Parallel Learning for Legal Intelligence: A HANOI Approach Based on Unified Prompting

Zhuoyang Song, Min Huang, *Member, IEEE*, Qinghai Miao, *Senior Member, IEEE*, and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*— Pretrained language models (PLMs) have made significant progress on various NLP tasks recently. However, PLMs encounter challenges when it comes to domain-specific tasks such as legal AI. These tasks often involve intricate expertise, expensive data annotation, and limited training data availability. To tackle this problem, we propose a human-oriented artificial–natural parallel system for organized intelligence (HANOI)-Legal based on the parallel learning (PL) framework. First, by regarding the description in PL as the pretraining process based on a large-scale corpus, we setup an artificial system based on a PLM. Second, to adapt the PLM to legal tasks with limited resources, we propose UniPrompt as a prescription. UniPrompt serves as a unified prompt-based training framework, enabling the utilization of diverse open datasets for these tasks. Third, we labeled a few task-specific legal data through distributed autonomous operations (DAO-II) for further fine-tuning. By combining a scalable unified-task-format reformulation and a unified-prompt-based training pipeline, HANOI-Legal leverages PLMs' linguistic capabilities acquired from a variety of open datasets to generate task-specific models. Our experiments in two legal domain tasks show that HANOI-Legal achieved an excellent performance in low-resource scenarios compared to the state-of-the-art prompt-based approach.**

***Index Terms*— Natural language processing (NLP), parallel learning (PL), parallel systems, pretrained language model (PLM), prompt tuning.**

I. Introduction

CURRENTLY, AI technology represented by ChatGPT [1] is propelling human society into a new era of linguistic intelligence. With unprecedented performance, these models have widely influenced the whole world and also profoundly affected everyone's work and life [2]. Through analyzing successful cases like AlphaGo and ChatGPT, we can observe three distinct characteristics of AI technology's development and principles.

First, the utilization of pretrained foundation models (PFMs) [3]. Training on massive data, PFMs have powerful generalization ability which enables zero-shot or few-shot learning with leading performance. GPT-3, the pretrained language model (PLM) used in ChatGPT, exemplifies such a PFM and demonstrates its immense power. An important aspect of PFMs is the scale effect, where larger models yield significantly improved performance across diverse tasks, showcasing the emergence of enhanced capabilities. As the scale of the model increases, the PFM's performance exhibits a remarkable upward trend. This makes large models, including large language models (LLMs) [4] and large vision models (LVMs) [5], the foundation of artificial intelligence. As a result, exploring methods to effectively harness the capabilities of PFMs for specific scenarios has emerged as a mainstream focus in AI research.

Second, the utilization of a systematic approach. It is evident that a single model alone lacks the capability to address intricate challenges effectively. However, through systematic approaches like AlphaGo, AlphaFold, ChatGPT, and others, significant breakthroughs have been achieved in solving real-world problems, leading to notable milestones in the field of artificial intelligence. The systematic approach integrates multiple models/algorithms to leverage their individual strengths under unified coordination. This paradigm shift, exemplified by the successes of AlphaGo, AlphaFold, and ChatGPT, has reshaped AI research and opened new avenues for advancements in the field [2].

Third, rethinking the role of humans. Deep neural networks can extract features automatically, accurately, and efficiently, making end-to-end learning models popular. However, in some applications, humans play an important role, such as making requirements, setting goals, providing knowledge, implementing control, performing evaluation, and so on. Therefore, end-to-end learning is not the ultimate solution, and humans cannot simply be removed from the machine-learning process. The value of human-in-loop is evident in the success of ChatGPT [1], where reinforcement learning from human feedback (RLHF) [6] played a key role.

While the attention toward PFMs is prominent, the latter two aspects are sometimes overlooked in traditional AI approaches. However, from the perspective of parallel learning (PL) [7], these three new changes have already been included in the concept of human-oriented artificial–natural parallel system for organized intelligence (HANOI) [8], [9], which is a PL

Manuscript received 25 April 2023; revised 13 July 2023; accepted 27 July 2023. Date of publication 14 August 2023; date of current version 3 April 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB2104001, in part by the National Natural Science Foundation of China under Grant 62271485, and in part by the Open Project of the State Key Laboratory for Management and Control of Complex Systems under Grant 20220117. *(Corresponding author: Qinghai Miao.)*

Zhuoyang Song, Min Huang, and Qinghai Miao are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: songzhuoyang20@mails.ucas.ac.cn; huangm@ucas.ac.cn; miaoqh@ucas.ac.cn).

Fei-Yue Wang is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feiyue.wang@ia.ac.cn).

Digital Object Identifier 10.1109/TCSS.2023.3301400

framework for organized intelligence, through the interaction between artificial systems and the natural world in a human-in-the-loop fashion.

HANOI emphasizes the role of humans in the machine-learning process [10]. It recognizes that AI should serve humans and aims to reposition its value in creating a better society. End-to-end learning is not always ideal, and a human-in-loop approach, leveraging experience, knowledge, and creativity, is critical for successful AI projects. HANOI's second aspect focuses on the parallel systems of the artificial and natural worlds. The artificial world provides a platform for machine learning to incorporate diverse knowledge and explore unknown spaces. It offers a sandbox for decision-making and guiding the natural world toward optimal development. HANOI's third element proposes decentralized autonomous organizations (DAO-I) and distributed autonomous operations (DAO-II) to organize AI development [11], [12]. DAOs ensure fairness, transparency, and accountability, democratize AI by involving the public, and use blockchain for immutable records of scientific discoveries. By embracing the value of humans, creating PL systems, and implementing decentralized organizational structures, HANOI offers a framework to harness the potential of AI while addressing ethical and practical challenges.

PL involves three principles: description, prescription, and prediction. First of all, the pretraining process of the large model corresponds to the description of PL. In PL, the artificial system is a broad concept, including 3-D virtual reality models that directly correspond to the real world, as well as high-dimensional neural network models that indirectly summarize the real world. As a result of training on massive data, the PFM is a form of artificial system. Second, the large model learns on massive data through self-supervised methods, forming the knowledge hidden in the parameters of the neural network. Since this knowledge cannot be directly understood and controlled by humans, people need to further guide the PFM to release its proper capabilities in a specified way for certain applications. This boot process includes the methods of introducing adapters or controllers to a PFM, as well as methods such as in-context learning, instruction learning, and prompt learning that do not modify PFMs. The boot process corresponds to the prescription in PL. Third, the aligned model through prescription can be further fine-tuned with specific task data or directly put online to give responses to users' requests. This step corresponds to prediction in PL. The initial model test or online trial operation usually cannot achieve optimal performance, so it needs to be further improved through user feedback.

These perspectives motivate us to solve legal tasks based on the HANOI framework. Legal artificial intelligence (LegalAI) aims to apply AI technologies to the legal domain, where the majority of data is in textual form. Nowadays, LegalAI is mainly based on natural language processing (NLP) methods to assist legal practitioners with various textual tasks and shows great potential in driving the efficient functioning of the judicial system [13].

Compared with general domains, the higher complexity of legal tasks challenges the model's comprehensive linguistic capability as well as its understanding of legal concepts and knowledge. Recently, many works attempted to apply PLMs in the legal domain, adapting their powerful linguistic capabilities to various legal downstream tasks [14], [15].

