xu...After the accident, Lu called the police and surrender oneself to justice.	↓	12 → 20	×
Example 3 Testing **Time Fairness** with ***INV**ariance*	Same pred. (inv) after ~~removals~~		
Test Case 3: ~~On June 15th, 2022~~, Defendant Zhang posed as a real estate agent and deceived Wang into transferring 50,000 RMB to a fake account.	inv	Fraud → Fraud 65 → 59	×

Table 2: Examples of LJPCHECK.

tors include voluntary surrender, mental illness, etc. For instance, if someone commits a crime with a mental illness, they may get a lighter punishment than a mentally sound person.

To test whether LJP systems can identify key elements, we derive two kinds of functionalities with respect to *Key Factor* and *Discretionary Sentencing Factor*. As *Key Factors* are relevant to the types of cases and it is impossible to test all types, we select six case types for testing, including *Theft*, *Traffic Offense*, *Passive Bribery*, *Intentional Injury*, *Defraud* and *Robbery* (**F13-F18**). These six types are chosen because there are subtle differences between them and others, which often cause models to make incorrect judgments and thereby make our tests more challenging. For instance, both illegally entering and theft can involve the fact of unlawfully entering a house. If LJP models only identify the fact of *entering a house*, the crime of illegally entering may be wrongly predicted. For *Discretionary Sentencing Factors*, we include voluntary surrender (**F19**), criminal attempt (**F20**), whether receiving forgiveness by victims (**F21**) and mental illness (**F22**) and recidivism (**F23**). These factors are chosen because they can be applied to any case type.

3.3 Test Types

We adopt four test types (i.e. Minimum Functionality test (**MFT**), Invariance test (**INV**), Directional Expectation test (**DIR**), and Real Case Test (**RCT**)) to test the 23 functionalities (Ribeiro et al., 2020). Table 2 illustrates examples of these test types with the corresponding functionalities. All functionalities and their test types can be found in Table 1.

INV and DIR are both inspired by software metamorphic tests (Segura et al., 2016): the former makes label-preserving perturbations to inputs with the expectation that outputs remain consistent, while the latter perturbs the input and anticipates a directional change in the output. For instance, to test *Time Fairness* with INV, removing or replacing the time information in the fact description should make no difference to the prediction. However, in Test Case 3 in Table 2, the predicted term of penalty decreases from 65 months to 59 months while the expectation is the output should remain the same as before. INV is used to test **F2-F10** as LJP models should not show any prediction change in term of fairness. As for DIR, for example, we expect that the term of penalty will decrease if we add the sentence involving *voluntary surrender* in the fact description (Test Case 2 in Table 2). DIR is used to test **F19-F23** as *Discretionary Sentencing Factors* have impacts on the term of penalty and charges, and changes are expected. MFT, which is derived from unit tests in software engineering, uses simple and focused test cases to check a behavior within a functionality. Taking the functionality *Key Factor* as an example, we assess whether a model can make correct predictions for a theft case only with key factors (Test Case 1 in Table 2). We expect that the model will predict a term of penalty that falls within the range of 0 to 36 months. MFT is used to test **F13-18** as we need to construct different test cases to determine whether LJP models can identify *Key Factors*. Besides, some functionalities can be tested directly using ground-truth cases. We refer to this kind of tests as RCT, which is used to test (**F1** and **F10-F12**). For instance, we collect multi-charge cases to construct test cases for testing the *Multi-charge* functionality.

3.4 Generating Test Cases

Two methods are used to generate desired test cases: (1) starting from scratch and (2) perturbing existing data. Using the first method, we can create a small number of high quality test cases, which is typically implemented by filling the templates. The second method generates test cases based on existing data but poses a challenge in defining perturbation functions. Next, we will detail how these two methods are implemented [1].

Templates We generate test cases based on templates for testing the *Key Factor* functionalities (**F13-F18**). In Table 2, we generate "*Defendant Wang sneaked into a house and stole wallets. The stolen items were worth 2,600 RMB.*" from the template "Defendant {Name} sneaked into a house and {Trigger} {Item}. The stolen items were worth {Value}.", where {Name}, {Trigger}, {Item} and {Value} are selected from a predefined list. We generate multiple templates for each *Key Factor Functionality* to improve diversity. Each template is designed to match with case types (e.g., Theft). To be specific, each template contains all the necessary key factors that are stipulated in law. For instance, theft-related templates must contain {Trigger}, {Item} and {Value}. We seek help from legal experts to derive these key factors from law articles. We then construct different contexts containing these key factors and add different confusing semantics to the templates to make the test cases challenging. As Example 1 in Table 2 shows, we add the phrase *sneaked into a house* to mislead LJP models to predict *Illegally Entering* instead of *Theft*, as *Illegally Entering* and *Theft* are both related to the fact of illegally entering. Here, we ask legal experts to identify the confusing semantics with respect to each case type. Regarding the potential candidate values for the key factors, we ask legal experts to identify options and establish their associations with penalty intervals (e.g., if the value of the stolen property is less than 10,000 RMB, the associated penalty is 0-36 months of prison). Gold charges are easily obtained for templates. For gold term of penalty, we identify penalty intervals given the associations with options.

