Information Processing and Management 63 (2026) 104319

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A multi-agent framework with legal event logic graph for multi-defendant legal judgment prediction

Weikang Yuan [a], Kaisong Song [b], Zhuoren Jiang [a,*], Junjie Cao [b], Yujie Zhang [a], Chengyuan Liu [a], Jun Lin [b], Ji Zhang [b], Kun Kuang [a], Xiaozhong Liu [c]

[a] *Zhejiang University, Hangzhou, China*
[b] *Tongyi Lab, Alibaba Group, Hangzhou, China*
[c] *Worcester Polytechnic Institute, USA*

ARTICLE INFO

Keywords:
Multi-defendant LJP
Large language models
Legal event logic graph
Multi-agent collaboration

ABSTRACT

Multi-defendant Legal Judgment Prediction (LJP) is a complex and challenging task in real-world legal scenarios. Existing approaches often struggle with analyzing intricate relationships among defendants and incorporating domain-specific legal expertise, particularly in penalty prediction. To address these challenges, we propose MAGLJP, a novel Multi-Agent framework with Legal Event Logic Graph for multi-defendant LJP. Our framework systematically decomposes the task into a Standard Operating Procedure, employing three specialized LLM-based agents: the Conviction Agent for law article and charge prediction, the Legal Knowledge Assistant Agent for legal knowledge integration, and the Sentencing Agent for penalty prediction. To support legal reasoning and effectively integrate domain knowledge, we introduce the Legal Event Logic Graph (LELG), a directed acyclic graph structure designed to represent and infer the complex relationships among criminal facts, legal knowledge, and sentencing outcomes. Additionally, we construct a comprehensive legal knowledge base that incorporates multiple levels of judicial interpretation. Extensive experiments on two benchmark multi-defendant LJP datasets show that MAGLJP significantly outperforms strong baselines, achieving state-of-the-art performance across all evaluation metrics. Tests conducted across diverse scenarios and case studies further demonstrate the robustness, generalization ability, and interpretability of MAGLJP, highlighting its strong applicability in real-world judicial settings.

1. Introduction

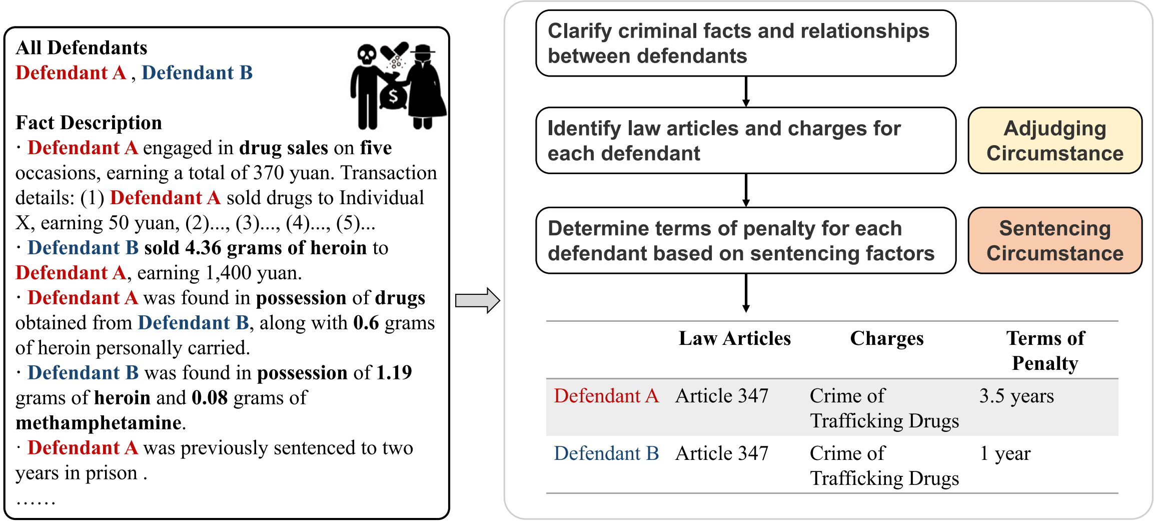
The judicial system serves as the key mechanism for safeguarding fundamental rights and maintaining social equity. However, judicial systems are facing intensifying pressures from explosive case growth versus limited judicial expertise. In China, there were approximately 130,000 judges handling over 46 million adjudicated cases in 2024, resulting in an average annual caseload of 354 cases per judge.[1] This substantial disparity threatens the efficiency and sustainability of legal systems and highlights the urgent need for technological support. Similar challenges have also been observed in various legal systems globally (Cui, Shen, & Wen, 2023), where judicial resources are strained under increasing caseloads and legal complexity.

Legal Artificial Intelligence (Legal AI) has emerged as a promising solution to assisting legal professionals by automating repetitive tasks and improving overall efficiency. At the same time, it enhances the accessibility and equity of legal services by

* Correspondence to: School of Public Affairs, Department of Information Resources Management, Zhejiang University, Hangzhou 310058, China.
E-mail address: jiangzhuoren@zju.edu.cn (Z. Jiang).

[1] Refer to China Court Net, available at https://www.chinacourt.org/article/detail/2025/03/id/8731367.shtml

https://doi.org/10.1016/j.ipm.2025.104319
Received 14 April 2025; Received in revised form 16 July 2025; Accepted 21 July 2025
Available online 8 August 2025



All Defendants
Defendant A , Defendant B

Fact Description
· Defendant A engaged in drug sales on five occasions, earning a total of 370 yuan. Transaction details: (1) Defendant A sold drugs to Individual X, earning 50 yuan, (2)..., (3)..., (4)..., (5)...
· Defendant B sold 4.36 grams of heroin to Defendant A, earning 1,400 yuan.
· Defendant A was found in possession of drugs obtained from Defendant B, along with 0.6 grams of heroin personally carried.
· Defendant B was found in possession of 1.19 grams of heroin and 0.08 grams of methamphetamine.
· Defendant A was previously sentenced to two years in prison .
……

Clarify criminal facts and relationships between defendants
Identify law articles and charges for each defendant
Adjudging Circumstance
Determine terms of penalty for each defendant based on sentencing factors
Sentencing Circumstance

	Law Articles	Charges	Terms of Penalty
Defendant A	Article 347	Crime of Trafficking Drugs	3.5 years
Defendant B	Article 347	Crime of Trafficking Drugs	1 year

Fig. 1. An illustration of multi-defendant LJP. Generally, a judge needs to reason on the fact description to clarify the complex interactions among different defendants and make accurate judgments for each defendant.

providing non-experts with reliable legal references and affordable support options (Zhong et al., 2020). With the advancement of Large Language Models (LLMs), significant progress has been made in areas such as text comprehension, complex reasoning, and text generation (Chowdhery et al., 2023; OpenAI, 2023; Touvron et al., 2023). These capabilities have facilitated Legal AI applications such as similar case retrieval (Sun, Zhang, et al., 2024), judgment prediction (Wu et al., 2023; Yuan et al., 2024), document summarization & generation (Deroy et al., 2023), and legal consultation services (Louis et al., 2024).

Legal Judgment Prediction (LJP) is one of the most common tasks in judicial decision-making. It aims to predict final judgment outcomes based on given case facts. The LJP task primarily involves determining two main parts: the Adjudging Circumstance, which predicts the corresponding verdicts (i.e., law articles and charges), and Sentencing Circumstance, which determines the terms of penalty (Yue et al., 2021). In practice, a substantial proportion of real-world criminal cases involve multiple defendants and multiple charges.[2] However, most existing LJP approaches focus solely on single-defendant scenarios and are ill-equipped to handle the increased procedural and reasoning complexity posed by multi-defendant cases. As illustrated in Fig. 1, multi-defendant LJP differs fundamentally from traditional single-defendant LJP in that it requires disentangling each defendant's individual criminal facts and clarifying their relationships to others, followed by making separate legal predictions for each individual.

The task of multi-defendant legal judgment prediction (LJP) presents two primary challenges:

Increased procedural complexity. In multi-defendant cases, the fact descriptions of different defendants are usually presented together in one narrative. However, the charges and penalties must be individually determined based on each defendant's specific circumstances. To ensure fair adjudication, judges must accurately identify the relationships among defendants and their respective roles in the case. This process necessitates a deep understanding of judicial decision-making logic and the ability to model the complex chains of reasoning involved (Lyu et al., 2023).

Insufficient integration of legal knowledge affecting subtask performance. LJP comprises three core subtasks: law article prediction, charge prediction, and penalty prediction. Among them, penalty prediction has consistently proven to be the most challenging, as evidenced by numerous studies in the field (Lyu et al., 2023; Yue et al., 2021; Zhong et al., 2018). This difficulty largely results from the complex and context-dependent nature of sentencing decisions, which involve numerous factors such as the nature and social harm of the crime, degree of involvement (principal or accomplice), the offender's identity (e.g., minors are often given lighter sentences), and other circumstances subject to judicial discretion. Moreover, legal knowledge relevant to sentencing is often fragmented and dispersed across various sources, including statutory provisions and interpretive guidelines, making it difficult for general models to incorporate and reason over effectively. These challenges highlight the urgent need for more structured and efficient methods to integrate legal knowledge, particularly to improve the accuracy and reliability of penalty prediction.

To address the above challenges, we propose a Multi-Agent framework with Legal Event Logic Graph for Multi-Defendant Legal Judgment Prediction (MAGLJP).

To tackle the first challenge, we conceptualize the multi-defendant LJP task as a Standard Operating Procedure (SOP), which includes three components, (1) Charge and Law Article Prediction, (2) Legal Event Logic Graph Generation, and (3) Sentence Prediction. Consequently, we leverage large language models (LLMs) to develop three agents: the Conviction Agent, the Legal Knowledge Assistant Agent, and the Sentencing Agent. These agents collaboratively address the multi-defendant LJP task, reflecting

[2] Based on empirical research examining disclosed legal documentation, nearly 30% of cases involve multi-defendants (Pan et al., 2019).

the reasoning logic of legal experts as they analyze cases and reaching judgment decisions. Specifically, the Conviction Agent is responsible for predicting both applicable charges and law articles based on the defendants' criminal facts. Based on these adjudging circumstances, the Legal Knowledge Assistant Agent retrieves relevant legal knowledge and constructs a Legal Event Logic Graph (LELG). Finally, the Sentencing Agent integrates the criminal facts, the predicted charges and legal articles from the Conviction Agent, and the LELG from the Legal Knowledge Assistant Agent to more accurately predict the defendants' terms of penalty.

To address the second challenge, we propose a two-level solution. **Firstly**, to tackle the complexity of domain-specific legal knowledge required for penalty prediction, we have constructed a comprehensive legal knowledge base. This knowledge base integrates authoritative data from Chinese Criminal Law, guidelines from the Supreme People's Court, and directives from provincial high courts across most regions. This resource serves to clarify inherent ambiguities in legal rules and definitions. For instance, while Chinese Criminal Law provides general sentencing ranges (e.g., "3-7 years imprisonment for severe drug trafficking circumstances"), it often lacks specific quantitative criteria. The precise thresholds, such as the amount of drugs that constitutes a particular severity level, are typically detailed in supplementary legal documents, which our knowledge base comprehensively incorporates. **Secondly**, we employ a Legal Event Logic Graph (LELG) to efficiently utilize legal knowledge. Specifically, the LELG is designed to represent and reason about causal relationships in legal cases. It is structured as a Directed Acyclic Graph (DAG), comprising nodes and edges that map the relationships between criminal facts, legal knowledge, and penalty outcomes. This graph-based structure enables more effective representation and reasoning of complex legal information, facilitating more accurate and interpretable legal judgment prediction. By fine-tuning the Conviction and Sentencing Agent, we have substantially enhanced their performance in legal judgment prediction.

Our contributions can be summarized as follows:

1. Multi-agent framework for multi-defendant LJP: We propose MAGLJP, the first multi-agent framework with Legal Event Logic Graph that systematically formalizes multi-defendant LJP task. We address this task as a Standard Operating Procedure, utilizing multi-agents to collaboratively address the problem. This approach breaks down complex procedural multi-defendant LJP task by emulating legal experts' operational logic in complex case analysis. MAGLJP not only enhances prediction performance but also provides valuable procedural insights for future developments in legal judgment prediction.

2. LELG for efficient legal knowledge integration: We introduce the LELG, a novel directed acyclic graph-based structure that explicitly models the causal relationships among criminal facts, legal knowledge, and sentencing outcomes. The graph's three-tier architecture achieves superior interpretability through hierarchical knowledge representation while enhancing prediction performance. It offers a new perspective for the efficient integration of legal knowledge in future research on LJP and explainable legal AI systems.

3. Comprehensive legal knowledge base for enhanced judgment prediction: We meticulously construct a knowledge base of judicial interpretations at various levels, related to different charges. This knowledge base extends and complements the legal knowledge traditionally derived solely from law articles regarding charges and sentencing. This comprehensive knowledge base provides fine-grained legal insights that enhance LLMs. It represents a valuable resource that facilitates future research in legal judgment prediction, particularly in the challenging domains of precise penalty prediction.

4. Empirical validation of model performance: We validated our proposed MAGLJP on two multi-defendant LJP datasets, and MAGLJP achieved state-of-the-art results across the three sub-tasks of LJP. Specifically, on the MultiLJP dataset, our model improved the predictions F1-score for law articles, charges, and penalties by 36.87%, 44.62%, and 49.67%, respectively. On the CAIL2024-DRDZ dataset, the improvements were 18.07%, 17.46%, and 47.12%, respectively. These significant improvements demonstrate the effectiveness and robustness of our model, offering a flexible path to improve the performance of LLMs in domain-specific applications.[3]

The rest of the paper is organized as follows. Section 2 reviews the existing LJP methods from previous studies, highlighting the differences of our work. Sections 3 and 4 introduce the preliminaries of multi-defendant LJP and the core idea of our framework. Section 5 presents the baselines, data, experimental setting, and results. Finally, in Section 6, we conclude the paper.

