# KnowJudge: A Knowledge-Driven Framework for Legal Judgment Prediction

**Zhitian Hou**[1] **(houzht@mail2.sysu.edu.cn)**
**Jinlin Li**[1] **(lijlin57@mail2.sysu.edu.cn)**
**Ge Lin**[2] **(linge3@mail.sysu.edu.cn)**
**Kun Zeng**[1,*] **(zengkun2@mail.sysu.edu.cn)**
[1]Department of Computer Science and Engineering, Sun Yat-sen University
[2]National Engineering Research Center of Digital Life, Sun Yat-Sen University
[*]Corresponding Author

## Abstract

Large Language Models (LLMs) have been extensively employed in Legal Judgment Prediction (LJP) in recent years. However, existing LLM-based methods often fail to effectively simulate the cognitive processes of human judges, particularly in keyword extraction, leading to suboptimal predictions. Inspired by cognitive science, we propose KnowJudge, a knowledge-driven framework, which explicitly models the cognitive process of legal decision-making, leveraging keyword extraction and precedent-based enhancement to guide LLMs in structured legal reasoning. By integrating external legal knowledge tailored to fact descriptions, it refines keyword identification and selects relevant case precedents, thereby mitigating ambiguity in legal judgment. Unlike conventional methods that rely on fine-tuning, KnowJudge improves performance purely through cognitive-process simulation. Experiments on five benchmarks show that KnowJudge outperforms baseline methods, including both general and legal LLMs.

**Keywords:** information integration; large language models; keyword extraction; legal judgement prediction

## Introduction

Legal Judgment Prediction (LJP) stands as one fundamental task in legal domain. The task aims to predict legal judgement, such as charge, based on fact descriptions (Feng, Li, & Ng, 2022; Zhong et al., 2020). Over the past few years, inspired by the success of Large Language Models (LLMs) in a range of domains (Ehara, 2023; Marjieh, Sucholutsky, Sumers, Jacoby, & Griffiths, 2022; Collins, Wong, Feng, Wei, & Tenenbaum, 2022; Waldon, Brodsky, Ma, & Degen, 2023; Prystawski, Thibodeau, Potts, & Goodman, 2023), many Legal AI researchers have endeavored to integrate PLMs or LLMs into LJP tasks (Hwang, Lee, Cho, Lee, & Seo, 2022; Shui, Cao, Wang, & Chua, 2023; Sun, Huang, & Wei, 2024; Y. Wu et al., 2023; Yue et al., 2023).

Though the application of LLMs enhances the interpretability of LJP, using naive legal LLMs to understand legal case documents may lead to hallucinations. Cognitive load theory (Sweller, 1988) suggests that external knowledge guidance can reduce cognitive load and enhance reasoning efficiency. In this regard, providing explicit legal knowledge, such as case-specific keywords and precedents, serves as a structured cognitive scaffold that improves LLM performance without the need for additional fine-tuning.

The application of keyword extraction to various tasks has been widely demonstrated to be effective, particularly when incorporates within prompt templates (Sun et al., 2024; Apsel, Kumar, & Jones, 2022). In judicial practice, judges engage in comprehensive cognitive reasoning processes, leveraging keywords to retrieve relevant law articles and case precedents. For instance, distinguishing between ”dangerous driving” and ”traffic accident” often hinges on precise legal indicators. Fig. 1 illustrates a case where keywords such as ”small sedan” and ”159.65mg/100ml” serve as critical determinants for charge prediction.

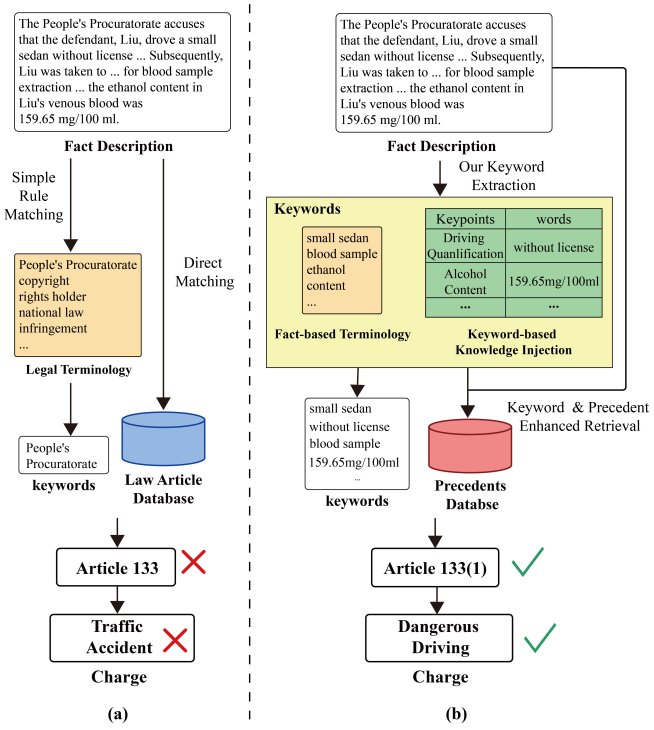


Figure 1: Comparison of main ideas between existing methods and KnowJudge. (a) illustrates the existing LLM method for LJP. (b) shows the innovations of KnowJudge.