However, due to the specialized nature of the legal domain, which leads to greater data annotation difficulty and cost than the general domain, it is not easy to obtain sufficient high-quality annotated data to directly fine-tune the PLMs. This makes PLM-based LegalAI encounter many challenges when applied to real legal tasks.

To tackle these problems, we propose **HANOI-Legal** (HANOI framework for Legal AI), as shown in Fig. 1. Briefly, we first build a PLM, which is a mapping from the real world to the artificial system, by condensing a huge amount of corpus into neural network parameters. Second, we group diverse open datasets by their required linguistic capabilities and integrate them into a unified dataset through the unified-task-format reformulation. Then, the vanilla PLM is trained to handle the unified task with unified prompting (UniPrompt), while acquiring rich prior knowledge and diverse linguistic capabilities from the unified dataset. By simply continuing training on a few labeled data, the unified PLM is applied to domain-specific tasks with much less data annotation overhead.

We applied the proposed approach to three complex tasks in the legal domain, that is, Chinese judicial reading comprehension, civil case event classification (CLS), and civil case information extraction (IE). The experiments show that our UniPrompt significantly improves PLMs' performance in low-resource settings, suggesting that HANOI-Legal is a feasible way for domain-specific tasks to alleviate the adverse effects of data scarcity and reduce labeling costs. The contributions of this article are as follows.

1) The HANOI-Legal PL framework for legal intelligence based on PLMs.
2) A unified prompt-based training method (UniPrompt), which enables a task-agnostic unified model leverage various open datasets to enhance its performance on the target tasks.
3) Experiments on complex LegalAI tasks show that UniPrompt significantly improves PLMs' performance in low-resource settings.

This article is organized as follows. Related works are presented in Section II. The methodology is given in Section III. In Section IV, experiments on three tasks are introduced in detail. The article is summarized with the remaining problems and future directions in Section V.

II. Related Works

A. Parallel Intelligence and PL

Parallel system [16], [17], [18], [19] is a methodology to execute evaluation, prediction, optimization, and control for an objective through interactions between the natural (real) subsystem and one or multiple corresponding virtual (artificial) subsystems. The real subsystem is guided by the artificial subsystem to obtain an effective solution. The parallel system framework technically integrates artificial systems (A), computational experiments (C), and parallel execution (P). The ACP methods in parallel systems are supported by technologies such as big data, cloud computing, the Internet of Things, and deep learning. Based on the infrastructure of cyber–physical–social systems (CPSSs) [20], a parallel system realizes knowledge representation, decision inference, as well as adaptive optimization with closed-loop feedback. The parallel system has been widely used in control, transportation, medical, automatic driving, urban management, military, and chemical. New advances in the fields such as scenarios engineering [21] are also emerging.

Digital twin and Metaverse are two terms related to parallel systems. The digital twin is the basic form of combining virtual and real systems. The Metaverse can be seen as the promotion of the digital twin system from industrial applications to social cyberspace. When knowledge automation and intelligence are realized in the Metaverse, a technically parallel system will be formed.

By introducing the idea of the parallel system into machine learning, Li et al. [7] proposed the concept of PL that searches for an optimized policy with virtual–real interactions in a setting of reinforcement learning. The PL framework comprises three components, namely, description, prediction, and prescription. Furthermore, in 2023, the PL framework was extended to general machine learning [22].

Given a machine-learning task, a description in PL means to construct a corresponding artificial system that can be used as a factory to generate virtual data with labels, or as a testbed to verify the learning performance. Traditionally, there are two ways to utilize the artificial system, that is, Syn2Real and Sim2Real. Syn2Real means generating virtual data by directly modifying the original data using basic Data Augmentation operations or using generative methods like GANs. Sim2Real means running simulations in 3-D virtual environments for reinforcement learning using domain randomization operations. Representative work by Li et al. [23] is to enhance foreground detection in highway surveillance scenarios by generating synthetic videos. Nowadays, PFMs are dominating both computer vision and NLP areas. The PFMs are artificial systems with high-dimensional representation and their pretraining process corresponds to the description in PL. Li et al. [24] presented a novel method related to foundation models and Metaverse.

B. Pretrained Language Models and Prompt Learning

In recent years, "pretraining and fine-tuning" has been the dominant methodology in NLP. With large-scale unsupervised corpus, PLMs can acquire powerful linguistic capabilities by training with the pretraining tasks, including masked language modeling (Masked-LM [25]), autoregressive language modeling (Autoregressive-LM [26], [27]), and so on. Among them, Masked LM trains models' ability to predict masked words from context, and Autoregressive LM trains the ability to generate the following text from the above. These pretraining tasks enable PLMs to capture the characteristics of the natural world from a linguistic perspective. On various NLP downstream tasks, PLMs such as GPT [26], BERT [25], and T5 [28] have achieved promising performance by performing fully supervised fine-tuning on data from these tasks, demonstrating strong task adaptation capabilities [4].

As scaling up the model parameters, PLMs emerge with a significant boosting on the zero-shot generalization capability and acquire amazing abilities [29] such as multistep reasoning with the Chain-of-Thought [30], [31], in-context learning [32], [33], and so on. For better motivating the linguistic capabilities of large-scale language models (LLMs) on downstream tasks, the recently popular prompt learning methods use textual prompts to reformulate tasks into formats similar to the pretraining tasks [34], bridging the gap between the pretraining phase and fine-tuning phase. Specifically, the prompt texts are of two types, discrete and continuous, which are concatenated into the raw data in the form of prefixes [28], [35] or cloze [36], [37]. The discrete prompts are texts with actual meanings [26], [38], while continuous prompts are virtual representations in the embedding space without semantics [35], [36], [37], [39].

By tuning the parameters in the models or prompts on the reformulated data, prompt learning serves as a prescription for more efficient transformation of PLMs to real-world tasks. Many current efforts unified various downstream tasks into the same format to further push the boundaries between them and used unified prompts to guide PLMs through the unified-format task. For question-answering (QA) tasks, UnifiedQA advocated for a unified view of different QA formats by building a unified system [38]. For CLS tasks, BinaryClfs converted various formats into a binary CLS for each class with a few task descriptions [40]. Due to similarities in the task format (Masked-LM versus cloze format, Autoregressive-LM versus prefix format), PLMs can be adapted more efficiently to a variety of downstream tasks simultaneously.

On this basis, it gradually evolved into multitask Instruction Tuning, a large-scale prompting engineering. Specifically, Instruction Tuning methods use a wider variety of tasks, large-scale data, and a more diversified of prompts to elicit the ability of PLMs to understand textual instructions and generate text that solves the tasks [41], [42], [43]. Many works related to the recently popular ChatGPT went further by introducing the RLHF approach to align the performance of PLMs on real tasks with human preferences [1], [44], [45], making PLMs exhibit more compliant and harmless capabilities.

Our proposed HANOI-Legal revisits PLMs from the PL perspective. Focusing on the problem of data scarcity and the difficulty of legal textual tasks, the HANOI-Legal framework reconstructs various textual tasks into unified instructions based on well-designed prompts and applies PLMs to the legal domain through staged training processes.

III. Methodology

In this section, we give the HANOI-Legal design in detail based on the framework introduced in Fig. 1. The learning pipeline of HANOI-Legal, as shown in Fig. 2, consists of the following procedures.