Perturbing Existing Data We generate test cases for INV and DIR by perturbing existing datasets through adding, replacing or removing key information. For example, to test a model's ability to identify the *Voluntary Surrender* (one of the *Discretionary Sentencing Factors*) functionality, we use regular expressions to extract and remove surrender-related words, such as "surrender", "active alarm" and "cooperation with accident investigation", from a predefined list created with the help of a legal expert. We then obtain a pair of test cases, the original case and a perturbed case without factors of voluntary surrender. The penalty term of the original case is supposed to be shorter than that of the perturbed one.

Validating Test Cases To evaluate the quality of the generated test cases, we recruit 3 legal professionals (one judge and two lawyers) and give them a one-hour tutorial on evaluation criteria. Our evaluation criteria include two aspects. The first involves whether the generated fact description is consistent with the generation method. For instance, the perturbation should successfully perturb the target words. The second involves whether the generated fact description is consistent with the expected prediction result. For instance, the generated fact description should lead to an increase of penalty. If a test case meets both criteria, it is positive; otherwise, it is negative. During evaluation, the two lawyers are required to rate test cases independently. The positive ratios are 89.5% and 91.3% respectively. Inter-annotator agreement Cohen's Kappa score (McHugh, 2012) is 0.712, which indicates substantial agreement. Finally, each case of disagreement is resolved by the judge.

In total, we collect 10,800 test cases (See Table 1 for statistics). The original facts and gold labels are collected from China Judgment Online[2].

4 Experiments

We test five SOTA LJP models using LJPCHECK.

4.1 Models

We find that SOTA models achieve different performance across papers. Thus, we conduct preliminary experiments under the same experimental setup to select reliable SOTA models. We evaluate models that are reported to achieve SOTA performance in recent years (2018-2023) on three Chinese LJP tasks using the CAIL-small dataset (Xiao et al., 2018) [3]. Finally, we obtain five SOTA models, i.e., TOPJUDGE (TOP) (Zhong et al., 2018), Lawformer (LAW) (Xiao et al., 2021), NeurJudge (Neur) (Yue et al., 2021), Dependent-T5

[1] See Appendix C for details on functionalities with the corresponding test types/examples and how test cases are generated.

[2] http://wenshu.court.gov.cn/

[3] See Appendix A for preliminary experiment setup details.

	Charges				Law Articles				Term of Penalty
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Log Distance
TOPJUDGE	75.65	71.29	59.73	62.31	71.68	65.80	54.23	56.66	2.07
Lawformer	77.33	72.30	64.05	66.15	74.01	70.40	60.57	62.79	2.12
NeurJudge	72.54	74.64	63.67	66.87	72.55	76.86	63.48	66.94	1.97
Dependent-T5	77.58	78.99	74.00	74.59	70.59	75.47	65.06	67.66	1.71
Rformer	**87.18**	**85.45**	**82.11**	**82.99**	**80.25**	**82.65**	**77.44**	**78.12**	**1.63**

Table 3: SOTA models' performance on CAIL-small. We cast law article and charge prediction as multi-label classification tasks with four metrics namely accuracy (Acc.), macro-precision (MP), macro-recall (MR), and macro-F1 (F1). The term of penalty prediction is cast as a regression task measured by log distance.

(T5) (Huang et al., 2021), and Rformer (R) (Dong and Niu, 2021). Table 3 shows their performance on CAIL-small.

4.2 Functional Test Setup

We ensure there is no overlap between our collected test cases and CAIL-small, and test the five SOTA models after training on CAIL-small in the preliminary experiment. We test the SOTA models on charge and term of penalty prediction tasks. Model performance on functionalities is evaluated by *failure rate*. For MFT and RCT, a "failure" in the charge prediction task occurs when the prediction does not match the gold label. In penalty prediction, any prediction that falls outside a predefined interval is considered a "failure". For INV and DIR, a "failure" is defined by determining the variation between the predicted and expected values. See Appendix D for details on the definition of failure rate.

4.2.1 Performance Analysis

Results are shown in Table 4. All models exhibit weaknesses on **F1**. Even Rformer, which has the highest accuracy score on CAIL-small, has a failure rate of up to 78.3%. A plausible reason is that the training set of CAIL-small only contains a limited number of such samples (6.4%), leading to a bias, i.e., a conviction of a charge results in a penalty.

Multi-charge cases (**F10**) also pose a challenge to all models. When collecting our test cases, we select several combinations of charges that don't exist in the training set of CAIL-small to add challenges. Dependent-T5, which may benefit from leveraging semantics of charges rather than treating them as atomic labels, gets the lowest failure rate (33.0% | 58.3%). Other models do not have any mechanism designed for multi-charge cases.

For fairness functionalities (**F2-F9**), NeurJudge performs the worst on penalty prediction on all fairness functionalities, while Dependent-T5 appears to be the least sensitive to fairness data perturbation. Generally, each model's failure rate significantly differs across fairness functionalities, making it difficult to conclude any obvious bias.

For *F11* and *F12*, we observe that models perform better on misdemeanor cases than felony cases w.r.t. charge prediction. One reason is that felony cases are typically more complicated than demeanor cases. However, we find that models perform better on felony cases than misdemeanor cases w.r.t. term of penalty prediction. This might be because judges handle felony cases with greater caution compared to misdemeanor cases, resulting in fewer disagreements on penalties.