2. Related work

2.1. Legal judgment prediction

2.1.1. Single-defendant and multi-defendant LJP

Legal Judgment Prediction (LJP) has been studied in both case law countries (Katz et al., 2017; Malik et al., 2021) and statutory law countries (Feng et al., 2022). Most research has focused primarily on single-defendant cases, as multi-defendant cases are generally skipped due to their complexity. Early research centered around rule-based systems (Kort, 1957; Segal, 1984), which struggled with the complexity of real-world cases, especially when multiple legal factors were involved. The advent of machine learning ushered in a second paradigm shift (Liu & Chen, 2018; Sulea et al., 2017), but shallow models failed to capture the semantic details of lengthy legal documents. Recent advancements in deep learning have changed this landscape. In the realm of single-defendant LJP, studies have focused on multi-task learning (Yang et al., 2019; Zhong et al., 2018), leveraging legal knowledge (Yue et al., 2021), and using pre-trained language models (Chalkidis et al., 2020; Xiao et al., 2021).

[3] The resource is available at https://github.com/yuanwk99/LegalKnowBase_MAGLJP.

For multi-defendant LJP, there has been limited research. MAMD (Pan et al., 2019) introduced a multi-scale attention model for charge prediction in multi-defendant cases, focusing on specific defendants based on their names and fact descriptions. MUD (Wei et al., 2024) introduced a new benchmark for cases involving multiple defendants, incorporating crime elements and legal rules for charge prediction. However, both MAMD and MUD are limited to charge prediction and do not address tasks such as law article prediction and penalty prediction, which to some extent limits its generalizability in LJP tasks. HRN (Lyu et al., 2023) followed hierarchical reasoning chains to determine criminal relationships, sentencing circumstances, law articles, charges, and penalties for each defendant and developed a multi-defendant LJP dataset called MultiLJP. CMDL (Huang et al., 2024) is another multi-defendant LJP dataset, highlighting that existing state-of-the-art methods struggle with cases involving multiple defendants.

2.1.2. Incorporating legal knowledge in LJP

Given the importance of legal knowledge in judicial decisions, several studies have attempted to integrate legal knowledge into models to improve performance. Luo et al. (2017) used legal articles as a knowledge base, employing attention mechanisms to aggregate relevant article representations to support judgment predictions. Xu et al. (2020) explored extracting distinctive knowledge from similar law articles using a graph-based method. Gan et al. (2021) proposed representing declarative legal knowledge as first-order logic rules, integrating these into a co-attention network-based model. Zhao et al. (2023) employed a multi-graph fusion mechanism to integrate legal article information. Li et al. (2025) proposed a contrastive knowledge fusion module to inject external statute knowledge into fact description embeddings. However, the knowledge primarily focuses on charge predictions and lacks detail for accurate penalty predictions.

In summary, previous studies on LJP tend to focus only on single-defendant scenarios and are often insufficient for multiple defendants. While some approaches leverage graphs or text to integrate legal knowledge, such knowledge is typically derived from law articles and may be insufficient for accurate penalty prediction. Furthermore, these approaches may fail to consider explicit causal interactions between defendants and knowledge, potentially undermining the predictive performance and reliability of LJP tasks. To tackle these limitations, we (1) construct a comprehensive legal knowledge base that encompasses a broader spectrum of judicial knowledge, and (2) introduce a Legal Event Logic Graph that explicitly models the interconnections between defendant case facts and legal knowledge to improve predictive capabilities.

2.2. LLM-based multi-agent collaboration framework

Large Language Models (LLMs) show great promise in legal contexts due to their ability to understand and generate text. Consequently, LLM-driven agents have been studied and developed rapidly, aiding complex decision-making across various contexts (Guo et al., 2024). Multi-agent collaboration reflects human cooperation in solving complex tasks, with applications in software development (Hong et al., 2023), mathematics, coding (Wu et al., 2024), and medical reasoning (Tang et al., 2024). In the legal domain, multi-agent frameworks have been employed in legal consultations (Cui, Li, et al., 2023; Sun, Dai, Luo, et al., 2024), confusing charge prediction (Yuan et al., 2024), and court simulation (He et al., 2024). However, there is still a notable gap in the exploration of multi-agent collaboration frameworks for multi-defendant LJP.

Our innovation lies in transforming multi-defendant LJP into a structured process, akin to a Standard Operating Procedure (SOP). By emulating legal experts' decision-making processes and designing multiple collaborative agents, we propose MAGLJP, a multi-agent collaboration framework to standardize workflows and decompose complex legal reasoning task, thereby improving prediction effectiveness.

3. Preliminaries

The multi-defendant legal judgment prediction task aims to predict the judgment results of multiple applicable law articles, charges and a term of penalty for each defendant. So the task can be divided into three subtasks: charge prediction, law article prediction and penalty prediction. The prediction of law articles and charges can be served as multi-label classification tasks, and term of penalty prediction is a multi-class classification task.

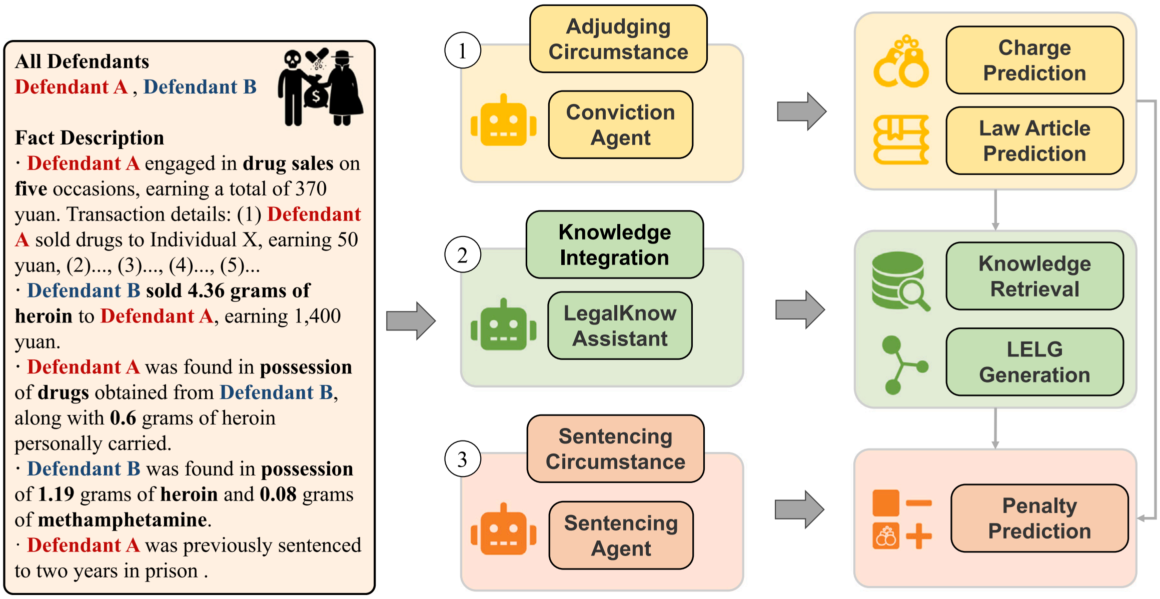
Formally, given the case fact (criminal fact) x from a real case and the set of involved defendants $\mathcal{D} = \{d_1, d_2, \dots, d_l\}$, where l is the number of defendants. The LJP task includes a set of charges $\mathcal{C}$, a set of law articles $\mathcal{A}$, and terms of imprisonment (penalty) $\mathcal{T}$ which are typically divided into several time intervals. The prediction model $\mathcal{M}$ aims to predict the applicable charges, law articles and the term of penalty for each defendant d based on the case fact x:

$$y_d^a, y_d^c, y_d^t = \mathcal{M}(x) \tag{1}$$

where y_d^a is a subset of $\mathcal{A}$, y_d^c is a subset of $\mathcal{C}$, and y_d^t is one of the classes of $\mathcal{T}$.

Sub-tasks Dependencies: In LJP, the prediction sub-tasks exhibit strong interdependencies. The identification of applicable law articles and charges traditionally precedes and informs sentencing decisions, following established legal principles. However, most existing studies treat these tasks as independent components, predicting each element separately as shown in Eq. (1). While some pioneering works like TopJudge (Zhong et al., 2018) have acknowledged and incorporated the topological dependencies among these sub-tasks, they do not explicitly leverage legal knowledge in their prediction process. This limitation is particularly evident in penalty prediction, where domain-specific legal expertise plays a crucial role in determining appropriate sentences.

In our approach to the LJP task, we introduce two key procedural enhancements: (1) We streamline the task process by clearly defining the logical relationships between three subtasks and define the Multi-defendant LJP as a multi-agent collaboration



All Defendants
Defendant A, **Defendant B**

Fact Description
· **Defendant A** engaged in **drug sales** on **five** occasions, earning a total of 370 yuan. Transaction details: (1) **Defendant A** sold drugs to Individual X, earning 50 yuan, (2)..., (3)..., (4)..., (5)...
· **Defendant B sold 4.36 grams of heroin** to **Defendant A**, earning 1,400 yuan.
· **Defendant A** was found in **possession** of **drugs** obtained from **Defendant B**, along with **0.6** grams of heroin personally carried.
· **Defendant B** was found in **possession** of **1.19** grams of **heroin** and **0.08** grams of **methamphetamine**.
· **Defendant A** was previously sentenced to two years in prison .

Fig. 2. Overview framework of MAGLJP. Given case facts involving multiple defendants, the framework processes legal judgment through three stages: (1) The Conviction Agent analyzes adjudging circumstances to predict charges and law articles, (2) The LegalKnow Assistant performs knowledge retrieval and LELG generation, and (3) The Sentencing Agent considers sentencing circumstances to predict appropriate penalties.

framework. (2) We identify an efficient way to assist predictive models in leveraging external legal knowledge for more accurate penalty prediction.

Our framework employs three specialized agents. First, a Conviction Agent $\mathcal{M}_{conv}$ predicts each defendant's charges and relevant law articles based on the facts (detailed in Section 4.2). Then, building upon these predictions, we explore an efficient way to incorporate relevant legal knowledge to aid in more accurate penalty prediction. Specifically, we define a Legal Knowledge Assistant Agent (LegalKnow Assistant) $\mathcal{M}_{legalknow}$ to retrieve relevant legal knowledge and generate a Legal Event Logic Graph (detailed in Section 4.3). Finally, a sentencing agent $\mathcal{M}_{sent}$ incorporates the LELG information to make accurate penalty predictions (detailed in Section 4.4). This process can be formalized as shown in Eq. (2):

$$\begin{cases} y_d^a, y_d^c = \mathcal{M}_{conv}(x) \\ LELG_f = \mathcal{M}_{legalknow}(x, K, y_d^a, y_d^c) \\ y_d^t = \mathcal{M}_{sent}(x, LELG_f) \end{cases} \tag{2}$$

where K can represent any external legal knowledge base, such as statutory provisions, legal theories and doctrines. In this work, we specifically utilize a comprehensive legal knowledge base that we construct as K.

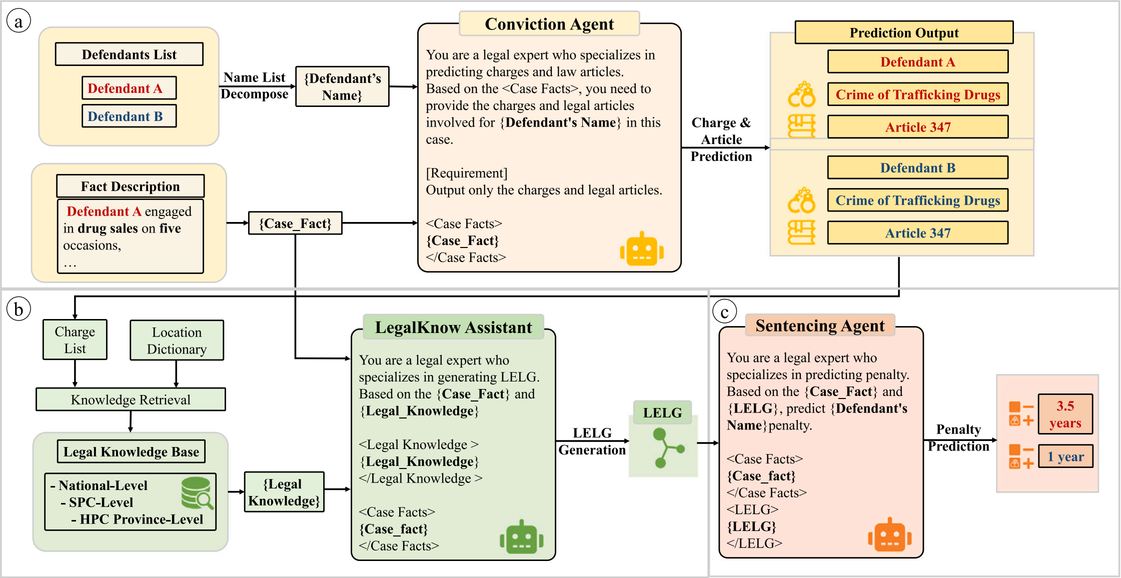
4. The proposed framework: MAGLJP

In this section, we will introduce our proposed model: Multi-Agent framework with Legal Event Logic Graph for Multi-Defendant Legal Judgment Prediction (MAGLJP). We first provide an overview of the overall framework, followed by detailed descriptions of each component's operational mechanisms.

4.1. Overall framework

The overall framework is illustrated in Fig. 2. Drawing inspiration from the judicial reasoning process employed by legal experts, we conceptualize multi-defendant legal judgment prediction as a three-stage Standard Operating Procedure (SOP):

1. **Adjudging Circumstance Analysis**: The Conviction Agent analyzes case facts and adjudging circumstances to predict applicable charges and corresponding law articles for each defendant.
2. **Legal Knowledge Integration**: The LegalKnow Assistant performs knowledge retrieval and Legal Event Logic Graph (LELG) generation to integrate domain-specific legal expertise into the reasoning process.
3. **Sentencing Prediction**: The Sentencing Agent evaluates sentencing circumstances and leverages the integrated legal knowledge to make penalty predictions.



Defendant A engaged in drug sales on five occasions, …

You are a legal expert who specializes in predicting charges and law articles. Based on the <Case Facts>, you need to provide the charges and legal articles involved for {Defendant's Name} in this case.