However, recent researches based on PLMs or LLMs (Sun et al., 2024; Shui et al., 2023; Y. Wu et al., 2023; L. Li, Liu, Zhao, Zhang, & Liu, 2024) suffer from several limitations, as illustrated in Fig. 1(a):

- The methods for obtaining keywords are limited to simple rule matching between the words in existing general



In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* 

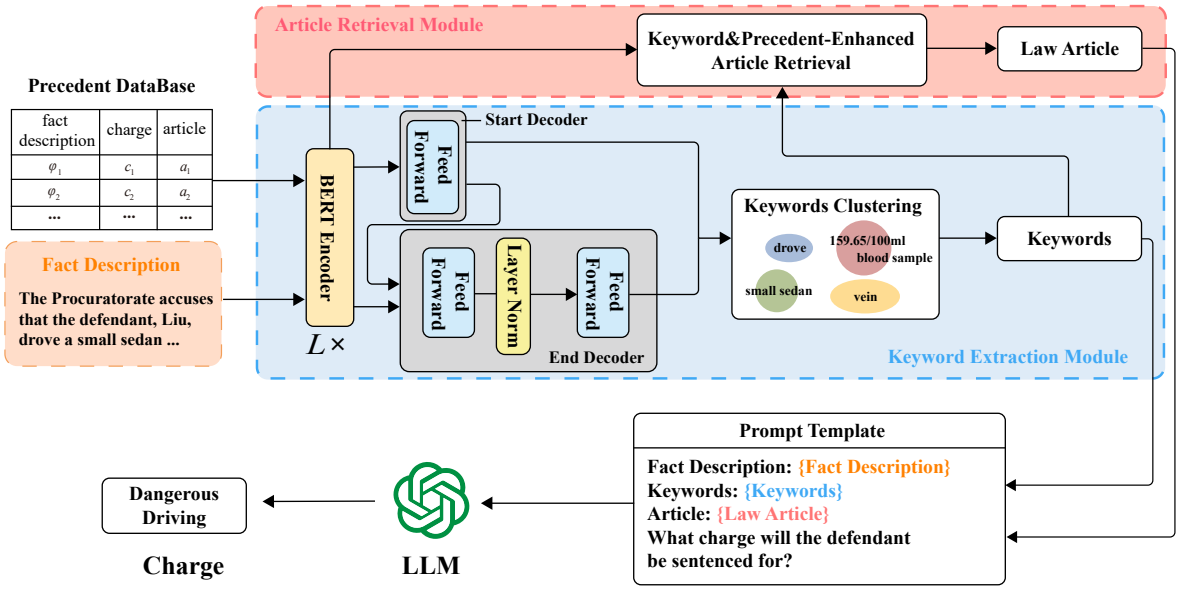


Figure 2: Illustration of our proposed KnowJudge framework. KnowJudge is composed of two modules: Keyword Extraction Module and Article Retrieval Module.

legal terminology and in the case fact descriptions. They can only match generic terms such as "People's Procuratorate", failing to capture truly impactful keywords such as "159.65mg/100ml" for LJP tasks.

- The fact description is directly matched with the legal articles. However, due to the significant differences between fact descriptions and law articles in document structure and description format, direct matching often fails to achieve accurate results.

To address the aforementioned issues, we propose KnowJudge, a Knowledge-Driven Cognitive Simulation Framework based on LLM for LJP. Our approach is based on cognitive science principles, particularly information integration theory (Noble & Shanteau, 1999). This theory posits that decision-making improves when key knowledge elements are structured in a way that reflects human reasoning processes. Unlike existing methods that rely solely on LLM fine-tuning, KnowJudge simulates judicial reasoning by leveraging structured legal knowledge. Specifically, we introduce a keyword extraction method that identifies case-specific terms and mimics how judges assess key legal aspects. By incorporating expert-annotated legal knowledge, KnowJudge captures essential charge-specific features for reasoning. Additionally, our unsupervised keyword-based retrieval method matches case precedents and relevant laws, enriching the LLM's understanding. KnowJudge then integrates fact descriptions, extracted keywords, and relevant laws into structured prompts, enabling LLMs to make accurate legal predictions without additional training.

By eliminating the need for LLM retraining, the performance gains in KnowJudge stem from cognitive process simulation rather than parameter optimization of LLM. This design ensures that improvements arise from a better alignment with human legal cognition, rather than from brute-force model adaptation.