Rformer outperforms other models significantly in all three subtasks on CAIL-small. However, it shows high failure rates on **F13-F18**, suggesting weaknesses in following the legal logic. Note Rformer's failure rate of charge prediction is the highest in identifying key factors for Passive Bribery cases (**F15**). An error analysis reveals that the error samples tend to be predicted as *Bribery Crime of Non-state Staffs* (confused with *Passive Bribery*). It demonstrates that while Rformer adopts strategies like global consistency graph to alleviate the problem of confusing charges, these mechanisms are still far from addressing the issue. NeurJudge and Dependent-T5, which contain confusing alleviation mechanism, also exhibit such problems.

Nevertheless, NeurJudge achieves the best performance on **F19-F23**, possibly due to its *circumstances of crime aware fact separation (CCFS)* method. This method separates the facts into parts, one of which is relevant to *Discretionary Sentencing Factors*.

4.2.2 Improvement Suggestions

We further test two Rformer variants: one replacing its backbone from bert-base-chinese to bert-large-chinese (denoted as R-L), and another trained on

	Functionality and Test Type	Failure Rate(Charge|Term of Penalty)(%)						
		TOP	**LAW**	**Neur**	**T5**	**R**	**R-L**	**R-M**
1 Exemption	**F1** Exemption: *RCT*	– | 94.9	– | *98.7*	– | 95.9	– |89.3	– | **78.3**	– | 82.8	– | 75.4
2 Fairness	**F2** Name: *INV*	5.6 | 11.3	4.7 | 19.3	*9.7* | *22.0*	**1.7** | **7.6**	2.8 | 11.1	3.1 | 10.9	1.7 | 8.1
	F3 Gender: *INV*	3.3| 3.1	*9.4* | 16.1	6.3 | *18.6*	**2.6** | **2.8**	5.4 | 14.6	7.9 | 12.6	2.0 | 13.5
	F4 Ethnicity: *INV*	5.2 | **2.2**	*9.5* | 20.8	5.2 | *22.1*	**3.4** | 4.4	5.3 | 17.8	3.3 | 18.1	2.2 | 20.2
	F5 Education: *INV*	**4.2** | **3.0**	*9.1* | 15.9	7.5 | *25.3*	4.8 | 4.9	6.4 | 21.4	8.5 | 25.7	4.2 | 20.3
	F6 Gender & Ethnicity & Education: *INV*	6.1 | **3.5**	*9.9* | 22.1	7.8 | *24.9*	**4.5** | 4.8	6.3 | 21.8	5.0 | 15.9	3.5 | 12.3
	F7 Time: *INV*	5.1 | 10.4	3.0 | 12.6	*11.8* | *29.0*	**4.3** | **6.9**	11.4 | 26.6	6.7 | 19.4	8.7 | 18.6
	F8 Location: *INV*	2.9 | 7.8	6.5 | 17.0	*7.6* | *20.2*	**1.5** | **6.1**	4.5 | 14.1	5.3 | 16.1	3.9 | 11.4
	F9 Time & Location: *INV*	10.4 | 17.1	9.1 | 19.3	*16.7* | *30.2*	**3.8** | **14.0**	12.9 | 25.8	8.7 | 18.8	9.0 | 12.3
3-1 Multi-charge	**F10** Multi-charge: *RCT*	56.7 | *73.9*	*66.7* | 73.8	60.9 | 71.7	**33.0** | **58.3**	60.8 | 64.3	51.4 | 56.2	47.0 | 57.8
3-2 Felony and Misdemeanor	**F11** Misdemeanor: *RCT*	25.7 | 37.2	24.0 | *40.0*	29.4 | 37.6	*34.7* | 32.4	**22.5** | **30.5**	25.5 | 38.7	15.3 | 33.2
	F12 Felony: *RCT*	34.8 | *36.8*	34.0 | 35.9	*41.0* | 30.1	40.2 | **28.8**	**33.0** | 36.4	31.0 | 32.3	23.2 | 30.5
4-1 Key Factors	**F13** Theft: *MFT*	27.7 | 24.0	*34.7* | *32.3*	22.8 | 28.8	**17.1** | **18.2**	23.0 | 21.0	15.7 | 14.1	8.7 | 11.9
	F14 Traffic Offence: *MFT*	25.0 | 26.7	*38.8* | *39.7*	28.2| 27.5	15.0 | **13.5**	**12.4** | 20.8	10.1 | 17.8	8.9 | 15.4
	F15 Passive Bribery: *MFT*	**20.1** | 26.7	37.8 | *38.0*	30.1 | 29.4	32.5 | 30.2	*39.0* | **25.2**	40.6 | 30.3	36.5 | 29.8
	F16 Intentional Injury: *MFT*	33.3 | *40.2*	40.2 | 47.7	*35.1* | 29.8	**28.2** | 29.7	31.0 | **25.2**	29.2 | 20.1	25.4 | 26.1
	F17 Defraud: *MFT*	35.8 | 36.7	*43.5* | *45.7*	34.3 | **22.0**	31.7 | 32.3	**25.5** | 36.2	25.6 | 36.8	30.2 | 19.5
	F18 Robbery: *MFT*	**23.0** | **30.0**	40.3 | *45.0*	37.0 | 43.2	*41.3* | 32.5	31.5 | 34.5	27.2 | 33.8	23.0 | 26.8
4-2 Discretionary Sentencing Factors	**F19** Voluntary Surrender: *DIR*	– | 50.1	– | 52.1	– | **34.8**	– | *59.0*	– | 54.5	– | 56.4	– | 53.2
	F20 Criminal Attempt: *DIR*	– | 66.9	– | 70.6	– | **54.4**	– | *72.1*	– | 66.5	– | 64.3	– | 67.6
	F21 Forgiveness: *DIR*	– | 51.2	– | 62.2	– | **49.8**	– | 65.9	– | *70.6*	– | 71.3	– | 72.7
	F22 Mental Illness: *DIR*	– | 54.8	– | 70.8	– | **49.4**	– | 76.5	– | *80.7*	– | 77.9	– | 81.0
	F23 Recidivism: *DIR*	– | 63.0	– | 53.9	– | **40.6**	– | *72.9*	– | 69.0	– | 62.5	– | 70.9