[Requirement]
Output only the charges and legal articles.

<Case Facts>
{Case_Fact}
</Case Facts>

You are a legal expert who specializes in generating LELG. Based on the {Case_Fact} and {Legal_Knowledge}

<Legal Knowledge >
{Legal_Knowledge}
</Legal Knowledge >

<Case Facts>
{Case_fact}
</Case Facts>

You are a legal expert who specializes in predicting penalty. Based on the {Case_Fact} and {LELG}, predict {Defendant's Name}penalty.

<Case Facts>
{Case_fact}
</Case Facts>
<LELG>
{LELG}
</LELG>

Fig. 3. Workflow of MAGLJP, showcasing the specific prompting strategies and work flow between components. The diagram illustrates (a) the prompt-based conviction analysis with defendant decomposition, (b) the legal knowledge integration and LELG generation process, and (c) the structured penalty prediction mechanism.

This procedural decomposition serves as the foundation for our proposed MAGLJP framework, which consists of three specialized yet interconnected components: (1) **Conviction Agent** for law article & charge prediction, (2) **LegalKnow Assistant** for legal knowledge integration and LELG generation, and (3) **Sentencing Agent** considers sentencing circumstances to predict appropriate penalties.

4.2. Conviction agent

In this section, we introduce the Conviction Agent $\mathcal{M}_{conv}$, which serves as the foundational component for identifying applicable charges ($\mathcal{C}_d$) and law articles ($\mathcal{A}_d$) based on criminal facts. Given the nature of multi-defendant LJP task, followed previous works (Lyu et al., 2023; Pan et al., 2019), we first decompose the defendant list $\mathcal{D} = \{d_1, \dots, d_l\}$ to enable individual predictions for each defendant. Formally, for each defendant d_i, the agent $\mathcal{M}_{conv}$ processes the concatenated input of defendant name name_d and associated factual descriptions to generate predictions. To enhance task-specific reasoning, we design a legal expert persona prompt p_{conv} containing role definition, task instructions and output format specifications. The workflow of the $\mathcal{M}_{conv}$ can be formally expressed as:

$$y_d^a, y_d^c = \mathcal{M}_{conv}(x, \text{name}_d, p_{\text{conv}}; \Theta_{conv}) \tag{3}$$

where Θ_{conv} is the parameters of the Conviction Agent $\mathcal{M}_{conv}$.

The agent $\mathcal{M}_{conv}$ should possess the capability to identify relevant actions and determine corresponding charges and law articles for each defendant based on their name. In our implementation, we utilize Qwen2.5-7B-Instruct (Yang et al., 2024) as the backbone model for $\mathcal{M}_{conv}$, which is an instruction-following open-source large language model pre-trained on over 18 trillion tokens, demonstrating strong fundamental capabilities and professional knowledge. Through supervised fine-tuning on our training dataset, we update the parameters Θ_{conv} of the conviction agent to enhance its instructions-following ability and prediction accuracy for charges and law articles effectively.

As illustrated in Fig. 3a, the conviction agent operates through the following workflow. The process begins with input comprising a list of defendants (e.g., Defendant A and Defendant B) and the associated case facts. After decomposing the defendant list, the agent processes each defendant's name alongside the case facts. Operating with a legal expert persona specialized in charge and law article prediction, the Conviction Agent analyzes the case details to determine appropriate legal outcomes for each defendant. The agent then generates structured predictions, assigning specific charges (such as "Crime of Trafficking Drugs") and corresponding legal articles (e.g., "Article 347") to each individual defendant.

4.3. LegalKnow assistant

We introduce the Legal Knowledge Assistant Agent (LegalKnow Assistant) in this section. In judicial practice, legal experts determine sentencing by first consulting relevant judicial knowledge, mapping atomic facts of criminal behavior (including both

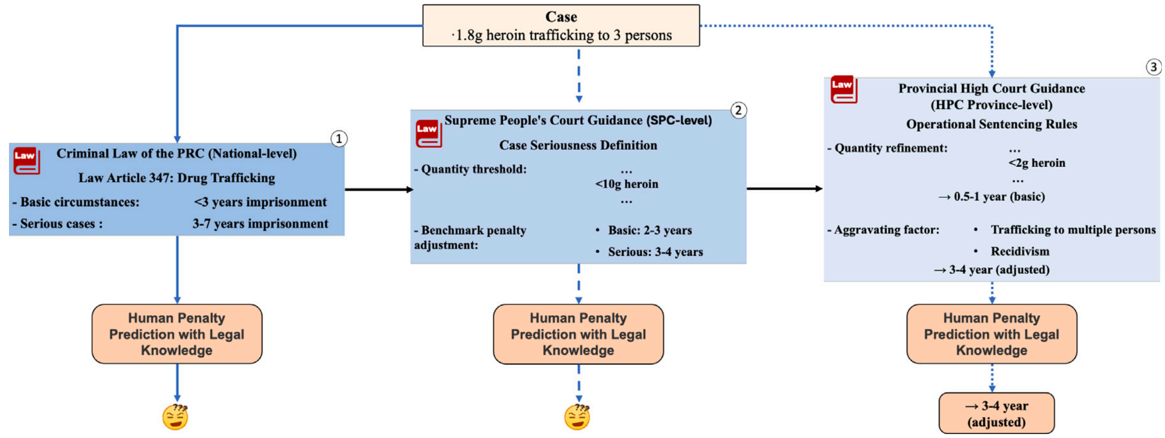

Case
·1.8g heroin trafficking to 3 persons
①
Law
Criminal Law of the PRC (National-level)
Law Article 347: Drug Trafficking
- Basic circumstances: <3 years imprisonment
- Serious cases : 3-7 years imprisonment
②
Law
Supreme People's Court Guidance (SPC-level)
Case Seriousness Definition
- Quantity threshold: ... <10g heroin ...
- Benchmark penalty adjustment: • Basic: 2-3 years • Serious: 3-4 years
③
Law
Provincial High Court Guidance (HPC Province-level)
Operational Sentencing Rules
- Quantity refinement: ... <2g heroin ...
→ 0.5-1 year (basic)
- Aggravating factor: • Trafficking to multiple persons • Recidivism
→ 3-4 year (adjusted)
Human Penalty Prediction with Legal Knowledge
Human Penalty Prediction with Legal Knowledge
Human Penalty Prediction with Legal Knowledge
→ 3-4 year (adjusted)

Fig. 4. Multi-Level Legal Framework for Drug Trafficking Penalties. The figure illustrates the integration of legal standards for a case involving 1.8 g of heroin trafficked to three persons in China. It starts with the national Criminal Law (Article 347) defining basic and serious offenses with penalties under three years and three to seven years, respectively. It moves to the Supreme People's Court's guidelines providing quantity thresholds and penalty adjustments. Finally, the Provincial High Court Guidance refines these rules, detailing adjustments for smaller quantities and aggravating factors such as multiple trafficking or recidivism. Human penalty predictions with legal knowledge are shown at each level, highlighting the effect of legal interpretation on sentencing.

individual sentencing factors and complex inter-defendant relationships) to corresponding sentencing circumstances from legal knowledge, ultimately arriving at an appropriate sentence for each defendant. To emulate this judicial reasoning logic, the Legal Knowledge Assistant agent (LegalKnow Assistant) $\mathcal{M}_{legalknow}$ serves two primary roles: First, the LegalKnow Assistant $\mathcal{M}_{legalknow}$ retrieves relevant legal knowledge that influences sentencing decisions based on case facts and the charges and legal articles predicted by the conviction agent $\mathcal{M}_{conv}$. Second, to establish connections between criminal facts and legal knowledge, the $\mathcal{M}_{legalknow}$ generates a Legal Event Logic Graph that maps these relationships, providing crucial auxiliary information for penalty prediction.

Given that existing legal resources, such as criminal law definitions, are often too broad to effectively differentiate between various circumstances and their corresponding penalties, we have meticulously constructed a comprehensive knowledge base incorporating legal knowledge and guidance from various levels of Chinese legal system. We will first introduce the construction of the Legal Knowledge Base and explain how the LegalKnow Assistant $\mathcal{M}_{legalknow}$ performs legal knowledge retrieval and generates the Legal Event Logic Graph.

4.3.1. Legal knowledge base construction

From a theoretical perspective, we formally define the legal knowledge base K as a tuple $K = (R, C, L)$, where R represents the set of geographical regions, C denotes the set of criminal charges and L is the set of legal texts including sentencing guidelines and interpretations. A mapping function $\phi : R \times C \to L$ that associates each region-charge pair with corresponding legal knowledge. This formal structure ensures the systematic organization and retrieval of legal knowledge.

Chinese criminal justice system operates through a multi-tiered legal knowledge architecture, where statutory provisions from the Criminal Law form only the foundational layer. In practice, sentencing decisions require synthesizing hierarchical judicial interpretations – from Supreme People's Court guidelines to provincial-level directives – that progressively refine abstract legal principles into operational rules.

Fig. 4 demonstrates the hierarchical structure of Chinese legal sentencing framework. At the national level, Article 347 of the Criminal Law establishes broad sentencing ranges (e.g., < 3 years for basic drug trafficking cases and 3–7 years for serious cases). However, these statutory provisions lack operational specificity for precise penalty determination. The Supreme People's Court's guidance introduces quantitative thresholds (e.g., < 10 g of heroin defining "serious cases") and benchmark penalty adjustments (2–4 years). This intermediate layer begins to operationalize the statutory language but remains insufficient for nuanced case handling. Provincial-level guidelines further refine these standards through three critical enhancements: (1) Lower quantity thresholds (< 2 g heroin), (2) Contextual aggravating factors (multiple recipients, recidivism), and (3) Differentiated sentencing brackets (0.5–1 vs. 3–4 years). This granularity exemplifies how multi-level legal knowledge integration enables precise penalty prediction — a defendant trafficking 1.8 g heroin to three recipients would receive 3–4 years under provincial rules, whereas national provisions alone could only suggest the broader 3–7 year range.

Construction details:

We construct a comprehensive legal knowledge base to directly address this challenge by providing detailed sentencing guidance. The implementation of the legal knowledge base consists of two main phases: **knowledge collection & extraction** and **structural organization**. Fig. 5 presents a structured formalization of the knowledge base construction process.

In the knowledge collection & extraction phase, we manually collected legal sentencing guidance documents from authoritative sources, including the Supreme People's Court and all provincial high courts across China. These official judicial documents follow a standardized structure, comprising: (1) general sentencing principles, (2) basic methodologies, (3) common circumstances, and (4)

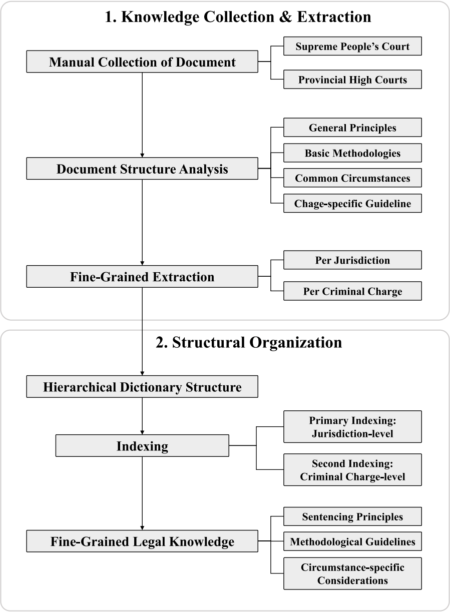

1. Knowledge Collection & Extraction
Manual Collection of Document
Supreme People's Court
Provincial High Courts
Document Structure Analysis
General Principles
Basic Methodologies
Common Circumstances
Chage-specific Guideline
Fine-Grained Extraction
Per Jurisdiction
Per Criminal Charge
2. Structural Organization
Hierarchical Dictionary Structure
Indexing
Primary Indexing: Jurisdiction-level
Second Indexing: Criminal Charge-level
Fine-Grained Legal Knowledge
Sentencing Principles
Methodological Guidelines
Circumstance-specific Considerations

Fig. 5. The construction process of the legal knowledge base.

charge-specific sentencing guidelines. This inherent structure facilitates systematic examination and extraction of legal information. Through careful examination of these structured documents, we extract fine-grained sentencing standards specific to each province-level jurisdiction and criminal charge combination. This systematic mapping ensures comprehensive coverage of charge-specific guidelines across different jurisdictions while preserving the nuanced standards for sentencing decisions.

The structural organization phase implements a hierarchical knowledge base. The knowledge base is architected as a multi-level hierarchical dictionary structure. In practice, we implement this knowledge base as a JSON-based file system, which provides a lightweight but efficient solution for our needs. Specifically, this hierarchical structure uses geographical jurisdiction and criminal charge as keys to map to their corresponding detailed sentencing guidelines. The primary indexing is based on jurisdictional levels (national and provincial) and secondary indexing by specific criminal charges. Each entry in this structure contains fine-grained legal knowledge that specifies the sentencing criteria and conditions for a particular charge within its jurisdiction, enabling efficient retrieval of relevant sentencing guidelines. Each entry in the knowledge base contains fine-grained legal information, including sentencing principles, methodological guidelines, and circumstance-specific considerations.

This flexible structure offers several advantages: (1) Efficient retrieval through province-charge key pairs for rapid integration with LLM analysis, (2) Clear hierarchical organization of geographical and charge-based relationships, and (3) High extensibility and maintainability for knowledge updates. The system enables swift access to relevant legal knowledge based on charge type and jurisdiction, facilitating precise and context-aware sentencing predictions.

This knowledge base integrates multi-level sentencing methodologies, application legal rules for common sentencing circumstances (e.g., sentence adjustments, roles in joint crimes), and jurisdiction-specific sentencing guidelines. Through careful organization into hierarchical levels, sentencing methodology types (such as baseline calculation and aggravating/mitigating factors), and charge-specific rules, the system facilitates efficient retrieval of relevant sentencing information based on criminal charges and jurisdictions.