The contributions of the paper are summarized as follows:

- A new legal fact description keywords dataset, Law Keyword Recognition (LKR), has been developed specifically for the LJP task. The dataset incorporates knowledge injection with a focus on keyword extraction from legal case documents. Additionally, a corresponding method for keyword extraction has been introduced.
- A new article retrieval method is introduced. It is an unsupervised method based on keywords and precedents. It can amplify the commonalities between two legal documents and achieve better results to yield the matching law articles.
- A new framework based on LLMs is proposed for the LJP task, named KnowJudge, which simulates the cognitive process of judges. Our proposed KnowJudge achieves a new state-of-the-art performance in LJP compared with a set of baseline methods based on LLM by comprehensive experiments.

## Method

The overall architecture of our proposed KnowJudge framework consists of two modules, as illustrated in Fig. 2. The fact description is first input into the Keyword Extraction Module to extract keywords. Consequently, the Article Retrieval Module uses the keywords to retrieve law articles with

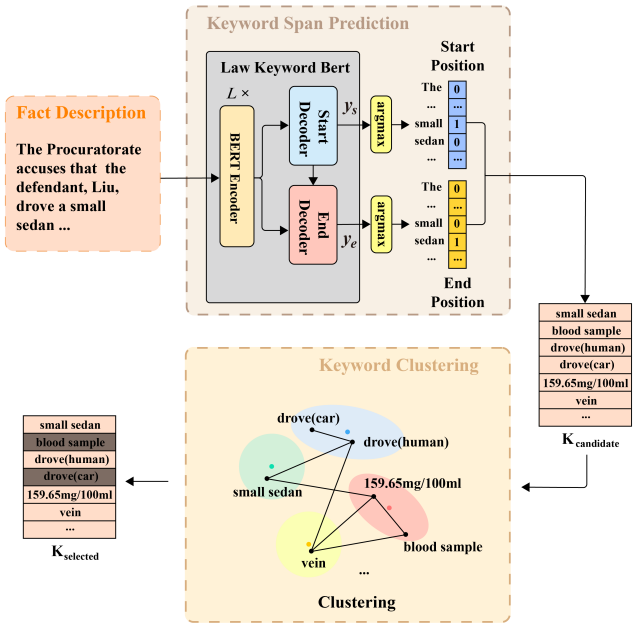


Figure 3: Illustration of Keyword Extraction Module. This module consists of two phases: Keyword Span Prediction phase and knowledge-driven Keyword Clustering phase. In the first phase, the Law Keyword BERT is fed with fact descriptions to predict the start position and end positions of keywords. Candidate keywords are then input into the second phase, where they are clustered to eliminate synonyms and ensure diversity of keywords by keyword clustering. The shaded words are those that need to be eliminated.

precedents. Finally, the keywords and law articles, along with the fact description, are used in the prompt template to guide LLM to predict corresponding judgements.

## Keyword Extraction Module

The Keyword Extraction Module is utilized to extract keywords from the fact description. As shown in Fig. 3, this module comprises two phases: Keyword Span Prediction phase and knowledge-driven Keyword Clustering phase. The Keyword Span Prediction phase recognizes the start and end positions of keywords such as "159.65mg/100ml" in fact descriptions. All candidate keywords $K_{candidate}$ are extracted for further optimization in the subsequent phase. The Keyword Clustering phase then clusters $K_{candidate}$ to eliminate semantically similar keywords, obtaining the most representative keywords $K_{selected}$ from each cluster as the output of Keyword Extraction Module.

In Keyword Span Prediction phase, we leverage Law Keyword Bert to predict the span of keywords or keyphrases. Law Keyword Bert consists of three components: Bert Encoder, Start Decoder, and End Decoder. BERT Encoder is employ to obtain embedding $E = [e_1, e_2, ..., e_T]$ of fact description, where $T$ denotes the number of tokens and $e_i (1 \leq i \leq T)$ denotes the embedding of $i$-th token. We utilize Start Decoder and End Decoder to predict the probabilities of keywords start position $y_s$ and end position $y_e$.

Keywords of fact description are extracted after obtaining their start and end positions. Nonetheless, a significant number of keywords exhibit synonymous or identical terms. Therefore, we utilize the knowledge-driven spectral clustering to enhance keyword diversity. We initially derive the embeddings of fact description and keywords using Law Keyword Bert. The $2 \times top_n$ keywords with the highest cosine similarity to the fact description $K_{candidate}$ are input to build the keyword knowledge graph. The graph is constructed with words as nodes and the similarity between words as the edges.

We employ Normalized Cut Spectral Clustering (Shi & Malik, 2000), which is suitable for graph-based data (X. Li, Kao, Luo, & Ester, 2018), to cluster the words. The average of the embeddings for each cluster are serve as the clusters' centers $C = [c_1, c_2, ..., c_{top_n}]$. The words nearest to the cluster center $c_i (1 \leq i \leq top_n)$ in each cluster are selected as the final keywords $K_{selected}$. The keywords selected in this method are diverse and are the most representative words in each clusters.