Table 4: Testing results of five SOTA models and two Rformer variants. Each cell contains two values divided by 'l', the left one refers to charge prediction and the right refers to penalty term prediction. **Bold** font denotes the best performance and the worst performance is highlighted in *italic red*. Improvements by variants are underlined.

the entire CAIL-big dataset (denoted as R-M, M stands for more data).

As shown in Table 4, R-M outperforms others on **F1** and **F10**, potentially benefitting from more training data, which is nearly eight times than that of Rformer and R-L. R-M learns more comprehensive and robust representations, which also explains why it performs better on **F13-F18** and **F2-F9** as robust representations are less sensitive to misleading semantics. However, R-M is not better at identifying *Discretionary Sentencing Factors*(**F19-F23**). In this case, simply increasing the amount of data might not be sufficient to improve these functionalities. However, R-L shows improvement as it utilizes a larger BERT. It suggests that a larger encoder is necessary to embed semantics.

4.3 Discussion

Building upon the insights gained from our results and analysis, we now delve into a broader discussion of various phenomena and implications.

One peculiar thing is that the performance of charge prediction and term of penalty prediction in the same functionality does not exhibit a clear correlation. The best performance in charge prediction may coincide with the worst performance in penalty prediction (e.g. Rformer on **F15**). Many models aim to improve charge prediction but struggle to enhance penalty accuracy as term of penalty is impacted by complex factors. This also suggests that models do not effectively leverage the correlations among subtasks.

It is also noteworthy that models that perform better on DIR tests (**F19-F23**) show poorer performance on INV tests (**F2-F9**). The sensitivity to fine-grained sentencing circumstances may be unavoidably accompanied by a similar sensitivity to other unrelated information. This implies that models may not accurately identify and understand the criteria but instead engages in a fine-grained learning of all semantic information.

Overall, each model has its own strengths and weakness, and none can pass all tests. For practical applications, these weaknesses can be vital and cannot be overlooked. As a suite of black-box tests, LJPCHECK only partially uncovers a model's shortcomings and offers indirect source of these shortcomings. Specific reasons for these failures and improvement methods require further investigation by the users. Nonetheless, we can still surmise that a significant portion of these reasons stems from the lack of training data and deficiencies in the model structure. If it is the biased training data, models could be improved by targeted data augmentation (Gardner et al., 2020). LJPCHECK users could, for instance, generate or select additional training cases to resemble test cases from functional tests where their model performs poorly. Note, however, that introducing additional data might introduce unforeseen biases.

5 Conclusion

We proposed using functional testing for a more comprehensive evaluation of LJP models. We introduced LJPCHECK, which contains 23 functionalities targeted at testing different aspects of LJP systems. Five SOTA LJP models are tested and weaknesses are revealed. In future work, we aim to explore testing model interpretability given its critical importance in LJP. We also plan to incorporate the results of law article prediction task, which were not utilized in our current tests, to enhance the completeness and reliability of our tests.

Ethical Statements

The real-world data used in this paper are all from publicly available resources and personal information (e.g. name, plate number, etc.) has been anonymized before use. Legal judgment prediction is not researched to replace judges and make final court decision solely by machine. Our motivation is to make the evaluation metrics more comprehensive rather than replace any of them. We do not intend to prove that LJP models have the potential of adjudging any cases without human verification, even if they achieve "perfect" scores on all functionalities.

Limitations

The current version of LJPCHECK covers only a small fraction of functionalities of LJP models which cannot entirely represents practitioners' expectations. For some functionalities, we only test limited charges, leaving a large amount of other charges to be further investigated. Fortunately, LJPCHECK is extensible that users can design their own test cases following our approaches. Regarding this, the test case generation methods in this paper are not diverse enough and can possibly lead to simplification of certain cases. Besides, some functionalities are limited to Chinese and might not be applicable to other jurisdictions. However, it doesn't mean that LJPCHECK lacks generalizability but instead, it highlights that individuals can formulate dedicated functionalities.

Acknowledgments

We would like to thank Xilu Chen, Xiaoyu Shen and Jiwei Shang for their sharing and help at the early stage of this work. We also appreciate the valuable comments from all the reviewers. This work was supported by the Overseas Project (KFKT2023A07, KFKT2024A07) and the General Innovation Project (ZZKT2024B02) of the National Key Laboratory for Novel Software Technology in China, the National Key R&D Program of China (2016YFC0800803) and the Cooperation Fund of NJU-Jinqiao Information Joint Research Center for Smart Justice and Artificial Intelligence.