4.3.2. Legal knowledge retrieval

Based on the legal knowledge base, $\mathcal{M}_{legalknow}$ implements rule-based legal knowledge retrieval according to the mapping function $\phi : R \times C \rightarrow L$, where jurisdiction R and criminal charges C serve as dual indices to retrieve related legal knowledge L. To ensure accuracy and comprehensiveness, we employ a systematic verification strategy:

1. Accurate Jurisdiction Identification: We develop a comprehensive province mapping dictionary that standardizes various forms of geographical references (standard names, abbreviations, and historical variations) in case facts. The system identifies the first provincial reference as the primary jurisdiction, ensuring accurate geographical indexing for knowledge retrieval.

2. Comprehensive Coverage Through Fallback Mechanism: To guarantee complete legal coverage, $\mathcal{M}_{legalknow}$ implements a hierarchical fallback strategy: (1) Defaults to nationally applicable "Sentencing Guidelines for Common Crimes (Trial)" issued by the Supreme People's Court when no explicit jurisdiction is mentioned; (2) References adjacent jurisdictions' knowledge when specific regional guidelines are unavailable. This ensures no cases lack applicable legal guidance.

3. Systematic Charge-based Retrieval: Using charges predicted by $\mathcal{M}_{conv}$, we implement iterative retrieval for multiple-charge cases, ensuring comprehensive coverage of all applicable sentencing rules.

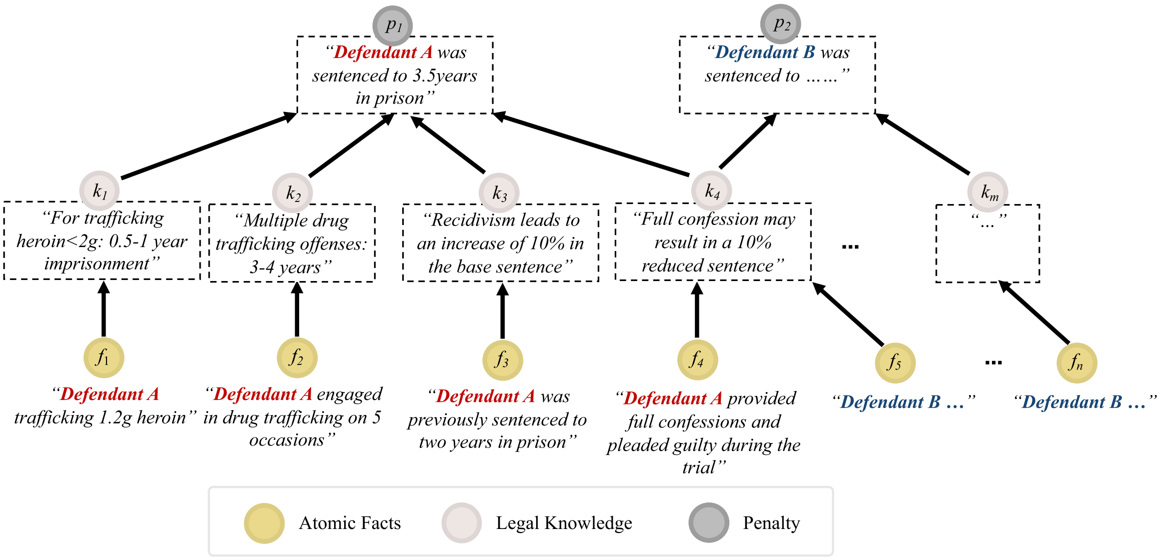

p_1
"Defendant A was sentenced to 3.5years in prison"
p_2
"Defendant B was sentenced to ……"
k_1
"For trafficking heroin<2g: 0.5-1 year imprisonment"
k_2
"Multiple drug trafficking offenses: 3-4 years"
k_3
"Recidivism leads to an increase of 10% in the base sentence"
k_4
"Full confession may result in a 10% reduced sentence"
...
k_m
"…"
f_1
"Defendant A trafficking 1.2g heroin"
f_2
"Defendant A engaged in drug trafficking on 5 occasions"
f_3
"Defendant A was previously sentenced to two years in prison"
f_4
"Defendant A provided full confessions and pleaded guilty during the trial"
f_5
"Defendant B …"
...
f_n
"Defendant B …"
Atomic Facts
Legal Knowledge
Penalty

Fig. 6. LELG Toy Example.

4.3.3. Legal event logic graph generation

To better capture the relationship between each defendant and legal knowledge, we propose the Legal Event Logic Graph (LELG), a specialized adaptation of event logic graph tailored for judicial reasoning. Traditional Event Logic Graphs (ELGs) serve as knowledge bases that describe patterns and rules of event evolution, representing these patterns through directed cyclic graphs where nodes represent events and edges denote various relationships (sequential, causal, conditional, or hierarchical) (Ding et al., 2019; Zhao et al., 2017). Our LELG differs fundamentally by employing a directed acyclic graph (DAG) structure specifically designed to capture causal dependencies in legal contexts. This structure formally maps the decision-making logic through three elements: criminal facts, legal knowledge and sentencing outcomes, providing an interpretable framework for legal knowledge representation.

LELG Definition:

Formally, the LELG is a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the directed edge set. Each node $v \in \mathcal{V}$ is formally represented as a tuple $v = (id, \tau, desc)$, where id is the node identifier, τ is the node type, and $desc$ is the text description of the node. The node type belongs to one of three types: (1) Atomic Fact Node (f), (2) Legal Knowledge Node (k) and (3) Penalty Node (p), denoted as $\tau \rightarrow \{f, k, p\}$. Each edge $e_{ij} \in \mathcal{E}$ connecting nodes v_i and v_j is typed by a relation attribute $\rightarrow \{\texttt{fact2knowledge}, \texttt{knowledge2penalty}\}$, encoding the causal relationships between nodes. `fact2knowledge` represents how factual evidence triggers legal knowledge, and `knowledge2penalty` maps Legal basis to sentencing decision. The toy example of LELG is shown in Fig. 6.

Therefore, the LELG adheres to the following key attributes. First, each LELG is designed to represent a complete multi-defendant case. Second, for each defendant, their associated fact descriptions, corresponding legal knowledge, and penalty results would form a connected subgraph within the larger structure. Third, the nodes in the LELG follow a strict sequential ordering from Atomic Fact Node through Legal Knowledge Node to Penalty Node, with penalty nodes serving as terminal nodes that cannot connect to any other nodes. Fourth, the graph encodes causal relationships, where edges from Atomic Fact Nodes to Legal Knowledge Nodes indicate how factual evidence aligns with specific legal knowledge, and edges from Legal Knowledge Nodes to Penalty Nodes demonstrate how these legal considerations contribute to specific sentencing outcomes.

LELG Generation:

The input to $\mathcal{M}_{legalknow}$ consists of the fact description of a case, the charges for each defendant involved in the case, the sentencing knowledge that impacts the sentencing corresponding to each charge, and the final sentences for each defendant. Through this analysis, the $\mathcal{M}_{legalknow}$ generates a case-level Legal Event Logic Graph (LELG) that captures the relationships between criminal facts, legal knowledge, and penalty outcomes. Formally, the generation process can be seen as:

$$LELG = \mathcal{M}_{legalknow}(x, K_d^c, y_d^a, y_d^c, p_{\text{legalknow}}, \hat{y}_d^t) \tag{4}$$

where $p_{legalknow}$ is the prompt for $\mathcal{M}_{legalknow}$. It is important to note that during the training process, the model benefits from access to known penalty values $\hat{y}_d^t$ for each defendant are known, which help better generation. During the test stage, the penalty values $\hat{y}_d^t$ are unknown and are set to None.

The LELG generation process requires $\mathcal{M}_{legalknow}$ to possess robust capabilities in legal knowledge integration, case information comprehension, and high-quality graph structure generation. In our implementation, we utilize the Qwen-max model as our foundation model, which excels in complex, multi-step reasoning tasks and demonstrates superior performance in sophisticated information processing. While we chose Qwen-max for its advanced capabilities, it is worth noting that $\mathcal{M}_{legalknow}$ can be instantiated with any large language model that exhibits strong text comprehension and generation abilities. Indeed, to demonstrate the generalizability of $\mathcal{M}_{legalknow}$, we conduct experiments using an alternative open-source LLM (Qwen2.5-32B-Instruct), which also achieved comparable performance (refer to Section 5.4.2).

The working principle illustration of the LegalKnow Assistant is shown in Fig. 3b. The detailed algorithm for LELG generation is presented in Algorithm 1.

Alg 1: LELG Generation

```
Initialize: LegalKnow Assistant: $\mathcal{M}_{legalknow}$
Trainset $\mathcal{D}_{\text{train}}$, Testset $\mathcal{D}_{\text{test}}$
if $stage = train$ then
    for $case$ in $\mathcal{D}_{train}$ do
        Get case fact $x$
        Get charges and legal knowledge $K_d^c$
        Get judgment results $\{y_d^a, y_d^c\}$
        Get ground truth penalties $\hat{y}_d^t$
        Generate LELG using $\mathcal{M}_{legalknow}(x, K_d^c, y_d^a, y_d^c, p_{\text{legalknow}}, \hat{y}_d^t)$
    end
else
    for $case$ in $\mathcal{D}_{test}$ do
        Get case fact $x$
        Get charges and legal knowledge $K_d^c$
        Get judgment results $\{y_d^a, y_d^c\}$
        Set $\hat{y}_d^t \leftarrow$ None
        Generate LELG using $\mathcal{M}_{legalknow}(x, K_d^c, y_d^a, y_d^c, p_{\text{legalknow}}, \hat{y}_d^t)$
    end
end
```

4.4. Sentencing agent

The Sentencing Agent $\mathcal{M}_{sent}$ predicts the terms of penalty by integrating multiple sources of information: the defendant's criminal facts, the charges and law articles predicted by the Conviction Agent, and the LELG generated by the LegalKnow Assistant. Formally, the workflow of the $\mathcal{M}_{sent}$ can be seen as:

$$y_d^t = \mathcal{M}_{sent}(x, \text{name}_d, p_{\text{sent}}, LELG; \Theta_{sent}) \tag{5}$$

where Θ_{sent} is the parameters of the Sentencing Agent $\mathcal{M}_{sent}$, and p_{sent} is the prompt of Sentencing Agent.

The working principle illustration of the Sentencing Agent is shown in Fig. 3c. Recent studies have shown that LLMs exhibit strong capabilities in understanding graph logic (Wang et al., 2024) and processing programming languages (Chen et al., 2023), which can enhance their reasoning abilities. Inspired by these findings, we structure the LELG generated by LegalKnow Assistant as programming code, consisting of a node dictionary and an edge list.

The node dictionary contains comprehensive information for each node, including a unique identifier (id), node type (τ), and textual description ($desc$). The edge list defines the logical relationships between nodes, following a structured reasoning pattern: atomic fact nodes (f) point to relevant legal knowledge nodes (k), which ultimately connect to penalty nodes (p). The programming language-based graph structure offers a clear representation of LELG, enabling the Sentencing Agent to effectively comprehend and systematically reason through the case information, from facts to legal principles and ultimately to penalty decisions.

Specifically, in our work, the $\mathcal{M}_{sent}$ is based on the Qwen2.5-7B-Instruct (Yang et al., 2024) as its backbone. Through supervised fine-tuning, we optimize the parameters of Θ_{sent} of the Sentencing Agent to enhance its ability to comprehend LELG content and better capture the relationships between legal reasoning and penalty prediction.

5. Experiments

In this section, we first describe our experimental setup, including datasets, baseline models, and implementation details. We then present a thorough evaluation of MAGLJP, encompassing experimental results, ablation studies, and comprehensive analyses that validate its effectiveness, advantages, and practical viability.

5.1. Dataset

To assess the performance of our proposed MAGLJP framework, we conduct experiments on two specialized multi-defendant LJP datasets: MultiLJP (Lyu et al., 2023) and CAIL2024-DRDZ (Huang et al., 2024). MultiLJP (Multi-defendant Legal Judgment

Table 1
Dataset statistics.

	MultiLJP	CAIL2024-DRDZ
Training set cases	18,968	12,000
Validation set cases	2379	1500
Testing set cases	2370	1500
Law articles	22	31
Charges	23	30
Terms of Penalty	11	14
Total defendants	80,477	48,815
Avg defendants	3.39	3.25

Prediction) (Lyu et al., 2023) dataset is constructed from the published legal documents in China Judgements Online.[4] Professional annotators are hired to manually produce law articles, charges, terms of penalty, criminal relationships, and sentencing circumstances for each defendant in multi-defendant cases. The MultiLJP contains 18,968 criminal cases with explicit annotations of inter-defendant relationships and charge allocations. CAIL2024-DRDZ is a high-quality dataset provided for the Chinese AI and Law Challenge 2024 (CAIL2024) DRDZ track,[5] containing 15,000 cases specifically curated for multi-defendant reasoning (Huang et al., 2024). We custom split it into 12,000 cases for training, 1,500 cases for validation and 1,500 for testing (8:1:1 ratio). Detailed statistics and characteristics of both datasets are presented in Table 1.

5.2. Baselines

To evaluate the effectiveness of our model MAGLJP in handling multi-defendant Legal Judgment Prediction (LJP), we select several baseline methods for comparison. Following previous evaluation frameworks, we primarily focus on the following baselines:

TopJudge (Zhong et al., 2018): This method employs a topological dependency learning framework tailored for single-defendant LJP. It formalizes explicit dependencies across subtasks using a directed acyclic graph.

NeurJudge (Yue et al., 2021): This approach leverages the outputs of intermediate subtasks to deconstruct the fact statement into distinct circumstances for single-defendant LJP, which are then utilized for predicting other subtasks.

BERT (Devlin et al., 2019): A bidirectional Transformer-based language model pre-trained on Chinese Wikipedia documents.

mT5 (Xue et al., 2021): This multilingual model is pre-trained by converting various language tasks into "text-to-text" tasks, incorporating Chinese datasets during training.

Lawformer (Xiao et al., 2021): Lawformer is a Transformer-based model and is pre-trained on a large-scale corpus of Chinese legal documents, particularly long case documents.

HRN (Lyu et al., 2023): HRN follows the hierarchical reasoning chains to determine criminal relationships, sentencing circumstances, law articles, charges, and terms of penalty for multi-defendant LJP.

