LawShift: Benchmarking Legal Judgment Prediction Under Statute Shifts

Zhuo Han[1*], Yi Yang[1*], Yi Feng[1†]
Wanhong Huang[1], Xuxing Ding[1], Chuanyi Li[1], Jidong Ge[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
fy@nju.edu.cn

南京大學 NANJING UNIVERSITY

Background

➢ What is LJP?

Legal Judgment Prediction (LJP) seeks to predict case outcomes given available case information.

➢ The Problem

1. A significant challenge in LJP is the **dynamic nature of legal systems**, where ongoing statutory revisions continually reshape legal norms.
2. A key limitation of existing LJP models is their **limited adaptability to statutory revisions**.

Fact Description: The defendant stole Tritium, a radioactive substance, from a local chemistry lab and deliberately introduced it into the public drinking water supply in his neighborhood, resulting in severe consequences, including ...
Old Article: ... commits arson, breaches dikes, cause explosion, spreads poisonous substances shall be sentenced ...
Judgment: Non-Violation *(obsolete)*
Revised Article: ... commits arson, breaches dikes, cause explosion, spreads poisonous or radioactive substances shall be sentenced ...
Judgment: Violation *(up-to-date)*

Table 1: An example of different judgment outcomes before and after revisions.

Solution: The LawShift Benchmark

The Contribution

To bridge this gap, we introduce **LawShift**, the first benchmark dataset for evaluating LJP adaptability under statutory revisions.

Benchmark Construction

1. We decompose law articles into **six key components**:
   - subject
   - action
   - object
   - objective condition
   - subjective condition
   - term sentence
2. Revisions are defined by three primary **Amendment Dimensions**:
   - scope changes (e.g., expansion, reduction)
   - condition changes (e.g., addition, removal)
   - sentence changes (e.g., numerical shift)
3. This results in **31 fine-grained revision types**.

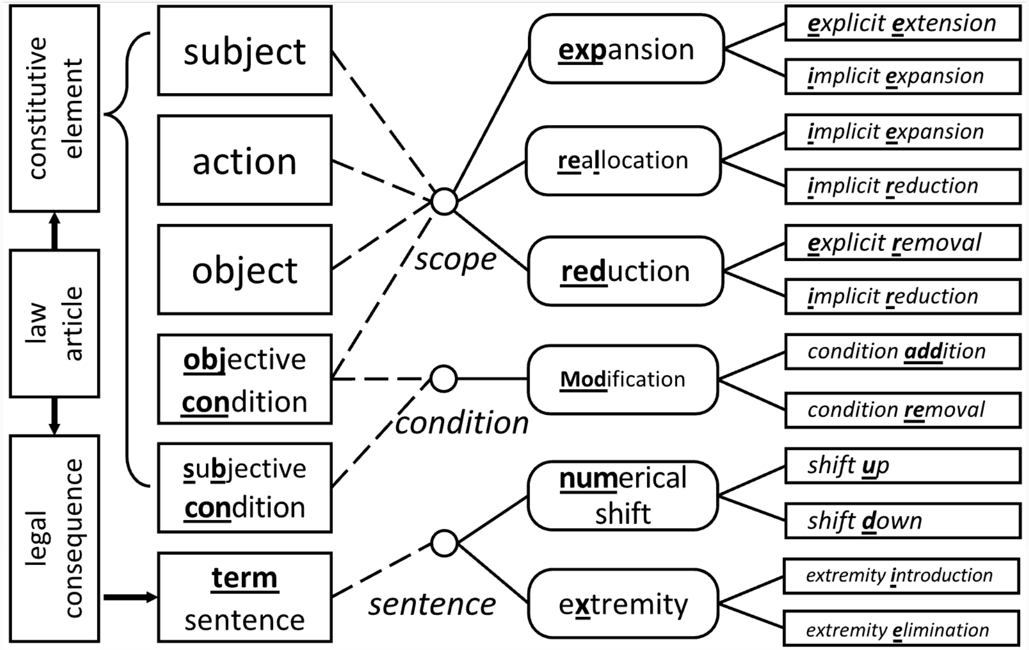

Figure 1: Law article template and amendment dimensions.

南京大學 NANJING UNIVERSITY

Methodology: Metamorphic Testing

Why?

1. **Problem**: Ground-truth labels are often missing or ambiguous for newly revised statutes.

2. **Solution**: We employ **Metamorphic Testing (MT)**, which assesses models by defining logical expectations (Metamorphic Relations) instead of relying on labels.

How?

We use two types of MT：

1. **Minimal Test Instances:** We construct test cases targeting a single capability.

2. **Comparing Outcomes:** We compare model outcomes before and after a revision.

TYPE	LAW RVSN.	FACT ALT.	EXPT
T1.1	*Article 271*: ... personnel of **state-owned companies**, ... → ... personnel of **private enterprises, state-owned companies**, ...	... the defendant while serving as person in charge of company A, which is a **state-owned company**, ... → ... the defendant while serving ... a **private enterprise** B ...	V
T6.1	*Article 232*: ... the sentence shall be **3 to 10 years** imprisonment. → ... the sentence shall be **more than 10 years** imprisonment.	N/A	T ↑

Table 4: Examples of metamorphic testing in LawShift.

Finding 1: Models Fail to Adapt to Change

Metric

- **Pass Rate:** An LJP model is considered to pass a test instance if it generates the expected output given the edited fact.

The Analysis

- Models are reliable when predictions should **remain consistent** (Fig. 3, gray background).
- Models **struggle significantly** when revisions require **different outcomes** (Fig. 3, beige background).
- LLMs show better adaptability only on **sentence changes** (Fig. 3, pink background).