Description refers to building an artificial system in the form of PLMs. Typically, a vanilla LLM is selected and then pretrained on a large-scale corpus by self-supervised learning

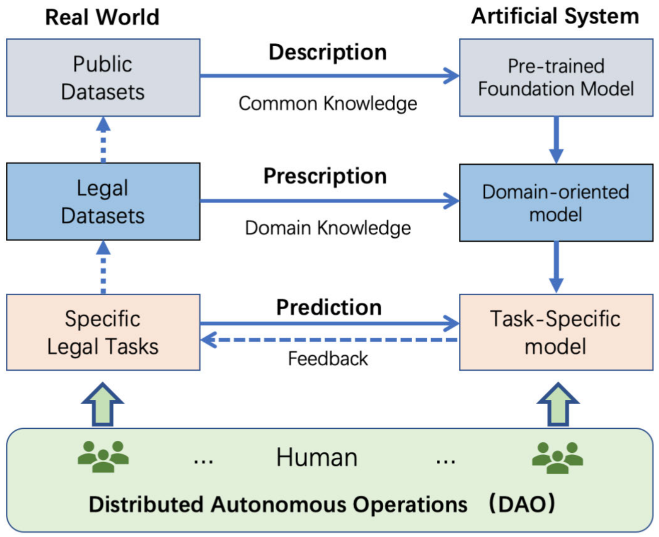


Fig. 1. HANOI-Legal is a four-level framework. On the top is the common knowledge abstraction level, where a mapping from the real world (massive data) to the artificial system (PFM) is processed by descriptive learning methods. The second is the domain knowledge abstraction level, where the PFM is further trained or tuned on the legal dataset. The third is the task-specific level, in which the users interact with finetuned PFM to make predictions on specific legal tasks such as IE, QA, and so on. At the bottom is the Organization layer that supports the operation of DAO-I or DAO-II with humans in a loop. The dashed lines indicate the inverse flow of information started from human feedback to correct the labels, enhance the dataset, and then improve the ability of PFM.

with vanilla pretraining tasks. The output is a PLM equipped with general linguistic capabilities.

Prescription is the focus of this article. There are several options to guide the PLM to release its capabilities, for example, fine-tuning, prompt-tuning, in-context-learning, instruct-learning, and chain-of-thought. In this article, we adopt a two-phase method. The first phase is prompt-based training with a highlight on unified prompting. Specifically, the model is trained on the unified dataset, covering various tasks of different categories. With similar training objectives, the PLMs softly apply the basic linguistic capabilities to handle tasks in the unified format, while purposefully acquiring specific capabilities potentially useful for target tasks. The second phase is prompt-based fine-tuning. For the target task, the unified PLMs are finally fine-tuned on the unified target dataset. With enriched prior knowledge and diverse linguistic capabilities, models quickly acquire the ability to solve a specific task in a unified format from limited labeled data.

As shown in Fig. 2, the proposed framework utilizes diverse public datasets with its scalable unified-task-format reformulation and produces task-specific models for each target task through a prompt-based training pipeline. Sections III-A and III-B will give more technical details on the description and prescription.

A. Description Based on PLM

The term "description" means the process to map the real world to an artificial system, which corresponds to the pretraining of language models for legal tasks. Equation (1) shows the abstraction of the task: given an input sentence s_{in}, we want an output sentence s_{out} from a PLM $\mathcal{M}_\theta$ by selecting and concatenating each generated word w_i

$$\begin{aligned} s_{\text{out}} &= \mathcal{M}_\theta(s_{\text{in}}) = \text{Concat}(w_1, w_2, \ldots, w_{n-1}, w_n) \\ w_i &= \underset{w}{\operatorname{argmax}}\, \mathcal{P}_\theta(w|s_{\text{in}}, w_1, \ldots, w_{i-1}). \end{aligned} \tag{1}$$

Currently, the basic building block of PLMs is the Transformer, which consists of an encoder stack and a decoder stack. There are three types of PLMs according to the use of encoders and decoders, that is, encoder-only, decoder-only, and encoder–decoder-based PLMs.

1) Encoder-Based PLMs: Equation (2) gives the training principle of encoder-only PLMs, whose task is to predict the word w_i which is masked out from the input sentence s [46]

$$\begin{aligned} &\mathcal{P}(w_i|w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n) \\ &\qquad s = \text{Concat}(w_1, w_2, \ldots, w_{n-1}, w_n). \end{aligned} \tag{2}$$

2) Decoder-Based PLMs: The training of decoder-only PLMs computes the probability of words w_i in sequence and concatenates the words as the generated output sentence [28]

$$\begin{aligned} \mathcal{P}(s_{\text{out}}) &= \mathcal{P}(w_1) * \mathcal{P}(w_2|w_1), \ldots, \mathcal{P}(w_n|w_1, \ldots, w_{n-1}) \\ s_{\text{out}} &= \text{Concat}(w_1, w_2, \ldots, w_{n-1}, w_n). \end{aligned} \tag{3}$$

3) Encoder–Decoder-Based PLMs: The training of encoder–decoder-based PLMs takes in a sentence s_{in} as an initial condition and then predicts each word in a generative manner

$$\begin{aligned} \mathcal{P}(s_{\text{out}}|s_{\text{in}}) &= \mathcal{P}(w_1|s_{\text{in}}) * \cdots * \mathcal{P}(w_n|s_{\text{in}}, w_1, \ldots, w_{n-1}) \\ s_{\text{out}} &= \text{Concat}(w_1, w_2, \ldots, w_{n-1}, w_n). \end{aligned} \tag{4}$$

Encoder-only PLM like BERT adopts the form of autoencoder, which is good at capturing the context features. When applying encoder-only PLMs to downstream tasks, it is usually necessary to add additional structures, such as multilayer perceptron (MLP) or conditional random fields (CRFs) to adapt PLMs to downstream tasks, including natural language understanding (NLU) tasks and natural language generation (NLG) tasks. However, encoder-only PLMs alone cannot meet the requirements of the PL framework, because the added extra structure is usually related to specific tasks, which will limit the form of prescription.

On the contrary, decoder-only PLMs based on autoregression, and encoder–decoder PLMs based on seq2seq, have the advantage to unify the pretraining task and downstream tasks. In this way, the prescription has higher versatility to cover a variety of tasks to stimulate the different language capabilities of PLMs. Specifically, in this article, we take the form of text-to-text transfer transformer (T5) [28], which is an encoder–decoder-based PLM.

B. Prescription by Unified Prompting

To release PLM's ability to apply to the legal field, we adopt the method of prompt tuning by reforming various legal tasks into one unified format.