**Algorithm 1** Keyword and Precedent Enhanced Article Retrieval (KPEAR)

Obtaining the vector of precedent database as $P$, where $p_i \in P$ is the representation of i-th precedent.
Extracting keywords of i-th precedent and obtain the vectors of keywords $w_i$. Let $M$ be the number of precedents in the database.
**for** each $i \in [1, M]$ **do**
    $\hat{p}_i \leftarrow AvgPool(p_i, w_i)$
**end for**
Obtaining the vector of fact description by Law Keyword Bert as $f$.
Extracting keywords of fact description and obtain the embedding of keywords $w_f$ by Law Keyword Bert.
$\hat{f} \leftarrow AvgPool(f, w_f)$
$score \leftarrow cossim(\hat{f}, \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_M\})$
**return** article of precedent with maximum score

## Article Retrieval Module

In this module, we employ precedent matching based on Keyword Extraction Module to retrieval the relevant law articles. Specifically, we design Keyword and Precedent Enhanced Article Retrieval (KPEAR) to retrieve law articles based on keywords and precedent. The pipeline of Article Retrieval Module is shown in Fig. 4. Firstly, the case precedents are stored as a triplet $(\varphi, c, a)$, which indicates the fact description, charge and article of the case precedent. Keyword Extraction Module and Law Keyword Bert are employed to obtain the precedent fact embedding $p_i$ and its keywords embeddings $w_i = \{w_1^i, w_2^i, ...\}$. These embeddings are stacked and used average pooling to generate the keyword-enhanced embeddings $\hat{p}_i$:

$$\hat{p}_i = AvgPool(p_i, w_i) \quad (1)$$

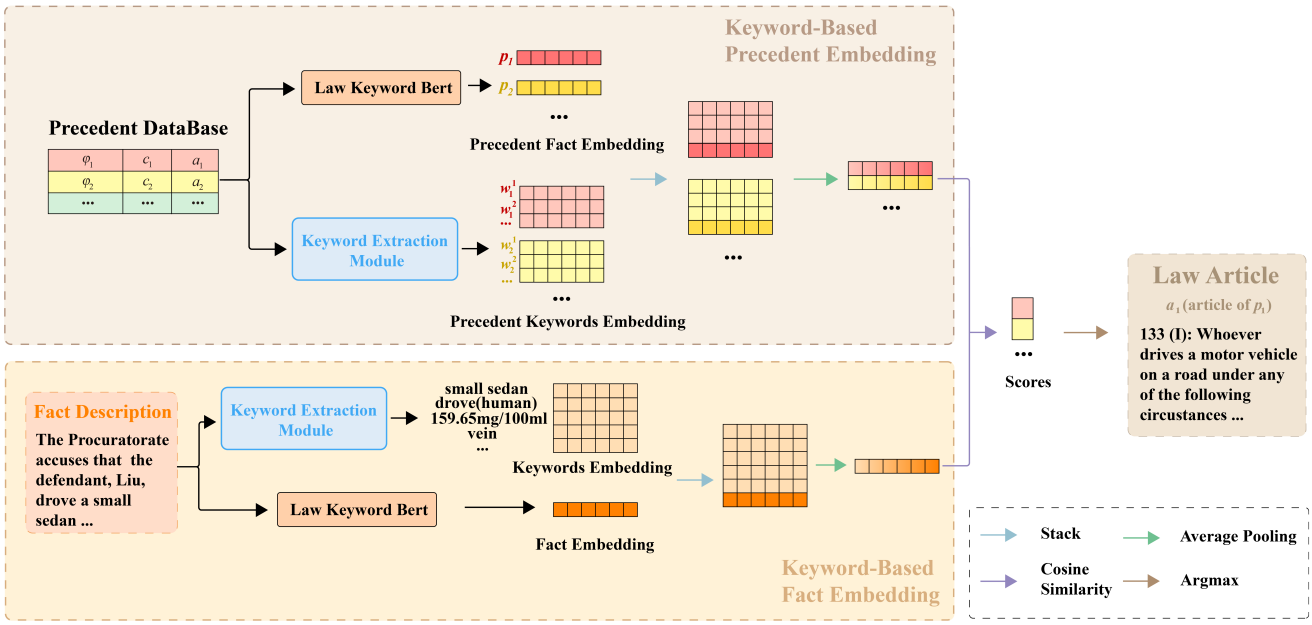


Figure 4: Illustration of Article Retrieval Module. The upper and lower parts of the figure represent the processes of keyword-based precedent embedding and keyword-based fact embedding, respectively. $w_i^j$ denotes the $j$-th keyword of the $i$-th precedent.

Table 1: The size of each test dataset.

| Dataset | Number |
|---|---|
| LKR | 581 |
| Criminal-S | 2377 |
| Criminal-M | 5922 |
| Criminal-L | 11875 |
| CAIL2018 | 17079 |

The keyword-enhanced embedding of the fact description to be predicted $\hat{f}$ is obtained through a similar process by its original embedding $f$ and keywords embedding $w_f$. Cosine similarity is employed to calculate similarity scores between $\hat{f}$ and $\hat{p}_i$:

$$score_i = Cossim(\hat{f}, \hat{p}_i) \quad (2)$$

The law article of precedent with max score is selected as output. The overall algorithm of KPEAR is summarized in Algorithm 1.