References

Zhenwei An, Quzhe Huang, Cong Jiang, Yansong Feng, and Dongyan Zhao. 2022. Do charge prediction models learn legal theory? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3757–3768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15513–15535. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Qian Dong and Shuzi Niu. 2021. Legal judgment prediction via relational learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 983–992, New York, NY, USA. Association for Computing Machinery.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022a. Legal judgment prediction: A survey of the state of the art. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5461–5469. ijcai.org.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022b. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. *Preprint*, arxiv:2004.02709.

Ece Gumusel, Vincent Quirante Malic, Devan Ray Donaldson, Kevin Ashley, and Xiaozhong Liu. 2022. An annotation schema for the detection of social bias in legal text corpora. In *Information for a Better World: Shaping the Global Future*, Lecture Notes in Computer Science, pages 185–194, Cham. Springer International Publishing.

Yunyun Huang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, and Bin Luo. 2021. Dependency learning for legal judgment prediction with a unified text-to-text transformer. *CoRR*, arXiv:2112.06370.

Lingzi Jin. 2009. Automated functional testing of search engine. In *Proceedings of the 4th International Workshop on Automation of Software Test, AST 2009, Vancouver, BC, Canada, May 18-19, 2009*, pages 97–100. IEEE Computer Society.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sergio Segura, Gordon Fraser, Ana B. Sanchez, and Antonio Ruiz-Cortés. 2016. A survey on metamorphic testing. *IEEE Transactions on Software Engineering*, 42(9):805–824.

Octavia-Maria Sulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the law area and decisions of french supreme court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 716–722. INCOMA Ltd.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *CoRR*, abs/2103.13868.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *Preprint*, arxiv:2105.03887.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 973–982, New York, NY, USA. Association for Computing Machinery.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

A Experimental Setup

We cast the law article and charge prediction tasks as multi-label classification tasks instead of single-label classification tasks as we want to test the multi-charge functionality. We cast the term of penalty prediction as a regression task instead of a classification task as we want to measure the degrees of increases and decreases. We keep hyperparameters, such as batch size, learning rate, training epochs and random seed, same to those in the original papers. We conduct our experiments on 4*Tesla V100 GPUs.

B Functionality Description

All functionalities, their detailed descriptions and applicable languages are shown in Table 5.

C Generation Method and Example

Details of generation methods and examples are shown in Table 6.

D Functional Test Setup

When testing on the term of penalty prediction task, the definition of '*failure*' differs depending on the functionality and test type, shown in Table 7.

Functionality	**Description**	**Applicable Jurisdiction**
F1: Exemption	According to Article 37 of Criminal Law of the People's Republic of China, if the circumstances of a person's crime are minor, he may be exempted from criminal punishment. LJP model should be able to predict zero punishment for these cases.	Chinese
F2: Name Fairness	The prediction of the LJP model should not be influenced by the names of the parties in the case description.	Chinese English
F3: Gender Fairness	The prediction of the LJP model should not be influenced by the gender of the defendant in the case description.	Chinese English
F4: Ethnicity Fairness	The prediction of the LJP model should not be influenced by the ethnicity of the defendant in the case description.	Chinese English
F5: Education Fairness	The prediction of the LJP model should not be influenced by the education level of the defendant in the case description.	Chinese English
F6: Gender & Ethnicity & Education Fairness	The prediction of the LJP model should not be influenced by the gender, ethnicity and education level of the defendant in the case description altogether.	Chinese English
F7: Time Fairness	The prediction of the LJP model should not be influenced by the time (when the case happened) in the case description. The premise is that the range of variation in time does not involve legal provisions changing.	Chinese English
F8: Location Fairness	The prediction of the LJP model should not be influenced by the location (where the case happened) in the case description. The premise is that the legal provisions involved in the case are not exclusive to a specific locality.	Chinese English
F9: Time & Location Fairness	The prediction of the LJP model should not be influenced by the time and location in the case description altogether.	Chinese English
F10: Multi-charge	One complex case scenario is when the defendant in a case commits multiple offenses simultaneously. The LJP model needs to identify the factual circumstances corresponding to each charge from the facts and match them with relevant legal provisions. When predicting the penalty, the combined result of multiple offenses needs to be taken into account for calculation.	Chinese English
F11: Misdemeanor	Studies show that judicial inconsistency is negatively correlated with the severity of the crime. Minor offense cases typically imply simpler criminal facts, allowing judges to arrive at verdicts more easily. In contrast, serious offense cases require more careful consideration. The LJP model should demonstrate reliable performance for both types of cases.	Chinese
F12: Felony	Refer to F11.	Chinese
F13-F18: Key Factors	In a civil law system, determining whether a specific offense has been committed requires examining if the act is involved in the clearly written provisions, which is known as circumstance. Therefore, recognizing circumstances from facts before searching for relevant law articles is a practice aligning with human mental model and thus enhancing interpretability. Key circumstances are the facts that directly determine the charge which the LJP model are expected to accurately recognize. Different charges involve different circumstances and need to be analyzed separately. We test 6 charges here, namely theft, traffic offence, passive bribery, intentional injury, defraud and rob.	Chinese
F19: Voluntary Surrender	Another kind of circumstance is facts influencing the penalty judgment, named Discretionary Sentencing Circumstance. If there is evidence of voluntary surrender, a more lenient sentence should be given.	Chinese
F20: Criminal Attempt	A "criminal attempt" occurs when an individual takes a significant step towards committing a crime but does not complete the crime itself. In this case, the penalty should be lighter than the one of completed crimes.	Chinese
F21: Forgiveness	Sometimes the defendant can obtain forgiveness from the victim or the victim's family. In such case, the penalty for the defendant will be lighter.	Chinese
F22: Mental Illness	In the legal system, mentally ill individuals who cannot recognize or control their own actions may lack criminal responsibility or have diminished criminal capacity. Such individuals can be subject to reduced or mitigated penalties.	Chinese
F23: Recidivism	In the legal systems, recidivists are subject to more severe punishment.	Chinese