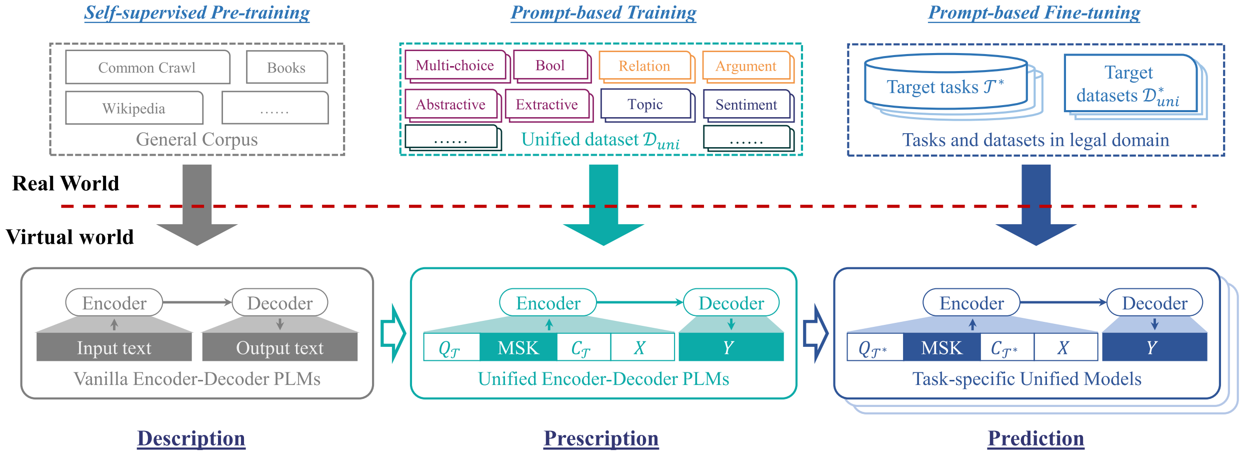


Fig. 2. HANOI-Legal learning pipeline with self-supervised pretraining, prompt-based training, and prompt-based fine-tuning for description, prescription, and prediction, respectively.

1) Unified Task Definition: With the flexible format, QA tasks can effectively express various types of NLP tasks [47]. For instance, a standard text CLS task can be represented as a QA format that asks for the text category. In this manner, the model completes the corresponding task by giving the answers to the questions. Therefore, our framework uses QA tasks as the unified task format for better compatibility with a broader range of datasets. Based on the PLMs' structure described in Section III-A, our framework handles all tasks in a text-to-text generation manner, that is, generating a new piece of text based on the input text.

According to the linguistic capabilities required for their tasks, the public datasets can be grouped into several categories. Let $\mathcal{T} = \{\mathcal{T}_n | n \in [1, N]\}$ represents N different task categories, each consisting of M_n datasets

$$\mathcal{D}^n_m = \{(x_i, y_i)\}, \quad m \in [1, M_n] \tag{5}$$

where x_i is the original text of a sample in the dataset, and y_i is the respective label or target text. Besides, the target task is denoted as $\mathcal{T}^*$, containing a small amount of labeled data. The samples of each collated dataset $\mathcal{D}^n_m$ are reformulated into the unified format and blended together into our unified dataset $\mathcal{D}_{\text{uni}}$

$$\mathcal{D}_{\text{uni}} = \bigcup_{n}^{N} \bigcup_{m}^{M_n} \left\{ \left(s^{\mathcal{T}_n}_{\text{in}}(x_i), s^{\mathcal{T}_n}_{\text{out}}(y_i) \middle| (x_i, y_i) \in \mathcal{D}^n_m \right) \right\} \tag{6}$$

where $s^{\mathcal{T}_n}_{\text{in}}$ and $s^{\mathcal{T}_n}_{\text{out}}$ are reformulated text of the input text x and target label y from the original dataset

$$s^{\mathcal{T}_n}_{\text{in}} = \text{Concat}\left(Q_{\mathcal{T}_n}, \langle\text{extra}\rangle, C_{\mathcal{T}_n}, x\right) \tag{7}$$
$$s^{\mathcal{T}_n}_{\text{out}} = \text{Concat}(\langle\text{extra}\rangle, y) \tag{8}$$

where $Q_{\mathcal{T}_n}$ represents the query, $C_{\mathcal{T}_n}$ is the additional candidate information, and $\langle\text{extra\_}i\rangle$ is a special token in the vocabulary of PLM, which varies in different PLMs. Of these, $Q_{\mathcal{T}_n}$ and $C_{\mathcal{T}_n}$ are the prompt templates corresponding to the task category $\mathcal{T}_n$. In some tasks, $C_{\mathcal{T}_n}$ is optional. The design of $(Q_{\mathcal{T}_n}, C_{\mathcal{T}_n})$ is related to its task category $\mathcal{T}_n$ and specific task objective. In Section III-B2, our framework presents a generic and simple design of the template.

On the basis of the operations above, our unified task is defined as follows:

$$\mathcal{T}_{\text{uni}}{:}\mathcal{M}_\theta(s_{\text{in}}) = s_{\text{out}}, \quad (s_{\text{in}}, s_{\text{out}}) \in \mathcal{D}_{\text{uni}} \tag{9}$$

where the model $\mathcal{M}_\theta$ is trained to generate the target text s_{out} based on the input text s_{in} under the unified format.

2) Unified Prompt Template Design: Corresponding to three commonly used linguistic capabilities, we group various NLP tasks into three categories, namely (**CLS**), (**IE**), and (**QA**). In other words, $\mathcal{T}$ can be simply defined as $\{\mathcal{T}_{\text{cls}}, \mathcal{T}_{\text{ie}}, \mathcal{T}_{\text{qa}}\}$. Table I shows several manual prompt templates $(Q_{\mathcal{T}_n}, C_{\mathcal{T}_n})$ designed for each category.

a) Classification: Q_{cls} queries the class of the input text x followed by a series of candidate class descriptions C_{cls}. For the consistency of format, we fix the number of candidates contained in C_{cls} without guaranteeing the inclusion of the correct answer, using *do not know* as an alternative, which increases the difficulty of the unified task. As the first example in Table II, each query Q_{cls} expresses the objective of the task in natural languages, such as the topic of text or the attitude of the comment, and the context C_{cls} provides the available candidates.

b) Information Extraction: Q_{ie} asks for specific information, such as trigger of events, relation, and property, requiring models to extract the correct segment from the inputs. The second example in Table II shows the reformulation of a sample from the IE task.

c) Question Answering: Q_{qa} is the original question followed by its options C_{qa} if they exist. For those unanswerable questions, we use *do not know* as their answer. The third and fourth examples in Table II show the reformulation of generative and selective QA tasks, respectively.

Based on the prompt templates above, various task-specific linguistic capabilities are aligned into a unified form: either extracting or summarizing the answer y of the question $Q_{\mathcal{T}_n}$ from the context $C_{\mathcal{T}_n}$ and x. The datasets of each category we use are described in Section IV-B.

C. Training Details

Aiming at minimizing the gap among the phases above, we adopt the same training objective in each phase.