Finally, the fact description to be predicted, along with the keywords and legal articles extracted by the two modules described above, are incorporated into the prompt template. This template guides the LLM in discerning the corresponding charge of the fact description.

## Experiments

### Datasets

In this paper, we test our approach on four external datasets: Criminal-S, Criminal-M, Criminal-L (Hu, Li, Tu, Liu, & Sun, 2018), CAIL2018 (Xiao et al., 2018). Each case encompasses both fact description and legal judgment, which is further divided into three components: law articles, charges, and prison terms. Given that the charge represents a key outcome in criminal cases, we focus on evaluating the judgment related to the charge. Furthermore, We construct a dataset, namely Law Keyword Recognition (LKR), which is utilized for training Law Keyword Bert to identify keywords or key phrases from fact descriptions. The LKR dataset is collected from China Judgments Online[1], spanning the years 1985 to 2021 and was filtered cases with a single charge. Additionally, cases that were suspended, not accepted, dis-

After identification, it was determined that Zhao, while intoxicated and without a driver's license,
Driving Qualifications of Perpetrators
was driving an overloaded car carrying 16 barrels of a substance suspected to be gasoline. Zhao ran a
Whether the vehicle is overcrowded or overloaded
red light at approximately 38 kilometers per hour, hitting and knocking down a pedestrian, Yao.
Vehicle Speed
According to the assessment, the 16 barrels of the suspected substance were confirmed to be gasoline,
with a total weight of 120 kilograms. According to the assessment, Zhao's blood alcohol content was
Loading of Chemicals
163.15 mg/ml, classifying it as drunk driving.
Alcohol Content of the Accused

Figure 5: Examples of key points of charge. It pertains to the fact description in a legal document regarding dangerous driving. The 5 highlighted sections correspond to the 5 key points of the dangerous driving.

[1]https://wenshu.court.gov.cn/

Table 2: Results of all dataset, the best is bolded and the second best is underlined.

| Model | LKR | | | | Criminal-S | | | | Criminal-M | | | | Criminal-L | | | | CAIL2018 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F |
| Chinese-Alpaca-2-1.3B | 0.22 | 0.61 | 0.32 | 0.21 | 0.30 | 0.28 | 0.21 | 0.18 | 0.29 | 0.38 | 0.21 | 0.18 | 0.29 | 0.51 | 0.21 | 0.19 | 0.22 | 0.44 | 0.31 | 0.20 |
| Chinese-Alpaca-2-7B | 0.42 | 0.75 | 0.39 | 0.35 | 0.51 | 0.83 | 0.28 | 0.27 | 0.50 | 0.74 | 0.26 | 0.24 | 0.51 | 0.83 | 0.26 | 0.25 | 0.41 | 0.59 | 0.36 | 0.32 |
| ChatGLM | 0.22 | 0.59 | 0.31 | 0.24 | 0.21 | 0.60 | 0.23 | 0.12 | 0.22 | 0.62 | 0.23 | 0.13 | 0.22 | 0.60 | 0.23 | 0.12 | 0.16 | 0.71 | 0.26 | 0.17 |
| ChatGPT(GPT-3.5-turbo) | 0.75 | 0.73 | 0.78 | 0.73 | 0.85 | 0.86 | 0.79 | 0.77 | - | - | - | - | - | - | - | - | - | - | - | - |
| Fuzi-Mingcha | 0.85 | **0.93** | 0.78 | 0.70 | 0.84 | 0.94 | 0.77 | 0.81 | 0.85 | 0.94 | 0.79 | 0.82 | 0.85 | 0.94 | 0.79 | 0.82 | 0.77 | 0.81 | 0.77 | 0.71 |
| LexiLaw | 0.77 | 0.77 | 0.63 | 0.62 | 0.74 | 0.69 | 0.49 | 0.49 | 0.74 | 0.63 | 0.48 | 0.49 | 0.74 | 0.83 | 0.48 | 0.49 | 0.62 | 0.65 | 0.55 | 0.50 |
| Hanfei | 0.82 | 0.85 | 0.78 | 0.72 | 0.91 | 0.93 | 0.83 | 0.87 | 0.90 | 0.92 | 0.83 | 0.87 | 0.90 | 0.92 | 0.83 | 0.87 | 0.72 | 0.79 | 0.74 | 0.71 |
| Lawyer-LLaMA | 0.80 | 0.74 | 0.88 | 0.76 | 0.93 | 0.86 | 0.86 | 0.84 | 0.93 | 0.87 | 0.86 | 0.85 | 0.93 | 0.86 | 0.85 | 0.84 | 0.73 | 0.74 | 0.83 | 0.72 |
| KnowJudge (Chinese-Alpaca-2-1.3B) | 0.36 | 0.44 | 0.36 | 0.29 | 0.49 | 0.36 | 0.24 | 0.25 | 0.48 | 0.46 | 0.24 | 0.27 | 0.48 | 0.54 | 0.26 | 0.28 | 0.37 | 0.55 | 0.35 | 0.29 |
| KnowJudge (Lawyer-LLaMA) | **0.92** | **0.93** | **0.91** | **0.89** | **0.97** | **0.95** | **0.89** | **0.91** | **0.97** | **0.96** | **0.89** | **0.92** | **0.97** | **0.96** | **0.88** | **0.91** | **0.89** | **0.86** | **0.91** | **0.85** |