Qwen2.5 SFT (Yang et al., 2024): Qwen2.5-7B-Instruct is an instruction-following open-source large language model pre-trained on over 18 trillion tokens, demonstrating strong fundamental capabilities and professional knowledge. We perform supervised fine-tuning (SFT) on Qwen2.5-7B-Instruct using our training data, treating multi-defendant LJP as a generation task.

Following (Lyu et al., 2023), we adapt single-defendant LJP baselines to handle multi-defendant cases by concatenating each defendant's name with their corresponding fact description as input, training the models to predict judgment results. However, the HRN model relies on manually annotated data, which is exclusively available in the MultiLJP dataset. Hence, we only compare HRN performance on MultiLJP, excluding it from the CAIL2024-DRDZ evaluation. Additionally, MAMD (Pan et al., 2019) is designed specifically for charge prediction tasks, making it incompatible with our comprehensive evaluation framework. Consequently, we exclude MAMD from our comparisons.

5.3. Experiment setting

In our experiments, model training was conducted utilizing four NVIDIA A100 GPUs, each equipped with 80 GB of memory. For MAGLJP, both the Conviction Agent and Sentencing Agent utilize Qwen-2.5-7B-Instruct as their backbone models. We employ a batch size of 256 and adopt the Low-Rank Adaptation (LoRA) strategy, with a rank value of 8 and an alpha parameter set to 16 for model fine-tuning. The maximum input sequence length is set to 4,096 tokens. The training process continues for up to 10 epochs using the AdamW optimizer with an initial learning rate of 5e-5, coupled with a cosine learning rate scheduler for learning rate decay. For baseline models, we follow the experimental settings as described in their respective original papers to ensure fair comparison.

To assess model performance, we remain consistent with the evaluation approach used in prior multi-defendant LJP tasks (Lyu et al., 2023). We employ four key metrics to evaluate the effectiveness of our approach: accuracy (Acc), Macro-Precision (P), Macro-Recall (R) and Macro-F1 (F1) scores.

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

[5] http://cail.cipsc.org.cn/Attend.html

Table 2
Main Results. Model performance comparison on two multi-defendant LJP datasets. Note: The symbol "–" indicates that HRN results are not available for CAIL2024-DRDZ, as the HRN model requires manually annotated data which is only available in the MultiLJP dataset.

Model	Article				Charge				Penalty			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
					MultiLJP							
TopJudge	69.32	35.60	39.13	36.93	64.42	24.96	35.28	28.34	28.36	23.16	22.25	22.00
NeurJudge	65.21	41.72	36.96	38.15	59.51	34.19	25.36	27.55	30.06	27.56	25.63	25.95
BERT	51.38	34.19	29.68	30.70	44.80	36.80	20.10	25.14	29.60	23.95	22.68	21.55
mT5	87.49	74.28	53.65	58.84	81.52	63.33	49.94	52.86	33.66	39.13	24.23	23.04
Lawformer	75.50	36.18	35.33	34.00	65.94	38.97	29.12	32.76	32.37	22.66	20.68	18.30
HRN	91.46	69.87	70.95	69.20	89.54	71.80	71.83	70.70	42.74	41.33	40.20	40.62
Qwen-sft	93.58	73.92	72.38	71.59	91.69	75.82	77.97	75.59	44.44	45.70	40.85	42.16
MAGLJP (Ours)	**94.72**	**77.48**	**79.17**	**76.8**	**92.67**	**81.27**	**83.52**	**80.72**	**50.62**	**56.65**	**53.77**	**54.96**
					CAIL2024-DRDZ							
TopJudge	54.68	62.91	52.27	53.06	57.71	63.48	55.44	55.75	24.08	9.92	11.03	9.24
NeurJudge	70.98	79.13	71.44	73.52	71.41	79.30	71.71	74.05	26.09	20.06	14.91	14.77
BERT	68.89	80.82	74.19	76.62	69.02	81.49	73.57	76.55	27.80	21.59	17.01	17.62
mT5	74.42	82.46	80.55	80.70	74.42	82.27	80.27	80.44	23.98	16.53	13.07	13.43
Lawformer	70.15	81.12	75.92	77.90	70.03	81.50	75.73	78.00	29.81	26.57	21.25	22.73
HRN	–	–	–	–	–	–	–	–	–	–	–	–
Qwen-sft	89.82	92.35	91.82	92.01	89.82	92.00	91.45	91.65	36.03	29.13	25.22	25.54
MAGLJP (Ours)	**90.33**	**92.58**	**92.17**	**92.32**	**90.33**	**92.37**	**92.10**	**92.17**	**39.49**	**36.51**	**30.91**	**32.57**

5.4. Result and analysis

5.4.1. Main results

As shown in Table 2, our proposed MAGLJP model achieved state-of-the-art performance on both MultiLJP and CAIL2024-DRDZ datasets, outperforming all baseline methods across three subtasks, demonstrating the effectiveness of our approach.

Specifically, on the **MultiLJP** dataset, for the Article prediction task, MAGLJP achieved the best performance with 94.72% accuracy, 77.48% precision, 79.17% recall, and 76.8% F1-score, surpassing all baseline models across all metrics. In the Charge prediction task, MAGLJP demonstrated superior performance with 92.67% accuracy, 81.27% precision, 83.52% recall, and 80.72% F1-score. For the Penalty prediction task, while MAGLJP achieved the highest accuracy (50.62%) and significantly outperformed other models in precision (56.65%), recall (53.77%), and F1-score (54.96%), the overall performance was lower than in the previous two tasks.

On the **CAIL2024-DRDZ** dataset: For Article prediction, MAGLJP maintained its superior performance with 90.33% accuracy, 92.58% precision, 92.17% recall, and 92.32% F1-score, with particularly impressive precision and recall rates. In Charge prediction, MAGLJP again achieved the best results with 90.33% accuracy, 92.37% precision, 92.10% recall, and 92.17% F1-score. For Penalty prediction, while MAGLJP still outperformed other models with 39.49% accuracy, 36.51% precision, 30.91% recall, and 32.57% F1-score, the lower absolute values indicate the task's inherent difficulty on this dataset.

Performance Improvements: Our model demonstrated substantial improvements across all subtasks. For Article Prediction, MAGLJP showed increases of 19.47% in accuracy and 36.87% in F1-score on MultiLJP, and improvements of 20.86% and 18.07% respectively on CAIL2024-DRDZ. In Charge Prediction, the model achieved gains of 23.32% in accuracy and 44.62% in F1-score on MultiLJP, with improvements of 20.22% and 17.46% on CAIL2024-DRDZ. Most notably, despite the inherent difficulty of Penalty Prediction, MAGLJP demonstrated remarkable improvements with 31.92% accuracy and 49.67% F1-score increases on MultiLJP, and 29.18% and 47.12% improvements on CAIL2024-DRDZ, highlighting the significant benefits of our approach even in challenging prediction tasks.

5.4.2. Ablation analysis

In our proposed framework MAGLJP, a multi-agent framework is to decompose the complex tasks and LELG effectively incorporate professional judicial knowledge to enhance Multi-defendant LJP performance. Specifically, the Conviction Agent was designed to improve charge and article predictions, while the LegalKnow Assistant and Sentence Agent were introduced to enhance penalty prediction through efficient legal knowledge integration. To validate the effectiveness of our proposed modules, we conducted ablation tests on the MultiLJP dataset (see Table 3).

For article and charge prediction tasks, we compared our model with a baseline that directly predicts all three subtasks without task decomposition (***w/o Conviction Agent***). The results showed significant performance drops, with the F1 score decreasing from 76.80% to 71.59% for article prediction and from 80.72% to 75.59% for charge prediction. These results validate the effectiveness of our Conviction Agent design, demonstrating that task decomposition effectively reduces the complexity of LJP reasoning tasks and improves prediction accuracy.

Our core contribution lies in effectively incorporating judicial knowledge to enhance penalty prediction. When removing the Multi-Agent framework (***w/o Multi-Agent***), the F1 score dropped substantially from 54.96% to 42.16%, representing the largest

Table 3
Ablation Test Results. Note: The symbol "–" indicates metrics that are not applicable for the specific ablation setting.

Model	Article				Charge				Penalty			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
HRN	91.46	69.87	70.95	69.20	89.54	71.80	71.83	70.70	42.74	41.33	40.20	40.62
Ablation for Article and Charge Prediction												
MAGLJP (Ours)	**94.72**	**77.48**	**79.17**	**76.80**	**92.67**	**81.27**	**83.52**	**80.72**	–	–	–	–
w/o Conviction Agent	93.58	73.92	72.38	71.59	91.69	75.82	77.97	75.59	–	–	–	–
Ablation for Penalty Prediction												
MAGLJP (Ours)	–	–	–	–	–	–	–	–	**50.62**	**56.65**	**53.77**	**54.96**
w/o Multi-Agent	–	–	–	–	–	–	–	–	44.44	45.70	40.85	42.16
w/o LELG	–	–	–	–	–	–	–	–	49.60	47.22	45.93	46.43
w/o Case Fact	–	–	–	–	–	–	–	–	44.62	45.30	44.34	44.67
w/o entire LELG	–	–	–	–	–	–	–	–	50.56	53.32	53.07	53.17
w Qwen2.5 Full	–	–	–	–	–	–	–	–	46.64	50.46	50.44	50.03
w Qwen2.5 Test	–	–	–	–	–	–	–	–	47.61	51.60	52.09	51.60

performance decrease in our ablation study. This significant reduction validates the effectiveness of task decomposition and the Sentencing Agent's role in handling complex reasoning tasks.

To validate whether the event logic graph is an efficient approach for legal knowledge incorporation, we conducted a comparative experiment. Instead of using LELG for knowledge organization, we directly supplemented the corresponding legal knowledge text from the Legal Knowledge Base as enhanced knowledge (***w/o LELG***). The results showed that the F1 score decreased significantly from 54.96% to 46.43%, representing a substantial drop of more than 8 percentage points. This notable decline demonstrates that LELG's structured representation plays a crucial role in explicitly connecting sentencing knowledge with criminal behaviors in case facts. The graph-based organization of legal knowledge proves to be more effective than simple text augmentation, further validating the effectiveness of our LegalKnow Assistant's approach to knowledge integration.

Since LELG captures the comprehensive mapping between multiple defendants' behaviors and corresponding legal knowledge within a complete case, we conducted an additional experiment to investigate whether individual defendant information is sufficient for penalty prediction. While the Sentencing Agent predicts penalties for specific defendants, we tested a variant using sub-graphs extracted from sub-LELG that only contained the behaviors and legal knowledge corresponding to the target defendant (***w/o entire LELG***). This experiment aimed to determine whether accurate penalty prediction requires only individual defendant information or benefits from complete case context. The results showed that using sub-graphs led to a slight performance decrease, with the F1 score dropping from 54.96% to 53.17%. This decline, though modest, reveals an important insight into multi-defendant cases: defendants' behaviors are often interdependent, with factors such as co-perpetration, primary offender status, and accessory roles significantly influencing sentencing decisions. The superior performance of complete LELG validates our approach of maintaining comprehensive case information and demonstrates the importance of considering the intricate relationships between multiple defendants when predicting penalties.

Finally, we investigated the importance of case facts by predicting sentences using only LELG information without case facts (***w/o Case Fact***). This resulted in a decrease from 54.96% to 44.67% in F1 score, indicating that while LELG effectively incorporates legal information, textual case descriptions remain vital. This finding suggests that LELG and case facts complement each other in improving prediction accuracy.

To further validate the robustness and generalizability of LELG generation, we conducted experiments using an open-source LLM, Qwen2.5-32B-Instruct, which offers a balanced trade-off between model capability and computational efficiency. We designed two experimental settings to evaluate the impact of the generated LELG quality. First, we trained and tested a new Sentencing Agent using LELG generated entirely by Qwen2.5-32B-Instruct (***w Qwen2.5 Full***), resulting in a slight performance decrease from 54.96% to 50.03% in F1 score. Second, we evaluated the cross-model generalization by using Qwen2.5-32B-Instruct to generate LELG only for the test set, while maintaining the Sentencing Agent trained on Qwen-max generated LELG (***w Qwen2.5 Test***). This setting showed a minimal performance drop to 51.60%. Besides, both ***w Qwen2.5 Full*** and ***w Qwen2.5 Test*** significantly outperform the methods without LELG structure integration (46.43% for ***w/o LELG***) or (42.16% for ***w/o Multi-Agent***). These results demonstrate that LELG's effectiveness is not strictly dependent on the specific LLM used for generation, suggesting good generalizability of our approach across different language models.

In sum, our extensive ablation studies on each core module have validated the effectiveness of MAGLJP. The experiments demonstrate that both the multi-agent framework and the legal event logic graph contribute substantially to model performance, while maintaining complete case information proves crucial for accurate multi-defendant judgment prediction. These findings comprehensively support the design choices in our proposed architecture.

5.4.3. Performance on different scenarios

To provide insights for future research directions, we conduct a comprehensive analysis of our MAGLJP's performance across various scenarios. Our investigation focuses on three aspects: (1) the impact of defendant numbers on model performance, (2) the model's performance in handling cases involving different criminal charges, and (3) performance on confusing charges.

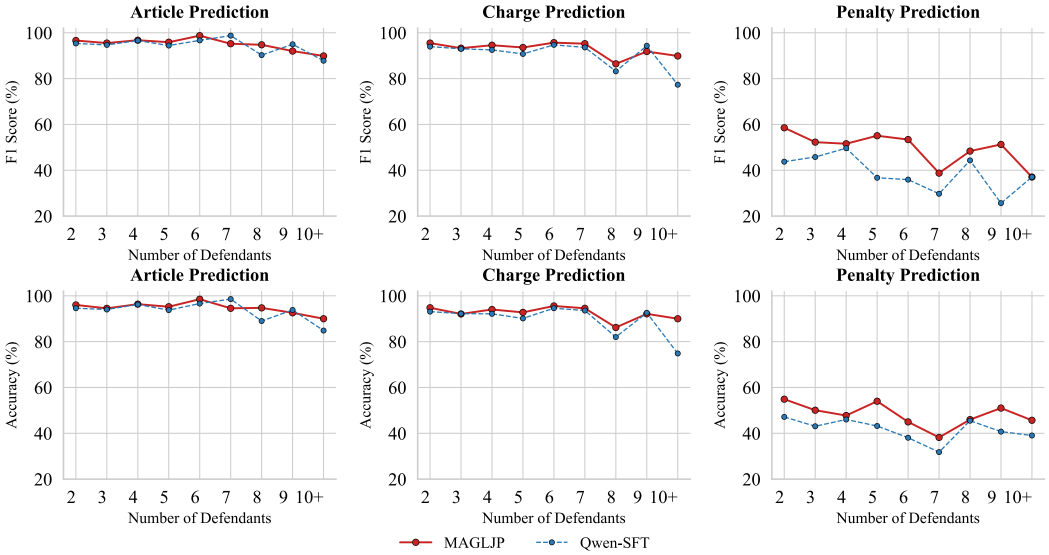


Fig. 7. Model Performance across Different Numbers of Defendants. The figure shows the F1 Score and Accuracy of our model in three sub-tasks (article, charge, and penalty prediction) as the number of defendants varies from 2 to 10+ (≥ 10).