TABLE I
PARTIAL PROMPT TEMPLATES FOR EACH CATEGORY

Category	**Task**	**Query** Q	**Context** C
CLS	Topic	What is the topic of the text? 以下文本的主题是什么?	$[description_1, description_2, ..., don'tknow]$. 选项：【主题描述1；主题描述2；……；不知道】。
	Sentiment	What is the attitude of this comment? 请问评论的态度怎么样?	positive, negative, neutral. 选项：积极；消极；中立。
		What about the attitude towards $[aspect]$? 以下文本中针对【方面】的态度如何?	positive, negative, neutral. 选项：积极；消极；中立。
	Event Type	What type of event happened? 请问以下事件属于哪种类型?	$[event\_type_1, event\_type_2, ..., don'tknow]$. 选项：【事件类型1；事件类型2；……；不知道】。
	Relation Type	What is the relationship of $[ent_1]$ and $[ent_2]$? 请问【实体1】和【实体2】有什么关系?	$[relation\_type_1, event\_type_2, ..., don'tknow]$. 选项：【关系类型1；关系类型2；……；不知道】。
IE	Trigger	What is the trigger of $[event\_type]$? 请问发生的【事件类型】的触发词是什么?	
	Argument	What is $[role\_type]$? 请问【事件属性】是什么?	
		What is the $[role\_type]$ of $[event\_type]$? 在发生的【事件类型】中，【事件属性】是什么?	
QA	Generative	Question: $[original\_question]$? 问题：【原始问题】？	
	Multi-choices	Question: $[original\_question]$? 问题：【原始问题】？	$[choice_1, choice_2, ..., don'tknow]$. 选项：【选项1，选项2，……，不知道】。

Specifically, given a single input s_{in} and its expected output s_{out}, we train the model using the teacher forcing loss

$$\mathcal{L}_{\text{tf}}(s_{\text{in}}, s_{\text{out}}) = -\frac{\sum_{i=1}^{\text{len}(s_{\text{out}})} \log p_\theta(s_{\text{out}}[i] | s_{\text{in}}; s_{\text{out}}[1{:}i])}{\text{len}(s_{\text{out}})} \tag{10}$$

where $s_{\text{out}}[i]$ denotes the ith token of s_{out}, and $s_{\text{out}}[1{:}i]$ represents all token before $s_{\text{out}}[i]$. We calculate the cross-entropy on the conditional probability distribution p_θ of $s_{\text{out}}[i]$ given by the PLMs based on preceding tokens $s_{\text{out}}[1{:}i]$ as the loss.

D. Data Labeling

To collect labeled data for task-specific fine-tuning, we setup a distributed operation platform to enable all team members to label the raw data concurrently. For the CLS task (CLS), we labeled 4204 events, including 2969 for training and 1235 for testing. For the IE task, we labeled 22 983 attributes, including 15 426 for training and 7557 for testing. Details can be found in Table III.

IV. EXPERIMENTS

A. Tasks in Legal Intelligence

Specifically, LegalAI covers three major tasks: Legal text CLS (LegalCLS), Legal IE (LegalIE), and Legal QA (LegalQA).

1) Civil Case Event CLS: LegalCLS aims to predict the label of the given legal text, such as case type, event class, relation category, and so on. Of these, legal judgment prediction (LJP) is the most characteristic and complex task, which requires models to analyze cases based on the existing laws and to speculate the judgments such as relevant statutes, charges, and penalties [48], [49], [50].

2) Civil Case Argument Extraction: LegalIE targets extracting specific valuable information from lengthy legal texts, such as key elements of the case present in the case description and the plaintiff's claims in legal documents [51]. As with general IE, LegalIE requires models to predict information like legal entities, event elements, and entity relationships that appear in a given legal text.

3) Legal Question Answering: LegalQA features a more flexible and diverse task format, requiring models to read and comprehend a legal passage and answer questions by selecting preset options or generating texts [52]. Specifically, most of these questions relate to the specific circumstances of a case [53] or ask about the interpretation of some legal concepts [54].

B. Datasets

We collected seven commonly used Chinese datasets from two Chinese evaluation benchmarks (CLUE[1] and LUGE[2]) and a legal domain competition (CAIL, Challenge of AI in Law[3]).

In addition, we adopted CivilEE, an event extraction dataset of civil cases annotated by ourselves, containing two subtasks: CivilEE-cls and CivilEE-args. CivilEE-cls requires models to classify cases into eight events that are frequently considered when handling civil cases, such as loan, bankruptcy, marriage, pledge, guarantee, and so on. Each event has several attributes, for example, a *guarantee* event has attributes such as the guarantor, creditor, obligor, guarantee type, content, and so on. There are 70 different attributes in CivilEE-args, requiring

[1] https://www.clue.ai/
[2] https://www.luge.ai/
[3] http://cail.cipsc.org.cn/

TABLE II
EXAMPLES OF UNIFIED TASK FORMAT REFORMULATION

No.	Category	Before		After	
1	CLS	Input X:	This is a strategy war game with western magic as the world-view background, combining wonderful RPG adventure and exciting SLG conquest.The game is based on building and building castles, raising heroes and carrying out siege and plunder, so that...	Input s_{in}:	**What is the topic of the text?** $< extra >$ ***Choice: Fantasy, Fiction, Sports, don't know.*** This is a strategy war game with western magic as the world-view background, combining wonderful RPG adventure and exciting SLG conquest.The game is ...
		Label l:	12 (Fantasy)	Output s_{out}:	$< extra >$ Fantasy
		Input X:	这是一款以西方魔幻为世界观背景的策略战争游戏，融合了精彩的RPG冒险和刺激的SLG征服。 游戏以建设和建造城堡、养成英雄、进行攻城掠地为主，使...	Input s_{in}:	文本的主题是什么？ $< extra >$ 候选项***:*** 幻想、小说、体育，不知道***.*** 这是一款以西方魔幻为世界观背景的策略战争游戏，融合了精彩的RPG冒险和刺激的SLG征服。 游戏以建设和建造城堡、养成英雄、进行攻城掠地为主，使...
		Label l:	12 (幻想)	Output s_{out}:	$< extra >$ 幻想
2	IE	Input X:	XXX announced that the Group expects to achieve an increase in loss attributable to owners of the Company by more than 170% for the year ending 31 March 2020, compared to a loss attributable to owners of the Company of approximately HK$16.1 million for the same period in 2019. The announcement stated ...	Input s_{in}:	**What is the net loss for this loss-making event?** $< extra >$ XXX announced that the Group expects to achieve an increase in loss attributable to owners of the Company by more than 170% for the year ending 31 March 2020, compared to a loss attributable to owners of the Company of approximately HK$16.1 million for the same period in 2019. The announcement stated ...
		Target t:	HK$16.1 million	Output s_{out}:	$< extra >$ Approximately HK$16.1 million
		Input X:	XXX宣布，集团预计截至2020年3月31日止年度，公司拥有人应占亏损将增加170%以上，而2019年同期公司拥有人应占亏损约为1610万港元。 公告称...	Input s_{in}:	这一亏损事件的净损失是多少？ $< extra >$ XXX宣布，集团预计截至2020年3月31日止年度，公司拥有人应占亏损将增加170%以上，而2019年同期公司拥有人应占亏损约为1610万港元。 公告称...
		Target t:	1610万港元	Output s_{out}:	$< extra >$ 大约1610万港元
3	QA	Input X:	The Shanghai Railway Museum is located at..., which was completed and opened to the public in August 2004. The entire museum includes an outdoor square area of about 1,300 square meters and the main museum building with a floor area of more than 3,000 square meters. The main building of the museum has 4 floors...	Input s_{in}:	**How large is the floor area of the main building of the Shanghai Railway Museum?** $< extra >$ The Shanghai Railway Museum is located at... The entire museum includes an outdoor square area of about 1,300 square meters and the main museum building with a floor area of more than 3,000 square meters. The main building of the museum ...
		Question q:	How large is the floor area of the main building of the Shanghai Railway Museum?	Output s_{out}:	$< extra >$ About 3,000 square meters
		Answer t:	About 3,000 square meters		
		Input X:	上海铁路博物馆位于......， 已于2004年8月建成并向公众开放。整个博物馆包括约1300平方米的室外广场和建筑面积超过3000平方米的博物馆主楼。博物馆的主楼有4层...	Input s_{in}:	请问上海铁路博物馆主楼的建筑面积有多大？ $< extra >$ 上海铁路博物馆位于......， 已于2004年8月建成并向公众开放。整个博物馆包括约1300平方米的室外广场和建筑面积超过3000平方米的博物馆主楼。博物馆的主楼有4层...
		Question q:	上海铁路博物馆主楼的建筑面积有多大?	Output s_{out}:	$< extra >$ 约3000平方米
		Answer t:	约3000平方米		
4	QA	Input X:	World Thrift Day, October 31, was established by the United Nations. The 2006 State of World Population report released by the United Nations Population Fund shows that the world's population has surpassed 6.5 billion and will reach 6.56 billion this year. In terms of global resource consumption alone, mankind must also be frugal. This is the far-reaching significance of setting October 31 every year as World Thrift Day.	Input s_{in}:	**Question: Which statement is correct?** $< extra >$ ***Choice:*** ***A. In 2005, ... B. There are many ... C. The world population ... D. The United Nations has established World Thrift Day.*** World Thrift Day, October 31, was established by the United Nations. The 2006 State of World Population report released by the United Nations Population Fund shows that the world's ...
		Question q:	Please select the option consistent with the article. A. In 2005, the world population was already 6.56 billion. B. There are many resources and does not need to be thrifty. C. The world population did not exceed 6.5 billion in 2006. D. The United Nations has established World Thrift Day.	Output s_{out}:	$< extra >$ D. The United Nations has established World Thrift Day.
		Label l:	D		
		Input X:	10月31日的世界节俭日是由联合国设立的。联合国人口基金会发布的2006年世界人口状况报告显示，世界人口已超过65亿，今年将达到65.6亿。仅就全球资源消耗而言，人类也必须节俭。这就是将每年的10月31日定为世界节俭日的深远意义所在。	Input s_{in}:	问题：哪种说法是正确的？ $< extra >$ 候选项***:*** ***A.*** 在***2005***年，世界人口已经达到***65.6***亿。 ***B.*** 有很多资源，不需要节俭。 ***C. 2006***年，世界人口没有超过***65***亿。 ***D.*** 联合国已经设立了世界节俭日。 10月31日的世界节俭日是由联合国设立的。联合国人口基金会发布的2006年世界人口状况报告显示，世界人口已超过65亿，今年将达到65.6亿。仅就全球资源消耗而言，人类也必须节俭。这就是将每年的10月31日定为世界节俭日的深远意义所在。
		Question q:	请选择与文章内容一致的选项。 A. 在2005年，世界人口已经达到65.6亿。 B. 有很多资源，不需要节俭。 C. 2006年，世界人口没有超过65亿。 D. 联合国已经设立了世界节俭日。		
		Label l:	D	Output s_{out}:	$< extra >$ D. 联合国已经设立了世界节俭日。