Table 5: Detailed descriptions of functionalities.

ID	Type	#	Generating Method	Example
F1	RCT	200	Collecting real criminal cases containing the element of 'exemption from punishment' from China Judgment Online (CJO: wenshu.court.gov.cn)	Defendant Niu XX, in November 2012, working at a school bus office, accepted a 20,000 yuan rebate from two companies during a tire purchase. Niu used this money for personal gain. After the event, Niu voluntarily surrendered and returned all 2,000 yuan of the taken money. Niu is exempt from criminal punishment for taking bribes.
F2	INV	500	Collecting real cases from CJO then remove (replace actual name with indefinite designation X) or keep the name of all parties using named entity tools.	Defendant ~~Zhang~~ XX conspired to defraud ~~Wang~~ XX, a student of Fengnan No.1 Middle School. Zhang took Wang to the rented room in Tienan Building and threatened him with words. and used ~~Wang~~ XX's mobile phone to call his father and ask for RMB 30,000. Expected: Invariant
F3	INV	500	Collecting real cases from CJO then remove or add the gender of the defendant using regular expression.	Defendant Kou XX, female. Defendant Kou, driving a small car, collided with a motorcycle driven by He X, who was speeding in the same direction. Motorcycle rider He suffered bone marrow injuries in his neck and died after rescue efforts were unsuccessful. Expected: Invariant
F4	INV	500	Collecting real cases from CJO then remove or add the ethnicity (e.g. Han) of the defendant using regular expression.	Defendant Kou XX, Han nationality. Defendant Kou, driving a small car, collided with a motorcycle driven by He X, who was speeding in the same direction. Motorcycle rider He suffered bone marrow injuries in his neck and died after rescue efforts were unsuccessful. Expected: Invariant
F5	INV	500	Collecting real cases from CJO then remove or add the education level (e.g. junior high school level) of the defendant using regular expressions.	Defendant Kou XX, junior high school education. Defendant Kou, driving a small car, collided with a motorcycle driven by He X, who was speeding in the same direction. Motorcycle rider He suffered bone marrow injuries in his neck and died after rescue efforts were unsuccessful. Expected: Invariant
F6	INV	500	Collecting real cases from CJO then remove or add the gender, ethnicity and education level of the defendant all at once using regular expression.	Defendant Kou XX, femal, Han nationality, junior high school education. Defendant Kou, driving a small car, collided with a motorcycle driven by He X, who was speeding in the same direction. Motorcycle rider He suffered bone marrow injuries in his neck and died after rescue efforts were unsuccessful. Expected: Invariant
F7	INV	500	Collecting real cases from CJO then remove (i.e. replace the actual time with indefinite designation of x) or keep the time in case description using regular expression.	Defendant Zhang XX conspired to defraud Wang x, a student of Fengnan No.1 Middle School, who was surfing the Internet in Baile Internet Cafe in Fengnan District at ~~around 23:00 on January 21, 2013~~ XX, and took him to the rented room in Tienan Building, Fengnan District, threatened him with words, and used Wang mou's mobile phone to call his father and ask for RMB 30,000. Expected: Invariant
F8	INV	500	Collecting real cases from CJO then remove (replace the actual location with indefinite designation of 'X place') or keep the location in fact description with named entity tools.	Defendant Zhang XX conspired in advance to defraud Wang X, a student of ~~Fengnan District No.1 Middle~~ XX School, who was surfing the Internet in ~~Baile~~ XX Internet Cafe in ~~Fengnan~~ XX District at around 23:00 on January 21, 2013, and take him to the rented room in ~~Tienan~~ XX Building, ~~Fengnan~~ XX District, threatened him with words, and used Wang X's mobile phone to call his father and ask for RMB 30,000. Expected: Invariant
F9	INV	500	Collecting real cases from CJO then remove or keep the time and location in fact description all at once.	Defendant Zhang XX conspired to defraud Wang XX, a student of ~~Fengnan No.1 Middle~~ XX School, who was in ~~Baile~~ XX Internet Cafe at ~~around 23:00 on January 21, 2013~~XX, and take him to the rented room in ~~Tienan~~ XX Building, threatened him with words, and used Wang's mobile phone to call his father and ask for RMB 30,000. Expected: Invariant
F10	RCT	200	Collecting real criminal cases from CJO where the defendant is convicted of multiple charges (i.e. the charge label consists more than one item).	Defendant Xu X collided with a taxi driven by Wang on the road. Later, the defendant Xu caught up with Wang, they got out of the car and quarreled and fought. During the tussle, the defendant Xu injured Wang's nose with his fist. ... Xu was charged with the crime of **traffic offense** and the crime of **intentional injury**.
F11	RCT	550	Collecting real criminal cases from CJO where the term of penalty is less than 3 years.	The defendant Fan, who was drunk and driving a small car, hit the guardrail, damaging both. After identification, the alcohol content in the blood of him is 210.2mg/100ml... The defendant Fan committed the crime of dangerous driving and was sentenced to two months of detention, with a fine of RMB 2000.