Table 3: Results of the ablation study on the proposed methods, the best is bolded and the second best is underlined.

| Model | LKR | | | | Criminal-S | | | | Criminal-M | | | | Criminal-L | | | | CAIL2018 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F | Acc | Ma-P | Ma-R | Ma-F |
| w/o Keywords & Article | 0.83 | 0.59 | 0.66 | 0.62 | 0.79 | 0.64 | 0.70 | 0.66 | 0.80 | 0.85 | 0.70 | 0.67 | 0.80 | 0.65 | 0.70 | 0.67 | 0.75 | 0.79 | 0.67 | 0.66 |
| w/o Keywords | 0.85 | 0.86 | 0.81 | 0.83 | 0.91 | 0.88 | 0.80 | 0.82 | 0.92 | 0.89 | 0.84 | 0.86 | 0.93 | 0.89 | 0.84 | 0.86 | 0.85 | 0.80 | 0.80 | 0.83 |
| w/o Article | 0.83 | 0.86 | 0.71 | 0.73 | 0.94 | 0.94 | 0.84 | 0.87 | 0.93 | 0.93 | 0.83 | 0.85 | 0.94 | 0.95 | 0.84 | 0.86 | 0.80 | 0.83 | 0.76 | 0.78 |
| KnowJudge | **0.92** | **0.93** | **0.91** | **0.89** | **0.97** | **0.95** | **0.89** | **0.91** | **0.97** | **0.96** | **0.89** | **0.92** | **0.97** | **0.96** | **0.88** | **0.91** | **0.89** | **0.86** | **0.91** | **0.85** |

missed, or withdrawn are excluded. To achieve knowledge injection, we transform relevant articles and judicial interpretations into key points and identify corresponding keywords or key phrases in fact descriptions. The model is implicitly injected knowledge by learning these keywords. Multiple legal professionals are tasked with setting up to five key points for each of the charges. The charges include dangerous driving, theft, intentional injury, traffic accident and smuggling-trafficking-transporting-manufacturing drugs. An examples of key points annotation of dangerous driving are illustrated in Fig. 5. To further enrich the keyword pool, we employed TF-IDF to construct a vocabulary for each charge. We sample 1000 cases per charge and utilize the BMEO scheme for annotation. The dataset are divided randomly into training, validation and test set in an 8:1:1 ratio. We extract subsets of the charges we annotated and excluded samples that appeared in the LKR training set for each external dataset. The number of samples for each test set is shown in Table 1[2].

## Metrics

Evaluation metrics include Accuracy (Acc), Macro-Precision (Ma-P), Macro-Recall (Ma-R) and Macro-F1 (Ma-F). Since KnowJudge and baselines are generative models, the responses may not contain the exact charges. To address this, we use SimCSE (Gao, Yao, & Chen, 2021), a sentence embedding model based on contrastive learning, to calculate the similarity of their response and exact charge.

## Baselines

We compared several general LLMs, legal LLMs for LJP tasks that have shown strong performance on datasets. We focus on investigating the effectiveness of LLMs in LJP by leveraging their ability to understand complex legal contexts. Therefore, we have not included comparisons with conventional deep learning models.

[2]Our data and code are available at https://github.com/ZhitianHou/KnowJudge

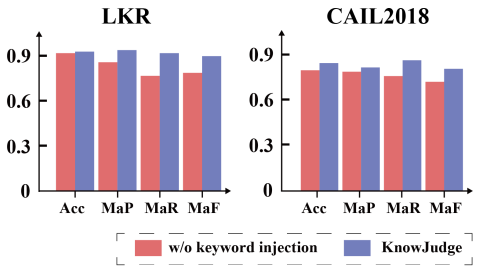


Figure 6: Results of impact of keyword injection on LKR and CAIL2018 datasets.

**Chinese-Alpaca-2-RLHF Series** (Cui, Yang, & Yao, 2023) expands and optimizes the Chinese vocabulary based on the vanilla LLaMA-2 (Touvron et al., 2023). **ChatGLM** (Du et al., 2022) is a Chinese Model fine-tuned on mixed Chinese and English datasets. **ChatGPT** [3] is a powerful conversational model exhibiting strong performance on various Q&A datasets. **Fuzi-mingcha** (S. Wu et al., 2023) is a Chinese legal model trained on a large corpus of Chinese judicial data. **LexiLaw** (Haitao Li & Liu, 2024) is trained on a general-domain data and professional legal data. **HanFei** (He et al., 2023) is trained on a diverse range of Chinese legal data sources. **Lawyer-LLaMA** (Huang et al., 2023) performs continual pre-training on a large-scale legal corpus and is fine-tuned using legal data.