models to extract the corresponding arguments from the case description.

We grouped these datasets into three categories and reformulated them with the template mentioned in Section III-B2. For CLS tasks, we adopt an application CLS dataset **IFLYTEK** [55], an aspect-level sentiment CLS dataset **ASAP_ASPECT** [56], and the event CLS part of CivilEE (**CivilEE-cls**). For IE tasks, we used **DuEE-Fin** [57], a financial domain event extraction dataset, and the argument extraction part of CivilEE (**CivilEE-args**). For QA tasks, we used three datasets from the general domain (**CMRC** [55], **C3** [55], and **DuReader** [58]) and **CJRC** [52], a dataset from the legal domain. After the reformulation, the details of each dataset are shown in Table III, where each category was limited to a similar scale. We targeted the legal domain with experiments on all task categories, using CJRC, CivilEE-cls, and CivilEE-args for fine-tuning and testing, while others for training.

C. Experimental Settings

1) Baseline: We used the prompt-based approach adopted in many state-of-the-art related studies [38], [40], [41], [59]

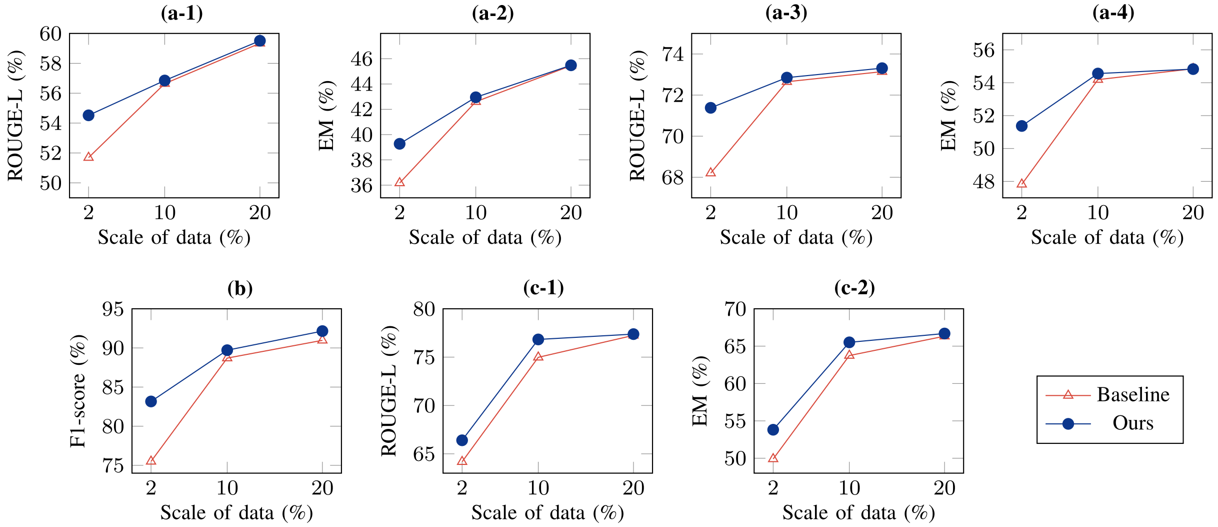


Fig. 3. Performance curves on target tasks as training data increases, where the blue curve is of the baseline and the red curve is ours. On each task, our approach yields considerable improvement. The less training data available, the more obvious improvement our approach achieves. (a-1) CJRC. (a-2) CJRC. (a-3) CJRC w/o imp. (a-4) CJRC w/o imp. (b) CivilEE-cls. (c-1) CivilEE-args. (c-2) CivilEE-args.

TABLE III
STATISTICS OF VARIOUS DATASETS INCLUDED IN THIS STUDY

Category	Dataset name	Domain	Train	Test
CLS	IFLYTEK[55]	General	12133	—
	ASAP-ASPECT[56]	General	45000	—
	CivilEE-cls	Legal	2969	1235
IE	DuEE-Fin[57]	Finance	48891	—
	CivilEE-args	Legal	15426	7557
QA	CMRC18[55]	General	10142	—
	C3[55]	General	6013	—
	DuReader-checklist[58]	General	3000	—
	CJRC[52]	Legal	44383	6000

as our *baseline*, which was only trained on the corresponding unified target dataset.

2) Settings: To precisely evaluate our proposed method, we conducted experiments from two perspectives.

1) *Low-Resource Settings:* We partitioned the training set of target tasks into different scales (0%, 2%, 10%, and 20%) to study the performance in the case of insufficient labeled data.
2) *Ablation Settings:* We trained several models (CLS, QA, IE, and Hybrid) using datasets from each category to evaluate the contribution of different task categories.