Continued on next page

ID	Type	#	Generating Method	Example
F12	RCT	550	Collecting real criminal cases from CJO where the term of penalty is more than 3 years.	The defendant Gong collided with Xu's electric bicycle while driving a truck in the rain. After, Gong drove over Xu's body and fled the scene. Xu died the same day after unsuccessful hospital treatment... The defendant Gong committed intentional homicide and was sentenced to fifteen years in prison.
F13	MFT	500	Using multiple templates including confusing-charge-expressions. e.g., "The defendant {NAME} sneaked in to the victim's house {Theft-Trigger} {item}. The stolen items were worth {Value}" contains information that may lead to predicting illegally entering.	The defendant Wang sneaked into the victim's house and stole two wallets and five electric devices. The stolen items were worth 7,800 RMB.
F14	MFT	500	Using multiple templates including confusing-charge-expressions. e.g., "The defendant {NAME} was driving drunk and {Accident-Trigger}, {Result}" contains information that may lead to predicting dangerous driving charge.	The defendant Ye ming was driving drunk and collided with the bicycle on the pavement ridden by the victim and crashed into the roadside, resulting in the death of one person.
F15	MFT	500	Using multiple templates including confusing-charge-expressions. e.g., "The defendant {NAME} took advantage of his job position, {Bribery-Trigger} {Value} from others and put into an official account. He later withdrew the money for personal use." contains information that may lead to predicting misappropriating public funds	The defendant Ma ming took advantage of his job position, illegally accepted bribes for a total of 20,000 RMB. from others and put into an official account. He later withdrew the money for personal use.
F16	MFT	500	Using multiple templates including confusing-charge-expressions. e.g., "The defendant {NAME} chased and insulted the victim, when the victim started to argue with him, he {Injury-Trigger}, After medical appraisal, the degree of injury to the victim is Injury-Degree." contains information that may lead to predicting creating disturbances	The defendant Yuan Yihang chased and insulted the victim, when the victim started to argue with him, he picked up a stone from the roadside and hit the victim in the head. After medical appraisal, the degree of injury to the victim is minor injury of Class II.
F17	MFT	500	Using multiple templates including confusing-charge-expressions. e.g., "The defendant {NAME} used the fake patent certificate he created to attract investment from others, resulting in a total amount of {Value} of fraud." contains information that may lead to predicting counterfeiting the patent	The defendant Huang Yichao used the fake patent certificate he created to attract investment from others, resulting in a total amount of 100,000 RMB of fraud.

Continued on next page

ID	Type	#	Generating Method	Example
F18	MFT	500	Using multiple templates including confusing-charge-expressions. e.g., "The defendant {NAME} grabbed the victim's neck with a knife and threatened him to hand over all his possessions. The taken items worth {Value} in total." contains information that may lead to predicting intentional injury.	The defendant Zhou Chenguang grabbed the victim's neck with a knife and threatened him to hand over all his possessions. The taken items worth 3,000 RMB in total.
F19	DIR	300	Collecting real criminal cases from CJO and remove the circumstances of voluntary surrender (e.g. turn oneself in, voluntarily called the police) using regular expression.	Defendant Lu X was driving a motorcycle along Qingzhao Highway in Lanshan District, when it collided with an unlicensed tricycle driven by Bi in front of him without a license...Defendant Lu bears the main responsibility for the accident. ~~After the accident, Lu turned himself in at the scene of the accident.~~ Expect: prison term ↑
F20	DIR	500	Collecting real criminal cases from CJO and remove the circumstances of criminal attempt using regular expression.	Defendant Ma X, together with others, stole a red Suzuki motorcycle at the gate of Baishun Electric Appliances in the city square, worth RMB 6,300; ... defendant Ma committed four thefts, ~~one of which was unsuccessful~~, and the amount of theft was RMB 25,800. Expect: prison term ↑
F21	DIR	500	Collecting real criminal cases from CJO and remove the circumstances of defendant being forgiven using regular expression.	The defendant Cai,driving an unlicensed taxi, stole Chang's bag from an open-windowed parked vehicle on Renmin West Road. The stolen mobile phone was worth 3655 yuan. Later, the items were returned and Cai compensated Chang 20,000 RMB. ~~Chang expresses his understanding for Cai's behavior.~~... Expect: prison term ↑
F22	DIR	500	Collecting real criminal cases from CJO and remove the circumstances of defendant being mentally ill using regular expression.	The defendant Liang, had an argument with and injured his girlfriend after drinking... ~~According to the Shandong Institute of Mental Illness Judicial Appraisal, the defendant Liang suffered from schizophrenia and his ability to identify illegal activities was weakened during the crime, which limits his criminal liability.~~ Expect: prison term ↑
F23	DIR	500	Collecting real criminal cases from CJO and remove the circumstances of recidivism using regular expression.	Defendant Zhang X was arrested by public security officers while selling drugs to He. 0.4 grams of heroin to be sold to He was seized from defendant Zhang. ...and is a ~~repeat offender or a repeat offender of a drug crime, and should be given a heavier punishment.~~ Expect: prison term ↓

Table 6: Test type, number, generating method and example test case of each functionality.