(1) The Number of Defendants.

The increasing number of defendants in a case potentially elevates the complexity of legal judgment prediction. To investigate this relationship, we conducted a comparative analysis between our MAGLJP model and a strong baseline, Qwen-SFT, examining their performance as the number of defendants varies from 2 to 10+ (a case involves more than 10 defendants) in the MultiLJP dataset. Fig. 7 presents both F1 scores and accuracy metrics across three prediction tasks.

For Article and Charge prediction tasks, both models demonstrate remarkably stable and high performance, maintaining scores around 90%–95% regardless of the number of defendants involved. MAGLJP consistently outperforms Qwen-SFT by a small margin throughout this range. While there is a slight downward trend as the number of defendants increases, this degradation is minimal, suggesting that these tasks remain relatively robust even with increased case complexity. This stability indicates that models can effectively distinguish and process conviction circumstances for different defendants, even as their numbers grow.

However, the Penalty prediction task reveals more nuanced and interesting patterns. Both models exhibit significantly lower performance (40%–60%) compared to other tasks, underscoring the inherent challenge of penalty prediction. More notably, MAGLJP demonstrates a substantial advantage over Qwen-SFT, maintaining relatively stable performance as the number of defendants increases. In contrast, Qwen-SFT shows marked performance degradation with more defendants, particularly beyond seven defendants.

This performance gap in Penalty prediction can be attributed to two key advantages of MAGLJP: (1) the multi-agent framework effectively handles the increased complexity of cases with multiple defendants, and (2) the LELG structure successfully captures the intricate relationships between defendants' roles and their corresponding legal knowledge. The results particularly validate our model's effectiveness in managing complex cases with multiple defendants, where traditional approaches often struggle.

Notably, there is a slight performance decline for both models in cases with 10+ defendants. This suggests that as the number of defendants increases, the criminal facts affecting sentencing become increasingly complex, involving more intricate relationships, distinctions between principal and accessory offenders, and broader social impacts. These complex factors significantly increase the difficulty of accurate penalty prediction. Despite these challenges, our proposed MAGLJP demonstrates consistent superiority over baseline approaches. Notably, while the strongest baseline (Qwen-SFT) shows performance degradation as task difficulty increases, particularly in charge prediction, our model exhibits smaller performance drops, indicating enhanced robustness. This performance differential suggests promising directions for future research.

(2) Cases involving Different Charges.

To evaluate the robustness of LJP models across different criminal contexts, we conducted a comprehensive analysis of our MAGLJP model's performance on diverse charge types. Specifically, we analyze the performance of our proposed MAGLJP model across the ten most frequent criminal charges in the MultiLJP dataset. These charges include: *Crime of Opening a Casino* (OC), *Crime of Theft* (TF), *Crimes of Smuggling, Trafficking, Transporting and Manufacturing Drugs* (STTMD), *Crime of Picking Quarrels and Provoking Troubles* (PQPT), *Crime of Intentional Injury* (II), *Crime of Gambling* (GAM), *Crime of Affray* (AFF), *Crime of Fraud* (FRA), *Crime of Illegally Holding Drugs* (IHD), *Crime of Disrupting Public Affairs* (DPA). The detailed performance results across these charges are presented in Fig. 8.

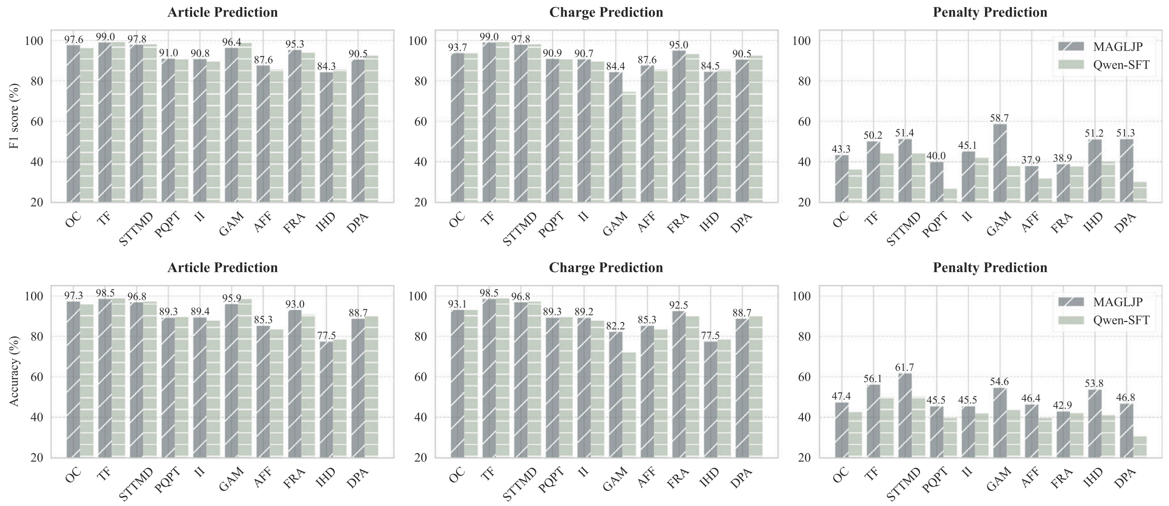


Fig. 8. Model Performance on Different Criminal Charges. The figure presents F1 scores and accuracy metrics for three sub-tasks (Article, Charge, and Penalty) across the ten most frequent charges in the MultiLJP dataset. The charges include: *Opening a Casino* (OC), *Theft* (TF), *Smuggling, Trafficking, Transporting and Manufacturing Drugs* (STTMD), *Picking Quarrels and Provoking Troubles* (PQPT), *Intentional Injury* (II), *Gambling* (GAM), *Affray* (AFF), *Fraud* (FRA), *Illegally Holding Drugs* (IHD), and *Disrupting Public Affairs* (DPA).

For Article prediction, our model demonstrates consistently high performance, with F1 scores ranging from 84.3% to 99.0% and accuracy from 77.5% to 98.5%. The model performs particularly well on cases involving Theft (TF: 99.0% F1) and Smuggling, Trafficking, Transporting and Manufacturing Drugs (STTMD: 97.8% F1). Slightly lower performance is observed in cases of Illegally Holding Drugs (IHD: 84.3% F1) and Affray (AFF 87.6%).

In Charge prediction, the model maintains robust performance with F1 scores between 84.4% and 99.0%. Similar to Article prediction, cases involving Theft (TF: 99.0% F1) and Smuggling, Trafficking, Transporting and Manufacturing Drugs (STTMD: 97.8% F1) show the highest performance. This consistency across Article and Charge prediction tasks indicates a strong correlation between these two aspects of legal judgment. However, our model exhibits relatively lower performance on certain charges, specifically Gambling (GAM: 84.4% F1), Illegally Holding Drugs (IHD: 84.5% F1), and Affray (AFF: 87.6% F1). This performance degradation may be attributed to the inherent similarity between these charges and their related offenses. For instance, there exists potential confusion between Gambling and Opening a Casino (GAM vs. OC), Illegally Holding Drugs and Smuggling, Trafficking, Transporting and Manufacturing Drugs (IHD vs. STTMD), and Affray and Intentional Injury (AFF vs. II). These pairs of charges often share similar contextual elements, making their distinction more challenging for the model.

The Penalty prediction task shows more significant variations across different charges, with F1 scores ranging from 37.9% to 58.7%. Overall, the performance on penalty prediction is substantially lower than that of the other two tasks, highlighting the particularly challenging nature of penalty prediction. Notably, the model performs best on Gambling cases (GAM: 58.7% F1) and Drug-related Crimes (STTMD: 51.4% F1), while showing lower performance on Fraud (FRA: 38.9% F1) and Affray (AFF: 37.9% F1). This variation likely reflects the inherent complexity in penalty determination, where factors such as circumstances, intent, and social impact play crucial roles. The lower performance in fraud cases might be attributed to the wide range of possible penalties depending on the amount involved and the sophistication of the scheme.

(3) Performance on Confusing Charges.

To further evaluate our model's discriminative capabilities, we examine its performance on confusing charges — pairs of criminal charges that are extremely similar. Specifically, we focus on two representative pairs from the MultiLJP dataset: *Opening a Casino* (OC) v.s. *Gambling* (GAM), and *Fraud* (FRA) v.s. *Contract Fraud* (COF). These pairs are particularly challenging as they share similar fundamental elements but differ in crucial details. For instance, while both gambling-related charges involve illegal gambling activities, *Gambling* typically refers to participation, whereas *Opening a Casino* involves organizational and operational aspects. Similarly, while both fraud types involve deceptive practices for financial gain, *Contract Fraud* specifically involves contractual relationships and typically larger monetary amounts (Yuan et al., 2024).

As shown in Fig. 9, MAGLJP demonstrates robust performance in distinguishing between these confusing charges, consistently outperforming the Qwen-SFT baseline across almost all three sub-tasks (Article, Charge, and Penalty prediction). Particularly noteworthy is the performance comparison on *Fraud* (FRA) versus *Contract fraud* (COF) cases. While Qwen-SFT achieves reasonable F1 performance on FRA cases (93.8% for article prediction, 93.2% for charge prediction, and 37.7% for penalty prediction), its performance drops significantly for COF cases, with F1 scores of only 45.1%, 41.8%, and 23.6% respectively. This substantial performance gap indicates Qwen-SFT's limitation in capturing subtle distinctions between confusing charges. In contrast, MAGLJP maintains consistently strong performance across both FRA and COF cases, demonstrating its superior ability to identify and utilize nuanced differences between similar charges for more accurate legal predictions.

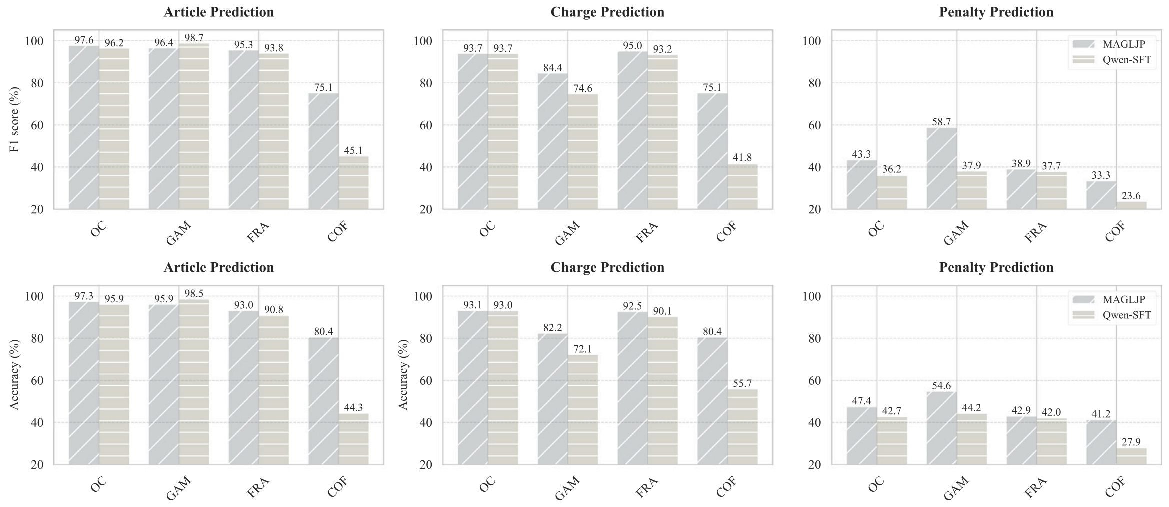
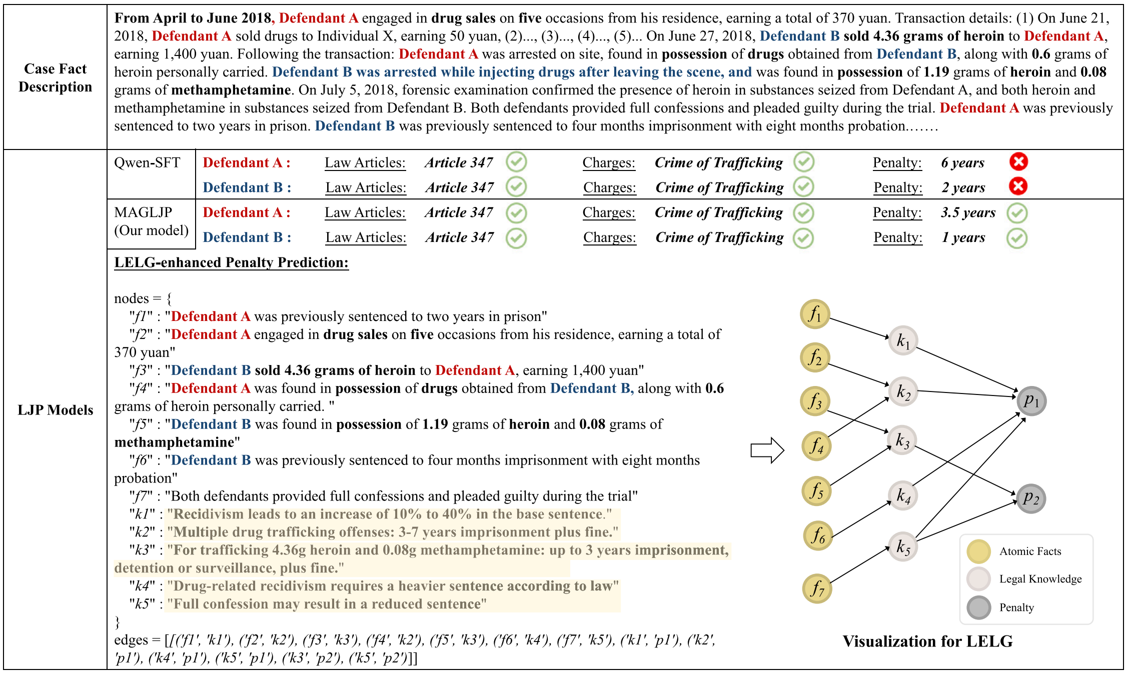


Fig. 9. Model Performance on Confusing Charges. The figure presents F1 scores and accuracy metrics for three sub-tasks (Article, Charge, and Penalty) across the two pairs of most confusing charges in the MultiLJP dataset. The charges include: *Opening a Casino* (OC) v.s. *Gambling* (GAM), *Fraud* (FRA) v.s. *Contract fraud* (COF).