3) Metrics: During the testing phase, we chose different metrics depending on the task category. For the CivilEE-cls task, we converted the generated class descriptions into their respective labels and evaluated them with a micro-F1 score. Specifically, if the generated text only contains keywords of one label, we consider the model's prediction as that label; otherwise, it is considered a failed prediction. For CJRC and CivilEE-args tasks, we used ROUGE-L and exact match (EM), which directly evaluates the similarity between generated answers and ground truths.

D. Implementation Details

We used *Randeng-T5-784M* model publicly available on Huggingface[4] as our base model. Based on mT5-large [28], *Randeng-T5-784M* has been trained on the 180 GB Chinese corpus under the pretraining objective of span corruption, equipping it with a decent linguistic capability in Chinese [60].

By default, we adopted a learning rate of 5e-6 for training all models with the same trend of linear warm-up and decay. We used a single NVIDIA A100 GPU for training, setting the batch size to 2, the gradient accumulation to eight steps, and the maximum sequence length to 512. In the training phase on the unified dataset, we randomly selected 5% of the training data for model selection. When fine-tuning the target dataset, models were initialized from the best one in the previous phase and trained on the corresponding data for two epochs without model selection. In the testing phase, models generated answers by greedy search.

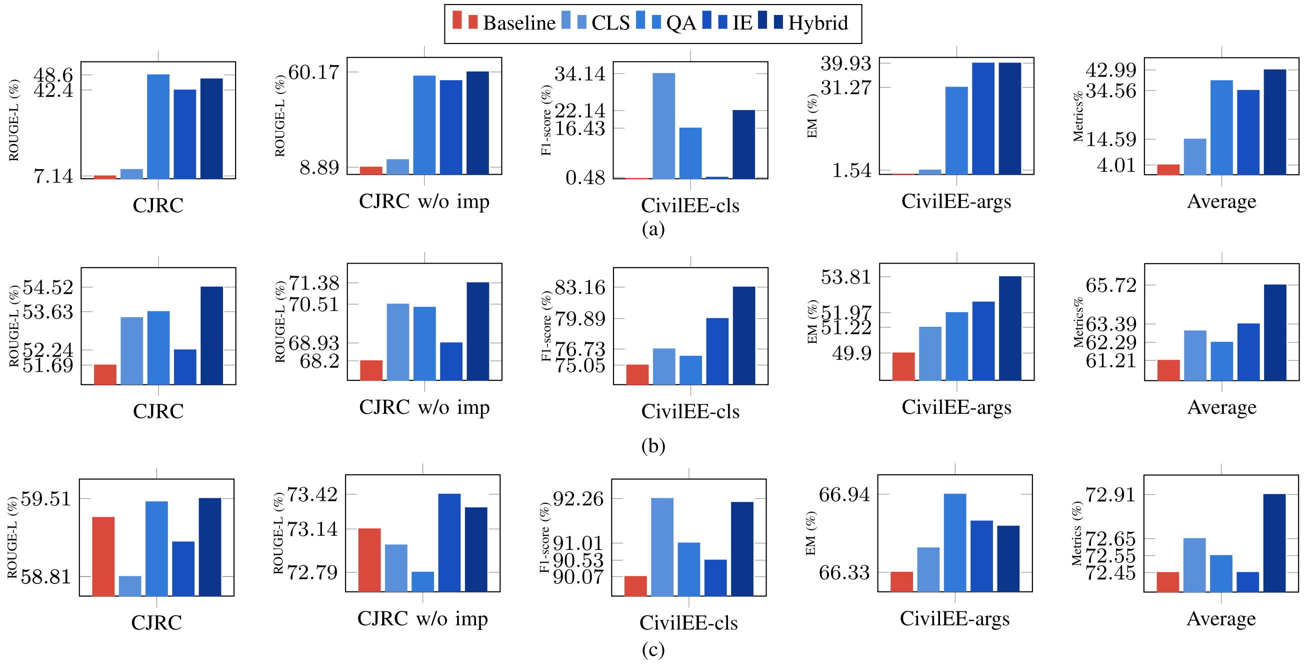
E. Results and Analysis

1) Performance in Low-Resource Settings: Detailed results on CJRC and CivilEE in low-resource settings are summarized in Tables IV and Fig. 3, showing that our method can effectively enhance the model in the case of insufficient labeled data.

In the zero-shot scenario, models trained with our method obtain the best results, significantly outperforming the baseline. As shown in Table IV, the baseline has a little zero-shot ability due to the similarity between our unified task format and the pretraining task. However, it is not effective when using the metrics that certain requirements for response format, such as the F1 score and EM. Benefiting from our framework, the model boosts up to 22.13% of the F1 score on CivilEE-cls, as well as 46.35% and 41.46% of ROUGE-L on CivilEE-args and CJRC, respectively. This demonstrates that PLMs trained on the unified dataset using our well-designed prescriptions can perform promising generalizations on various legal tasks.

Fig. 3 shows the performance curves of the model on each target task as training data scales up. By training on small amounts of data at 2%, 10%, and 20%, the baseline shows strong learning capabilities with its large scale of parameters

[4] https://huggingface.co/IDEA-CCNL/Randeng-T5-784M



(a)

(b)

(c)

Fig. 4. Ablation results on the contribution of data from different task categories to the model capabilities. We ablate according to the task category of training data, including CLS, QA, and IE. The experiments compare the performance of each model under different scales of fine-tuning datasets on target tasks. (a) Scale of data = 0%. (b) Scale of data = 2%. (c) Scale of data = 20%.

TABLE IV

PERFORMANCE IN THE ZERO-SHOT SCENARIO

		Baseline	**Ours**	
CJRC	ROUGE-L	7.14	46.93	+39.79
	EM	0.05	30.70	+30.65
w/o impossible	ROUGE-L	8.89	60.17	+51.28
	EM	0.07	39.80	+39.73
CivilEE-cls	F1-score	0.01	22.14	+22.13
CivilEE-args	ROUGE-L	11.90	58.25	+46.35
	EM	0.01	39.92	+39.91

TABLE V

PERFORMANCE IN HIGH-RESOURCE SETTINGS, WHERE CIVILEE-CLS USES THE F1 SCORE (%) AND OTHERS USE ROUGE-L (%). THE BEST AND SECOND BEST RESULTS ARE **BOLDED** AND UNDERLINED, RESPECTIVELY

	Baseline	**CLS**	**QA**	**IE**	**Hybrid**
CJRC	62.55	**63.35**	62.70	62.96	<u>63.11</u>
CJRC w/o	73.47	<u>73.84</u>	**73.89**	73.56	73.70
CivilEE-cls	92.15	<u>92.79</u>	92.31	**92.96**	92.63
CivilEE-args	78.17	78.25	<u>78.26</u>	78.18	**78.38**

and our prescriptions. Nevertheless, our method still yielded considerable improvements to models, where the less training data available, the more obvious improvement our approach achieves. Specifically, the improvement ranges from 0.17% to 2.83% ROUGE-L on CJRC (a-1), 0.17% to 3.18% ROUGE-L on CJRC w/o impossible (a-3), 1.18% to 8.11% F1 score on CivilEE-cls (b), and 0.36% to 3.91% on CivilEE-args (c-2). These indicate that with our prescriptions, the model can apply its acquired ability more effectively to low-resource tasks in the legal domain.