# Results

## Main Results

Table 2 shows the performance results of various models on the LKR, Criminal-S, Criminal-M, Criminal-L and CAIL2018 test sets. Notably, we solely report ChatGPT's performance on the LKR and Criminal-S datasets. The results

[3]https://platform.openai.com/docs/models/gpt-3-5-turbo

is shown that: 1) Our model consistently achieves nearly best performance across all test datasets, significantly outperforming baseline models. The performance of our model demonstrates an average improvement of 7%@Acc and 10%@Ma-F. 2)KnowJudge has been demonstrated to be effective with both general-domain and legal-domain LLMs. Upon applying our framework, a significant improvement in performance was observed across all metrics. 3) On the CAIL2018 dataset, the performance of all models has declined to varying degrees, indicating that the CAIL2018 dataset is a difficult dataset for most models. Our model achieves the state-of-the-art with 0.89@Acc, 0.86@Ma-P, 0.91@Ma-R and 0.85@Ma-F on CAIL2018 dataset, improving by 16%@Acc and 9%@Ma-F. 4) Legal domain LLMs have a better effect in LJP than general domain LLMs, with an average improvement of 11%@Acc, 19%@Ma-P, 11%@Ma-R and 9%@Ma-F on LKR and Criminal-S datasets.

### Ablation Study

To evaluate the individual contributions of each module in our approach, we designed ablation experiments. Specifically, we remove the corresponding modules from KnowJudge, and displayed the results in Table 3. From the results, the performance decreases after removing each part, indicating that all components in KnowJudge contribute to the final performance. As expected, the model performance drops most markedly after removing both keywords and law article, surpassing the cumulative decline observed when each module is removed individually. This observation underscores the synergistic effect of the modules within our framework, demonstrating a collective enhancement greater than the sum of its parts.

### Comparison of Knowledge Injection

To assess the effectiveness of the knowledge-injection keyword extraction method proposed in KnowJudge, we conducted comparative experiments with removing the keywords of corresponding key points. The results as shown in Fig. 6 demonstrate that our method consistently outperformed the approach that only used fact-based terminology across all datasets, underscoring the effectiveness of knowledge injection in keyword extraction.

## Discussion

### Case Study

We conducted a case study of the KnowJudge for prediction. We visualized our keyword extraction and precedent matching results as shown in Fig. 7. In this example, the charge label is dangerous driving. The highlighted words in the figure represent the extracted keywords. The darker the color, the higher the attention score of Law Keyword Bert for the word or phrase. The results show that Law Keyword Bert places significant attention on keywords related to the case's charge. This mirrors the human brain's attention system, which selectively allocates attention resources to enhance the processing of key information. On the other hand, the results revealed that the fact description and the precedents have similar document structures and keywords. This allows them to have better matching performance in the case of the keyword-enhanced method. However, the corresponding law articles and fact description have only a few similar words and different document structures. This result demonstrates the effectiveness of the keywords we extracted and keyword & precedent enhanced article retrieval, as well as their important role in distinguishing confusing charges.

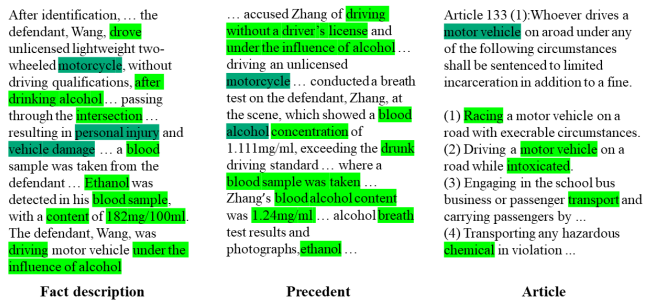


Figure 7: The visualization of the attention of Law Keyword Bert on the fact description, precedent fact and article. The darker the color, the higher the attention score.

### Ethical Discussion

In light of these concerns, we have to state that our work is an algorithmic exploration and will not be used directly in court at this time. Our goal is to provide advice to judges rather than make final judgments without human intervention. In practical applications, human judges should be the last guarantee for judicial fairness.

## Conclusion

To address the deficiency in keyword extraction for the existing LJP method based on LLMs, this paper proposes KnowJudge, a knowledge-driven cognitive simulation framework and introduces the Law Keyword Recognition (LKR) dataset. KnowJudge emulates the cognitive and reasoning processes of human judges by leveraging external knowledge for information focusing and integration and has been rigorously evaluated across five datasets. Experimental results demonstrate that KnowJudge achieves optimal performance, outperforming general domain LLMs and legal domain LLMs. Furthermore, we conduct experiments by removing the knowledge injection in keyword extraction of our framework for comparative analysis. The experimental results confirm the efficacy of the external knowledge injection approach we employ.

## Acknowledgements

This research was supported by National Key Research and Development Program of China (No.2021YFF0900900), and in part by the National Science Foundation of China (No.U1711266).