Case Fact Description

From April to June 2018, Defendant A engaged in **drug sales** on **five** occasions from his residence, earning a total of 370 yuan. Transaction details: (1) On June 21, 2018, **Defendant A** sold drugs to Individual X, earning 50 yuan, (2)..., (3)..., (4)..., (5)... On June 27, 2018, **Defendant B sold 4.36 grams of heroin** to **Defendant A**, earning 1,400 yuan. Following the transaction: **Defendant A** was arrested on site, found in **possession** of **drugs** obtained from **Defendant B**, along with **0.6** grams of heroin personally carried. **Defendant B was arrested while injecting drugs after leaving the scene, and** was found in **possession** of **1.19** grams of **heroin** and **0.08** grams of **methamphetamine**. On July 5, 2018, forensic examination confirmed the presence of heroin in substances seized from Defendant A, and both heroin and methamphetamine in substances seized from Defendant B. Both defendants provided full confessions and pleaded guilty during the trial. **Defendant A** was previously sentenced to two years in prison. **Defendant B** was previously sentenced to four months imprisonment with eight months probation.……

LJP Models

Qwen-SFT	**Defendant A :**	Law Articles: *Article 347* ✓	Charges: *Crime of Trafficking* ✓	Penalty: *6 years* ✗
	Defendant B :	Law Articles: *Article 347* ✓	Charges: *Crime of Trafficking* ✓	Penalty: *2 years* ✗
MAGLJP (Our model)	**Defendant A :**	Law Articles: *Article 347* ✓	Charges: *Crime of Trafficking* ✓	Penalty: *3.5 years* ✓
	Defendant B :	Law Articles: *Article 347* ✓	Charges: *Crime of Trafficking* ✓	Penalty: *1 years* ✓

LELG-enhanced Penalty Prediction:

nodes = {
"*f1*" : "**Defendant A** was previously sentenced to two years in prison"
"*f2*" : "**Defendant A** engaged in **drug sales** on **five** occasions from his residence, earning a total of 370 yuan"
"*f3*" : "**Defendant B sold 4.36 grams of heroin** to **Defendant A**, earning 1,400 yuan"
"*f4*" : "**Defendant A** was found in **possession** of **drugs** obtained from **Defendant B,** along with **0.6** grams of heroin personally carried. "
"*f5*" : "**Defendant B** was found in **possession** of **1.19** grams of **heroin** and **0.08** grams of **methamphetamine**"
"*f6*" : "**Defendant B** was previously sentenced to four months imprisonment with eight months probation"
"*f7*" : "Both defendants provided full confessions and pleaded guilty during the trial"
"*k1*" : "**Recidivism leads to an increase of 10% to 40% in the base sentence**."
"*k2*" : "**Multiple drug trafficking offenses: 3-7 years imprisonment plus fine.**"
"*k3*" : "**For trafficking 4.36g heroin and 0.08g methamphetamine: up to 3 years imprisonment, detention or surveillance, plus fine.**"
"*k4*" : "**Drug-related recidivism requires a heavier sentence according to law**"
"*k5*" : "**Full confession may result in a reduced sentence**"
}
edges = [*[('f1', 'k1'), ('f2', 'k2'), ('f3', 'k3'), ('f4', 'k2'), ('f5', 'k3'), ('f6', 'k4'), ('f7', 'k5'), ('k1', 'p1'), ('k2', 'p1'), ('k4', 'p1'), ('k5', 'p1'), ('k3', 'p2'), ('k5', 'p2')*]]

Fig. 10. Case study. The figure shows a drug trafficking case analysis comparing baseline Qwen-SFT with our MAGLJP model. The top section presents the case facts, followed by prediction results from both models. The bottom section illustrates the Legal Event Logic Graph (LELG) generated by MAGLJP, which maps atomic case facts (f1-f7, shown in yellow) to relevant legal knowledge nodes (k1-k5, shown in gray) and final penalty predictions (p1-p2). This visualization demonstrates how MAGLJP effectively structures legal reasoning and improves prediction accuracy through explicit fact-knowledge-penalty connections.

5.4.4. Case study

In this section, we demonstrate how our proposed MAGLJP effectively utilizes legal knowledge for precise penalty prediction through LELG by examining a complex case study. This case study illustrates how MAGLJP's structured approach to incorporating legal knowledge through LELG not only improves prediction accuracy but also provides valuable interpretability for practical legal applications.

Fig. 10 presents a case involving two defendants (A and B) charged with drug trafficking. First, we show the performance of our strongest baseline, Qwen-SFT. While this model accurately predicts the applicable Law Article (Article 347) and correctly identifies both defendants' charges as Crime of Trafficking, it fails to accurately predict their penalties. This limitation highlights the challenges in penalty prediction even when article and charge predictions are successful.

In contrast, our MAGLJP model, assisted by the LegalKnow Assistant, generates a comprehensive LELG that decomposes the case into atomic case facts for both defendants A and B. The graph structure explicitly maps these facts to relevant legal knowledge (highlighted in Fig. 10) that enhances penalty prediction. The edges in LELG establish clear connections between facts and legal knowledge, as well as between knowledge nodes and the defendants' penalty nodes (p1 and p2, initially empty before prediction).

The LELG visualization in Fig. 10 demonstrates two key advantages of our MAGLJP model: (1) **Enhanced Multi-defendant Processing**: Through multi-agent collaboration, the MAGLJP not only predicts judgment outcomes more accurately for each defendant but also captures the complex interactions between defendants' behaviors and their legal implications. (2) **Interpretable Legal Reasoning**: The LELG serves as an interpretable visualization module that explicitly maps defendants' behaviors (both individual actions and interactions with co-defendants) to relevant legal knowledge. This interpretability feature effectively assists human experts in understanding the reasoning behind predicted judgments by providing a clear visualization of how specific actions and legal principles contribute to the final decision.

5.4.5. *Error analysis*

To identify potential limitation of our framework and provide valuable insights for future improvements, we conducted a detailed error analysis. We randomly sampled 100 cases from each dataset and manually analyzed the error patterns, categorizing them into three general types: *1. Misunderstanding of Case Information* (39%): These include overlooking crucial case details, confusion in crime constituent elements, natural language understanding biases, entity recognition errors, missing critical information for penalty determination, and complexity issues in cases involving multiple defendants or lengthy descriptions. *2. Deficiencies in Legal Knowledge Application* (41%): This category encompasses insufficient legal knowledge, improper handling of concurrent punishments, and inaccurate sentence calculation methods. *3. Reasoning Logic Issues* (20%): These involve factual errors in causal reasoning graphs, misinterpretation of evidence, and limitations in model generalization or incomplete reasoning logic.

Based on these findings, we propose several directions for improving multi-defendant LJP systems: **1. Enhanced Understanding of Complex Cases**: Future work should focus on incorporating more specialized legal knowledge for distinguishing confusing charges. This could be achieved through either utilizing LLMs with longer context windows or developing case decomposition strategies to improve comprehension of case information. **2. Refined Legal Knowledge Integration**: While our MAGLJP automates legal knowledge integration, involving legal experts, lawyers, and judges specialized in sentencing could provide valuable guidance for LELG generation and refinement, leading to more comprehensive and accurate legal knowledge representation. **3. Advanced Reasoning Capabilities**: Incorporating more sophisticated LLMs with stronger reasoning abilities could address the logical reasoning limitations identified in our analysis.

5.4.6. *Computational cost analysis*

In this section, we present a comprehensive analysis of the computational costs associated with our MAGLJP framework across its three primary modules: Conviction Agent, LegalKnow Assistant, and Sentencing Agent. Our evaluation spans two validation datasets: MultiLJP and CAIL2024-DRDZ.

On the MultiLJP dataset, the Conviction Agent was trained on a cluster of four NVIDIA A100 GPUs for ten epochs, requiring approximately 46,146 s total, or roughly 1.28 h per epoch. During inference, the model demonstrates efficient performance, requiring only 0.09 s to process a single sample. The LegalKnow Assistant, implemented using the Qwen-max API, requires an average of 91 s to generate LELG for each case. However, thanks to the API's parallel processing capabilities with 1200 Queries Per Minute (QPM), the entire dataset can be processed in approximately 20 min. The average input and output tokens for LELG generation are 14,485.81 and 1,210.20, respectively, with a total API cost of 989 CNY (China Yuan) (approximately $136.41 at an exchange rate of 7.25). On average, processing one case takes 0.06 s with a cost of 0.046 CNY. To better demonstrate the practical cost implications, we analyze three representative cases from the dataset. For a simple case (minimum complexity) with 10,196 input and 911 output tokens, the cost is merely 0.0267 CNY. A typical case (median complexity) with 13,784 input and 1,028 output tokens costs 0.0355 CNY. Even for the most complex case (maximum complexity) requiring 91,805 input and 1,428 output tokens, the cost remains reasonable at 0.2238 CNY. The Sentencing Agent training stage requires 252,791 s, or about 3.51 h per epoch. Its inference time remains efficient at 0.14 s per sample.

For the CAIL2024-DRDZ dataset, we observed similar patterns with some variations. The Conviction Agent required 61,310 s for training, averaging 1.7 h per epoch, with a notably fast inference time of 0.03 s per sample. The LegalKnow Assistant maintained efficient processing times, averaging 82 s per case for LELG generation, with the entire dataset processed in approximately 13 min. The average input and output tokens were slightly lower at 12,364.74 and 1,141.40, respectively, resulting in a total API cost of 548 CNY (China Yuan) (approximately $75.61). Similarly, we examine three representative cases from this dataset: a simple case (9,730 input, 921 output tokens) costing 0.0256 CNY, a typical case (11,599 input, 818 output tokens) at 0.0298 CNY, and the most complex case (89,679 input, 1,798 output tokens) at 0.2195 CNY, all demonstrating the method's cost-effectiveness.

These results demonstrate that our framework will not require such substantial computational resources and achieve practical inference times suitable for real-world applications. The LegalKnow Assistant's API costs remain reasonable considering the enhanced performance and quality of legal knowledge incorporation. Moreover, the parallelization capabilities of the API enable efficient processing of large-scale datasets, making our framework both practical and scalable for real-world legal applications.

5.5. Further discussion

1. Further Analysis of Sub-task Performance.

In our main experiments, MAGLJP consistently outperforms all baselines across three sub-tasks. Compared to the strongest baseline, Qwen-SFT, MAGLJP achieves significant improvements: on the MultiLJP dataset, it increases F1 scores by 7% in both Article and Charge Prediction, and notably, by 30% in Penalty Prediction. Similarly, on CAIL2024-DRDZ, while showing modest gains in Article Prediction (0.3%) and Charge Prediction (1%), it demonstrates a substantial 28% improvement in Penalty Prediction.

These performance patterns merit further analysis. First, Article and Charge Prediction have traditionally been considered relatively straightforward tasks in Legal Judgment Prediction (LJP), while models consistently struggle with Penalty Prediction. This performance disparity is evident across various studies — for instance, on the CAIL2018 single-defendant dataset, models achieved 87.82% and 94.99% accuracy for Article and Charge Prediction respectively, while only reaching 48.72% for Penalty Prediction (Wu et al., 2023). Similarly, on the more complex MultiLJP dataset, previous state-of-the-art models achieved 91.46% and 89.54% accuracy for the first two tasks, but only 42.72% for Penalty Prediction (Lyu et al., 2023). Despite these already high baselines in Article and Charge Prediction, our Multi-Agent approach still manages to achieve modest but meaningful improvements, demonstrating its effectiveness even in well-solved tasks.

While MAGLJP brings substantial improvements to Penalty Prediction, the overall performance in this task remains challenging. This reflects the inherent complexity of penalty prediction, which stems from the multi-faceted and context-dependent nature of sentencing decisions. The task requires consideration of not only case facts and legal knowledge, but also various crucial factors: the nature and social harm of the crime, defendants' degrees of involvement (principal or accomplice), offender characteristics (e.g., age-related considerations), and other circumstances subject to judicial discretion. Furthermore, the interpretation and application of legal knowledge relevant to sentencing can vary across different temporal and geographical contexts, adding another layer of complexity to the prediction task. Our proposed MAGLJP makes significant strides in addressing this complexity through structured knowledge representation and utilization using LELG. These improvements, though substantial compared to existing methods, also highlight the considerable room for advancement in this challenging domain, pointing to promising directions for future research.

2. Discussion about LELG.

The integration of LELG represents our primary contribution to LJP tasks. Our adoption of LELG was primarily motivated by the challenges in penalty prediction, particularly the difficulty in effectively utilizing legal knowledge in sentencing scenarios. LELG's main advantage lies in its explicit representation of causal relationships between criminal facts, legal knowledge, and sentencing outcomes. This structured approach has significantly enhanced the performance of the sub-task. Moreover, the Event Logic Graph structure is particularly suitable for judicial cases, as it efficiently represents atomic case facts, relevant legal knowledge, and penalties as nodes, leveraging LLMs' text comprehension and generation capabilities to establish causal logical relationships in the sentencing decision-making process.