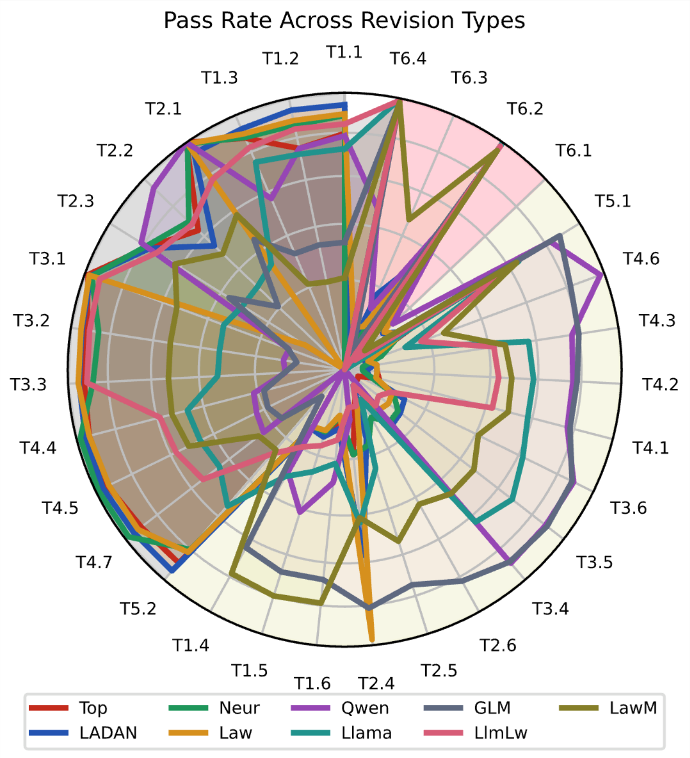

Figure 3: Pass rate across revision types.

Finding 2: Models Lack Semantic Understanding

Metric

- **Pass Rate:** An LJP model is considered to pass a test instance if it generates the expected output given the edited fact.

The Analysis

- We compare model performance on **explicit changes vs. implicit rephrasing**.
- **Finding**: All models perform **better on explicit changes** than on their implicit counterparts.
- **Implication**: Models struggle with semantic understanding and may rely on keyword matching.

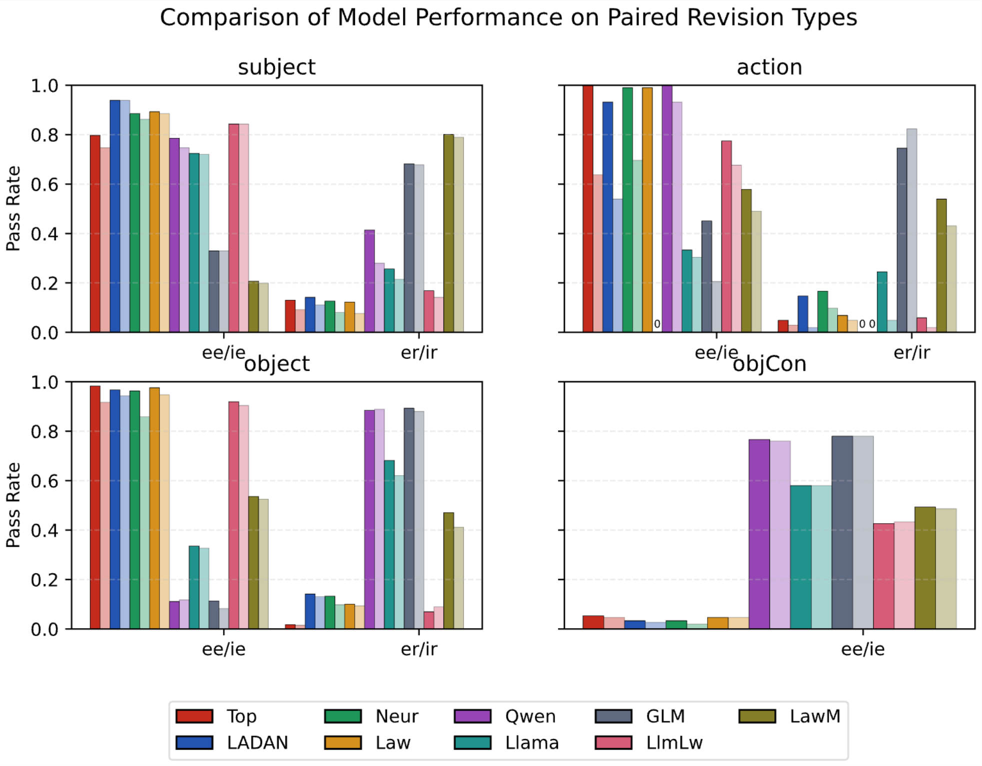

Figure 5: Pass rate differences between explicit and implicit changes. Lighter shades represent implicit changes (ie/ir).

Conclusion & Future Work

➢ Conclusion

1. **We introduced LawShift**, the first benchmark dataset for evaluating LJP adaptability under statutory revisions.

2. **Key Finding:** Existing SOTA models (Neural, PLMs, and LLMs) are **brittle** and **fail to adapt** to legal dynamics, especially when outcomes must change or the revision is semantically implicit.

➢ Future Work

1. Developing models that move beyond keyword matching to deeper semantic understanding.

2. Exploring adaptive RAG or continuous learning frameworks.

3. Creating timestamp-aware and reasoning-based models.

Q&A

Contact us: fy@nju.edu.cn

誠樸雄偉 勵學敦行