## References


Apsel, M., Kumar, A. A., & Jones, M. N. (2022). Finding the right words: A computational model of cued lexical retrieval. In J. Culbertson, H. Rabagliati, V. C. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society, cogsci 2022, toronto, on, canada, july 27-30, 2022.* cognitivesciencesociety.org.

Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In J. Culbertson, H. Rabagliati, V. C. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society, cogsci 2022, toronto, on, canada, july 27-30, 2022.* cognitivesciencesociety.org.

Cui, Y., Yang, Z., & Yao, X. (2023). Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022, May). GLM: General language model pretraining with autoregressive blank infilling. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 320–335). Dublin, Ireland: Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.26

Ehara, Y. (2023). Applying large language models to generate high-quality multiple-choice test questions. In M. B. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th annual meeting of the cognitive science society, cogsci 2023, sydney, nsw, australia, july 26-29, 2023.* cognitivesciencesociety.org.

Feng, Y., Li, C., & Ng, V. (2022, 7). Legal judgment prediction: A survey of the state of the art. In L. D. Raedt (Ed.), *Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22* (pp. 5461–5469). International Joint Conferences on Artificial Intelligence Organization. (Survey Track) doi: 10.24963/ijcai.2022/765

Gao, T., Yao, X., & Chen, D. (2021, November). SimCSE: Simple contrastive learning of sentence embeddings. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6894–6910). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552

Haitao Li, Q. D., Qingyao Ai, & Liu, Y. (2024). *Lexilaw: A scalable legal language model for comprehensive legal understanding.* Retrieved from `https://github.com/CSHaitao/LexiLaw`

He, W., Wen, J., Zhang, L., Cheng, H., Qin, B., Li, Y., ... Yang, M. (2023). *Hanfei-1.0.* `https://github.com/siat-nlp/HanFei`. GitHub.

Hu, Z., Li, X., Tu, C., Liu, Z., & Sun, M. (2018, August). Few-shot charge prediction with discriminative legal attributes. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 487–498). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Huang, Q., Tao, M., An, Z., Zhang, C., Jiang, C., Chen, Z., ... Feng, Y. (2023). Lawyer llama technical report..

Hwang, W., Lee, D., Cho, K., Lee, H., & Seo, M. (2022). A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, *35*, 32537–32551.

Li, L., Liu, D., Zhao, L., Zhang, J., & Liu, J. (2024). Evidence mining for interpretable charge prediction via prompt learning. *IEEE Transactions on Computational Social Systems*, *11*(4), 4556-4566. doi: 10.1109/TCSS.2022.3178551

Li, X., Kao, B., Luo, S., & Ester, M. (2018). Rosc: Robust spectral clustering on multi-scale data. In *Proceedings of the 2018 world wide web conference* (p. 157–166). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. doi: 10.1145/3178876.3185993

Marjieh, R., Sucholutsky, I., Sumers, T. R., Jacoby, N., & Griffiths, T. (2022). Predicting human similarity judgments using large language models. In J. Culbertson, H. Rabagliati, V. C. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society, cogsci 2022, toronto, on, canada, july 27-30, 2022.* cognitivesciencesociety.org.

Noble, S., & Shanteau, J. (1999). Information integration theory: A unified cognitive theory. *Journal of Mathematical Psychology*, *43*, 449–454.

Prystawski, B., Thibodeau, P. H., Potts, C., & Goodman, N. D. (2023). Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. In M. B. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th annual meeting of the cognitive science society, cogsci 2023, sydney, nsw, australia, july 26-29, 2023.* cognitivesciencesociety.org. Retrieved from `https://escholarship.org/uc/item/2q01t47h`

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888-905. doi: 10.1109/34.868688

Shui, R., Cao, Y., Wang, X., & Chua, T.-S. (2023, December). A comprehensive evaluation of large language models on legal judgment prediction. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 7337–7348). Singapore: Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.490

Sun, J., Huang, S., & Wei, C. (2024). Chinese legal judgment prediction via knowledgeable prompt learning. *Expert Systems with Applications*, *238*, 122177.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. doi: https://doi.org/10.1016/0364-0213(88)90023-7

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Waldon, B., Brodsky, M., Ma, M., & Degen, J. (2023). Predicting consensus in legal document interpretation. In M. B. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th annual meeting of the cognitive science society, cogsci 2023, sydney, nsw, australia, july 26-29, 2023*. cognitivesciencesociety.org.

Wu, S., Liu, Z., Zhang, Z., Chen, Z., Deng, W., Zhang, W., … Chen, Z. (2023). *fuzi.mingcha*. `https://github.com/irlab-sdu/fuzi.mingcha`. GitHub.

Wu, Y., Zhou, S., Liu, Y., Lu, W., Liu, X., Zhang, Y., … Kuang, K. (2023, December). Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 12060–12075). Singapore: Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.740

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., … others (2018). Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Yue, S., Chen, W., Wang, S., Li, B., Shen, C., Liu, S., … others (2023). Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020, July). How does NLP benefit legal system: A summary of legal artificial intelligence. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5218–5230). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.466