Compared to traditional knowledge representation methods such as knowledge graphs and ontologies, which often face limitations in predefined relationships, logical reasoning support, and flexibility (Ding et al., 2019; Zhao et al., 2017), LELG offers a more adaptable alternative. However, our analysis reveals several areas for potential improvement in LELG: 1. **Model Capability Dependencies**: As demonstrated through ablation experiments comparing Qwen-max and Qwen-32B models, LELG generation quality depends significantly on the underlying model's capabilities. Our manual evaluation of 100 randomly sampled LELGs from each model shows success rates of 88% for Qwen-max and 83% for Qwen-32B, confirming that stronger models generally produce higher-quality LELGs and subsequently improve penalty prediction performance. 2. **Generation Accuracy**: As identified in our error analysis, LELG generation may encounter issues such as factual errors, misinterpretation of evidence, improper handling of concurrent punishments, and inaccurate sentencing calculations. While current LLMs demonstrate promising capabilities in generating high-quality LELGs, there remains room for improvement.

To address these limitations, we propose two primary directions for future research: 1. Enhancement of LELG Generation: This could be achieved through utilizing more capable models and incorporating expert legal annotations for more precise LELG construction. 2. Integration with Other Knowledge Structures: Future research could explore combining LELG with ontologies and knowledge graphs (Ngo et al., 2024; Nguyen et al., 2022) to reduce LLM hallucinations and enhance reliability.

3. Discussion about Dataset.

Our study focuses on multi-defendant LJP, which represents a more prevalent and challenging form of legal judgment prediction. We evaluate our approach on two comprehensive datasets: MultiLJP, which encompasses 22 articles, 23 charges, and various penalty types, and CAIL2024-DRDZ, which covers 31 articles and 30 charges. While these datasets demonstrate considerable diversity in terms of charge types and case scenarios, those specifically addressing multi-defendant scenarios remain relatively scarce in the field.

Our experimental validation currently focuses on Chinese legal judgments. Future research could benefit from exploring the development of multi-defendant LJP datasets across different languages, jurisdictions, and legal systems (Marwala & Mpedi, 2024; Nguyen et al., 2024; Novelli et al., 2024). The cross-jurisdictional validation would be particularly valuable in establishing the framework's generalizability and robustness across different legal traditions and practices.

4. Practical Applications

MAGLJP demonstrates significant potential across various practical applications. Primarily, MAGLJP can serve as a decision support tool for judges during the judicial process. In the pre-trial phase, legal professionals can utilize MAGLJP to conduct preliminary analyses of complex multi-defendant cases, efficiently identify key evidence and relationships between defendants, and provide initial sentencing recommendations based on relevant legal knowledge, thereby enhancing judicial efficiency.

Furthermore, LELG provides a visualization of critical case elements, offering a window into model interpretability. Legal professionals can interact with the model by examining LELG's reasoning process, enabling a human-AI collaboration where experts can identify and correct errors in LELG, improve system accuracy, and reduce potential biases, ultimately supporting more reasonable and fair judicial decisions. Additionally, LELG's logical structure presents a novel approach to similar case retrieval. By decomposing specific cases into atomic facts and establishing their correspondence with legal provisions, it creates an abstract mapping from concrete cases to legal knowledge, enabling similar case retrieval based on essential case characteristics.

Moreover, MAGLJP serves as an effective tool for legal education and training. It assists less experienced legal practitioners in better understanding the processing logic of multi-defendant cases, particularly in comprehending the connections between case facts, legal application, and final judgment outcomes, thus providing robust support for case studies and practical training.

5. Ethical Considerations

Given the sensitive nature of the judicial domain, we emphasize that our framework demonstrates potential in assisting, rather than replacing, the legal judgment process. Several ethical considerations warrant careful attention: **First**, we acknowledge potential sources of bias in model training, particularly from training data. These biases could perpetuate or amplify existing prejudices in the legal system. Ensuring algorithmic fairness and preventing discrimination requires ongoing scrutiny and mitigation strategies. **Second**, regarding accountability and professional impact, we stress that this work is purely academic exploration. It is not intended for direct practical application or automated judicial decision-making. The irreplaceable role of judges and legal professions, with their professional expertise and ethical judgment, must be preserved. Any potential future applications should operate strictly under judicial supervision. The implementation of AI systems in legal contexts demands careful consideration of these ethical implications. While our research explores technological possibilities, we firmly maintain that judicial decisions must remain under human oversight, with judges' expertise, discretion, and ethical considerations at the forefront of the legal decision-making process.

6. Conclusion

In this paper, we introduce MAGLJP, a novel multi-agent framework with Legal Event Logic Graph for multi-defendant legal judgment prediction, aiming to address the challenges of increased procedural complexity and insufficient integration of legal knowledge in multi-defendant LJP. Mirroring the reasoning process of legal experts, our MAGLJP decomposes the complex multi-defendant LJP task into a Standard Operating Procedure. This framework comprises three specialized agents – the Conviction Agent, Legal Knowledge Assistant Agent, and Sentencing Agent – working collaboratively to handle different aspects of the judgment prediction process. To tackle the challenge of legal knowledge especially for penalty prediction, we first constructed a comprehensive legal knowledge base. This resource provides valuable support for precise penalty prediction and serves as a foundation for future research in legal AI. We also introduced the Legal Event Logic Graph (LELG), to effectively represent and reason about the relationships between criminal facts, legal knowledge, and sentencing outcomes. This novel approach significantly improves both the interpretability and accuracy of legal judgment prediction.

Through extensive experiments on two multi-defendant LJP datasets, we demonstrated the superior performance of MAGLJP. These results validate the effectiveness of our approach and suggest promising directions for future research in legal judgment prediction. In the legal domain, our framework presents an effective approach to assist in automating the legal judgment process. However, it is crucial to emphasize that our work is primarily academic research, and deep learning models must operate under human guidance and supervision. While they can serve as decision support tools in automated workflows, they cannot and should not replace judges' role in actual case adjudication. Future work could focus on exploring more diverse forms of legal knowledge integration to enhance LJP task performance more efficiently. This includes investigating various knowledge sources, representation methods, and reasoning mechanisms to further improve the framework's capabilities while maintaining its role as a supportive tool in legal decision-making.

CRediT authorship contribution statement

Weikang Yuan: Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Conceptualization. **Kaisong Song:** Writing – review & editing, Visualization, Methodology, Conceptualization. **Zhuoren Jiang:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Junjie Cao:** Writing – review & editing, Visualization, Methodology, Conceptualization. **Yujie Zhang:** Writing – review & editing, Visualization, Methodology, Data curation. **Chengyuan Liu:** Software, Methodology. **Jun Lin:** Writing – review & editing, Supervision, Resources, Project administration. **Ji Zhang:** Writing – review & editing, Resources, Project administration. **Kun Kuang:** Writing – review & editing, Supervision, Resources. **Xiaozhong Liu:** Writing – review & editing, Supervision.

Funding

We would like to acknowledge the funding by the National Natural Science Foundation of China (72104212) and the Fundamental Research Funds for the Central Universities, China. This work is also supported by Alibaba Innovative Research Program.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 2898–2904).

Chen, W., Ma, X., Wang, X., & Cohen, W. W. (2023). Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1–113.

Cui, J., Li, Z., Yan, Y., Chen, B., & Yuan, L. (2023). Chatlaw: Open-source legal large language model with integrated external knowledge bases. CoRR.

Cui, J., Shen, X., & Wen, S. (2023). A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, *11*, 102050–102071.

Deroy, A., Ghosh, K., & Ghosh, S. (2023). How ready are pre-trained abstractive models and LLMs for legal case judgement summarization?. arXiv preprint arXiv:2306.01248.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423/.

Ding, X., Li, Z., Liu, T., & Liao, K. (2019). ELG: An event logic graph. ArXiv abs/1907.08015. URL: https://api.semanticscholar.org/CorpusID:197544867.

Feng, Y., Li, C., & Ng, V. (2022). Legal judgment prediction: A survey of the state of the art. In *IJCAI* (pp. 5461–5469).

Gan, L., Kuang, K., Yang, Y., & Wu, F. (2021). Judgment prediction via injecting legal knowledge into neural networks. In *Proceedings of the AAAI conference on artificial intelligence*: *Vol. 35*, (14), (pp. 12866–12874).

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680.

He, Z., Cao, P., Wang, C., Jin, Z., Chen, Y., Xu, J., Li, H., Jiang, X., Liu, K., & Zhao, J. (2024). Simucourt: Building judicial decision-making agents with real-world judgement documents. arXiv preprint arXiv:2403.02959.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., et al. (2023). MetaGPT: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*.

Huang, W., Feng, Y., Li, C., Wu, H., Ge, J., & Ng, V. (2024). CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset. In *Findings of the association for computational linguistics* (pp. 5895–5906).

Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the supreme court of the United States. *PloS One*, *12*(4), Article e0174698.

Kort, F. (1957). Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, *51*(1), 1–12.

Li, S., Zhao, S., Zhang, Z., Fang, Z., Chen, W., & Wang, T. (2025). Basis is also explanation: Interpretable legal judgment reasoning prompted by multi-source knowledge. *Information Processing & Management*, *62*(3), Article 103996.

Liu, Y. H., & Chen, Y. L. (2018). A two-phase sentiment analysis approach for judgement prediction. *Journal of Information Science*, *44*(5), 594–607.

Louis, A., van Dijck, G., & Spanakis, G. (2024). Interpretable long-form legal question answering with retrieval-augmented large language models. *38*, In *Proceedings of the AAAI conference on artificial intelligence* (20), (pp. 22266–22275).

Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2727–2736).

Lyu, Y., Hao, J., Wang, Z., Zhao, K., Gao, S., Ren, P., Chen, Z., Wang, F., & Ren, Z. (2023). Multi-defendant legal judgment prediction via hierarchical reasoning. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 2198–2209).

Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., & Modi, A. (2021). ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. arXiv preprint arXiv:2105.13562.

Marwala, T., & Mpedi, L. G. (2024). Artificial intelligence and the law. In *Artificial intelligence and the law* (pp. 1–25). Springer.

Ngo, H. Q., Nguyen, H. D., & Le-Khac, N. A. (2024). Ontology knowledge map approach towards building linked data for Vietnamese legal applications. *Vietnam Journal of Computer Science*, *11*(02), 323–342.

Nguyen, T. H., Nguyen, H. D., Pham, V. T., Tran, D. A., & Selamat, A. (2022). Legal-onto: An ontology-based model for representing the knowledge of a legal document. In *ENASE* (pp. 426–434).

Nguyen, H., Pham, V., Ngo, H. Q., Huynh, A., Nguyen, B., & Machado, J. (2024). Intelligent search system for resume and labor law. *PeerJ Computer Science*, *10*, Article e1786.

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*, *55*, Article 106066.

OpenAI (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Pan, S., Lu, T., Gu, N., Zhang, H., & Xu, C. (2019). Charge prediction for multi-defendant cases with multi-scale attention. In *Computer supported cooperative work and social computing: 14th cCF conference, ChinesecSCW 2019, kunming, China, August 16–18, 2019, revised selected papers 14* (pp. 766–777). Springer.

Segal, J. A. (1984). Predicting supreme court cases probabilistically: The search and seizure cases, 1962–1981. *American Political Science Review*, *78*(4), 891–900.

Sulea, O. M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., & Van Genabith, J. (2017). Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306.

Sun, J., Dai, C., Luo, Z., Chang, Y., & Li, Y. (2024). Lawluo: A chinese law firm co-run by llm agents. arXiv preprint arXiv:2407.16252.

Sun, Z., Zhang, K., Yu, W., Wang, H., & Xu, J. (2024). Logic rules as explanations for legal case retrieval. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation* (pp. 10747–10759). Torino, Italia: ELRA and ICCL, URL: https://aclanthology.org/2024.lrec-main.939.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., & Gerstein, M. (2024). MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the association for computational linguistics* (pp. 599–621).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Wang, J., Wu, J., Hou, Y., Liu, Y., Gao, M., & McAuley, J. (2024). InstructGraph: Boosting large language models via graph-centric instruction tuning and preference alignment. In L. W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics* (pp. 13492–13510). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-acl.801, URL: https://aclanthology.org/2024.findings-acl.801/.

Wei, X., Xu, Q., Yu, H., Liu, Q., & Cambria, E. (2024). Through the MUD: A multi-defendant charge prediction benchmark with linked crime elements. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2864–2878).

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First conference on language modeling*. URL: https://openreview.net/forum?id=BAakY1hNKS.

Wu, Y., Zhou, S., Liu, Y., Lu, W., Liu, X., Zhang, Y., Sun, C., Wu, F., & Kuang, K. (2023). 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, http://dx.doi.org/10.18653/v1/2023.emnlp-main.740, URL: https://aclanthology.org/2023.emnlp-main.740.

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, *2*, 79–84.

Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., & Zhao, J. (2020). Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3086–3095).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 483–498). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.41, Online. URL: https://aclanthology.org/2021.naacl-main.41/.

Yang, W., Jia, W., Zhou, X., & Luo, Y. (2019). Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 4085–4091).

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., ... Qiu, Z. (2024). Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Yuan, W., Cao, J., Jiang, Z., Kang, Y., Lin, J., Song, K., Lin, T., Yan, P., Sun, C., & Liu, X. (2024). Can large language models grasp legal theories? Enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 7577–7597).

Yue, L., Liu, Q., Jin, B., Wu, H., Zhang, K., An, Y., Cheng, M., Yin, B., & Wu, D. (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* (pp. 973–982).

Zhao, Q., Gao, T., & Guo, N. (2023). LA-MGFM: A legal judgment prediction method via sememe-enhanced graph neural networks and multi-graph fusion mechanism. *Information Processing & Management*, *60*(5), Article 103455.

Zhao, S., Wang, Q., Massung, S., Qin, B., Liu, T., Wang, B., & Zhai, C. (2017). Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the tenth ACM international conference on web search and data mining* (pp. 335–344). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3018661.3018707, URL: https://doi.org/10.1145/3018661.3018707.

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2018). Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3540–3549).

Zhong, H., Xiao, C., Tu, C., et al. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5218–5230). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.466, Online. URL: https://aclanthology.org/2020.acl-main.466/.