ID	Name	Test Type	Definition of Failure on the task of	
			Charge Prediction	Term of Penalty Prediction
F1	Exemption	RCT	-	Failure when the predicted term of penalty is not 0.
F2-F9	Fairness	INV	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term changes after removals or additions of certain attribute values.
F10	Multi-charge	RCT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the $[0.75, 1.25]$ range of the actual value.
F11-F12	Felony and Misdemeanor	RCT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the $[0.75, 1.25]$ range of the actual value.
F13	Key Factors of Theft	MFT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the sentencing range specified in the law articles. According to [Law Article 264], the prison term intervals for different severity are: {"relatively large amount": $(0, 3]$ years}, { "huge amount": $[3, 10)$ years}, {"especially huge amount or other especially serious circumstances": $[10, +\infty)$ years}, where the degrees like "large" and "huge" are defined with specific numerical range in the law.
F14	Key Factors of Traffic Offence	MFT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the sentencing range specified in the law articles. According to [Law Article 133], the prison term intervals for different severity are: {"serious injuries or deaths or heavy losses of public or private property": $(0, 3]$ years}, { "run away from the spot or other especially flagrant circumstance": $[3, 7)$ years}, {"escape results in death of another person": $[7, +\infty)$ years}.
F15	Key Factors of Passive Bribery	MFT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the sentencing range specified in the law articles. According to [Law Article 383], the prison term intervals for different severity are: {"not less than 5,000 yuan": $(0, 2]$ years}, {"not less than 5,000 but less than 50,000) yuan": $[1, 7)$ years}, {"not less than 50,000 but less than 100,000 yuan": $[5, +\infty)$ years}, {"not less than 100,000 yuan": $[10, +\infty)$ years}.
F16	Key Factors of Intentional Injury	MFT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the sentencing range specified in the law articles. According to [Law Article 234], the prison term intervals for different severity are: {"basic": $(0, 3]$ years}, { "severe injury": $[3, 10)$ years}, {"death, resorting to especially cruel means, reducing the person to utter disability": $[10, +\infty)$ years}, where the degrees like "severe" are defined with specific injury level in the law.
F17	Key Factors of Defraud	MFT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the sentencing range specified in the law articles. According to [Law Article 266], the prison term intervals for different severity are: {"relatively large amount": $(0, 3]$ years}, { "huge amount or other serious circumstances": $[3, 10)$ years}, {"especially huge amount or other especially serious circumstances": $[10, +\infty)$ years}, where the degrees like "large" and "huge" are defined with specific numerical range in the law.
F18	Key Factors of Rob	MFT	Failure when the predicted charge is not the same with the actual one.	Failure when the predicted term falls outside the sentencing range specified in the law articles. According to [Law Article 263], the prison term intervals for different severity are: {"robs public or private property by violence, coercion or other methods": $[3, 10)$ years}, {"intruding residence; robbing on board the means of public transportation; robbing a bank or any other banking institution; repeatedly robbery or robbing a huge sum of money; causing serious injury or death, impersonating a serviceman or policeman in robbing; robbing with a gun; robbing military materials or the materials for emergency rescue, disaster relief or social relief": $[10, +\infty)$ years}.

Continued on next page

ID	Name	Test Type	Definition of Failure on the task of	
			Charge Prediction	**Term of Penalty Prediction**
F19	Discretionary Sentencing Factors of Voluntary Surrender	DIR	-	Failure when the predicted term does not increase compared to the original predicted value after removing voluntary surrender circumstances. Assessed by legal experts, the upper limit of the increased value is 1.8 times the original one.
F20	Discretionary Sentencing Factors of Criminal Attempt	DIR	-	Failure when the predicted term does not increase compared to the original predicted value after removing criminal attempt circumstances. Assessed by legal experts, the upper limit of the increased value is twice the original one.
F21	Discretionary Sentencing Factors of Forgiveness	DIR	-	Failure when the predicted term does not increase compared to the original predicted value after removing the defendant being forgiven circumstances. Assessed by legal experts, the upper limit of the increased value is twice the original one.
F22	Discretionary Sentencing Factors of Mental Illness	DIR	-	Failure when the predicted term does not increase compared to the original predicted value after removing the defendant mentally ill circumstances. Assessed by legal experts, the upper limit of the increased value is 1.8 times the original one.
F23	Discretionary Sentencing Factors of Recidivism	DIR	-	Failure when the predicted term does not decrease compared to the original predicted value after removing voluntary recidivism circumstances. Assessed by legal experts, the lower limit of the decreased value is 0.8 times the original one.

Table 7: The definition of ’failure’ on *charge prediction* task and *term of penalty prediction* task among different functionality and test types. **Predicted terms of penalty and the actual ones are all rounded to integer before failure checking.**