2) Ablation Study: Moreover, we conducted ablation experiments to explore the contribution of data from different task categories to the model capabilities under our prescriptions.

Fig. 4 shows that models trained with data from a single category (CLS, QA, IE) tend to perform better on their respective types of tasks, which implies a greater contribution of capabilities that match the requirements of the target tasks. In particular, these models can often be boosted on another task category, suggesting that our well-designed unified task format creates an overlap between tasks and enhances the transferability across them. Eventually, models trained with data from all categories (hybrid) can integrate various linguistic capabilities and achieve better overall performance in a wider range of situations.

3) Performance in High-Resource Settings: In addition, we carried out experiments in high-resource settings, training on the complete dataset for one epoch. As shown in Table V, each model achieves comparable results to the baseline with sufficient labeled data. Although the baseline gradually approaches the performance of our models as the data size increases, our models still have advantages by virtue of their preacquired capabilities.

V. CONCLUSION

In this article, we initiatively applied the PL framework to tackle data scarcity challenges in the legal domain. The proposed HANOI-Legal has several features. First, it takes PFM as the artificial system. Experiments showed that the PFM is a condensed knowledge base by learning from the large open-source corpus about the real world. Second, it takes unified prompt (UniPrompt) learning as a prescriptive method for the PFM. The UniPrompt enables the PFM to leverage various open datasets and enhances their performance on domain-specific tasks with low resources. Third, the learning process, especially in the fine-tuning stage for specific legal tasks, was organized by a distributed operation involving all team members, including a few lawyers, to do labeling concurrently. These three features show that the proposed framework is a typical HANOI form, which realizes organized intelligence based on the natural world and artificial system with humans in the loop. The experiments verified the proposed HANOI-Legal as an effective framework for legal tasks even in low-resource scenarios.

Although HANOI-Legal shows superiority, it is only a preliminary version. Future work includes several aspects of expansion. For example, taking a larger model is the first option worth trying. This article uses a pretraining model based on T5, which has a relatively small amount of parameters and does not yet have the ability to emerge from LLMs. It can be expected that using a large model with more than 100 B parameters will achieve better results. Corresponding to increasing the number of parameters, in-context learning and other methods can be used to reduce the huge computing consumption required for training or fine-tuning the model.

References

[1] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[2] F.-Y. Wang, Q. Miao, X. Li, X. Wang, and Y. Lin, "What does ChatGPT say: The DAO from algorithmic intelligence to linguistic intelligence," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 3, pp. 575–579, Mar. 2023.

[3] C. Zhou et al., "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," 2023, *arXiv:2302.09419*.

[4] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 1872–1897, 2020.

[5] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.

[6] P. F. Christiano et al., "Deep reinforcement learning from human preferences," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[7] L. Li, Y. Lin, N. Zheng, and F.-Y. Wang, "Parallel learning: A perspective and a framework," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 3, pp. 389–395, Jul. 2017.

[8] F.-Y. Wang, "Parallel intelligence in metaverses: Welcome to Hanoi!" *IEEE Intell. Syst.*, vol. 37, no. 1, pp. 16–20, Jan. 2022.

[9] Q. Miao, W. Zheng, Y. Lv, M. Huang, W. Ding, and F.-Y. Wang, "DAO to Hanoi via DeSci: AI paradigm shifts from AlphaGo to ChatGPT," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 877–897, Apr. 2023.

[10] F.-Y. Wang, W. Ding, R. Qin, and B. Hu, "Parallel philosophy for MetaOrganizations with MetaOperations: From Leibniz's monad to HanoiDAO," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 3, pp. 658–666, Jun. 2022.

[11] W. Ding et al., "DeSci based on web3 and DAO: A comprehensive overview and reference model," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 5, pp. 1563–1573, Oct. 2022.

[12] F.-Y. Wang et al., "The DAO to DeSci: AI for free, fair, and responsibility sensitive sciences," *IEEE Intell. Syst.*, vol. 37, no. 2, pp. 16–22, Mar. 2022.

[13] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit legal system: A summary of legal artificial intelligence," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5218–5230.

[14] C. Xiao et al., "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, Jan. 2021.

[15] Y. Huang, X. Shen, C. Li, J. Ge, and B. Luo, "Dependency learning for legal judgment prediction with a unified text-to-text transformer," 2021, *arXiv:2112.06370*.

[16] F.-Y. Wang, X. Wang, L. Li, and L. Li, "Steps toward parallel intelligence," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 4, pp. 345–348, Oct. 2016.

[17] F.-Y. Wang, "Artificial societies, computational experiments, and parallel systems: A discussion on computational theory of complex social-economic systems," *Complex Syst. Complex. Sci.*, vol. 1, no. 4, pp. 25–35, 2004.

[18] F. Fei-Yue, "Parallel system methods for management and control of complex systems," *Control Decis.*, vol. 19, no. 5, pp. 485–489, 2004.

[19] F.-Y. Wang, "Computational theory and methods for complex systems," *China Basic Sci.*, vol. 6, no. 41, pp. 3–10, 2004.

[20] F.-Y. Wang, "The emergence of intelligent enterprises: From CPS to CPSS," *IEEE Intell. Syst.*, vol. 25, no. 4, pp. 85–88, Jul. 2010.

[21] X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F.-Y. Wang, "From features engineering to scenarios engineering for trustworthy AI: I&I, C&C, and V&V," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 18–26, Jul. 2022.

[22] Q. Miao, Y. Lv, M. Huang, X. Wang, and F.-Y. Wang, "Parallel learning: Overview and perspective for computational learning across Syn2Real and Sim2Real," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 3, pp. 603–631, Mar. 2023.

[23] X. Li et al., "A novel framework to generate synthetic video for foreground detection in highway surveillance scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 5958–5970, Jun. 2023.

[24] X. Li, Y. Tian, P. Ye, H. Duan, and F.-Y. Wang, "A novel scenarios engineering methodology for foundation models in metaverse," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 4, pp. 2148–2159, Apr. 2023.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[26] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[27] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[28] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[29] J. Wei et al., "Emergent abilities of large language models," 2022, *arXiv:2206.07682*.

[30] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.

[31] T. Kojima et al., "Large language models are zero-shot reasoners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 22199–22213.

[32] Q. Dong et al., "A survey on in-context learning," 2022, *arXiv:2301.00234*.

[33] S. Min et al., "Rethinking the role of demonstrations: What makes in-context learning work?" 2022, *arXiv:2202.12837*.

[34] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.

[35] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4582–4597.

[36] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "PTR: Prompt tuning with rules for text classification," *AI Open*, vol. 3, pp. 182–192, Jan. 2022.

[37] G. Qin and J. Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 5203–5212.

[38] D. Khashabi et al., "UNIFIEDQA: Crossing format boundaries with a single QA system," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 1896–1907.

[39] X. Liu et al., "GPT understands, too," 2021, *arXiv:2103.10385*.

[40] R. Zhong, K. Lee, Z. Zhang, and D. Klein, "Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2021, pp. 2856–2878.

[41] H. Xu et al., "ZeroPrompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization," 2022, *arXiv:2201.06910*.

[42] V. Sanh et al., "Multitask prompted training enables zero-shot task generalization," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, 2022, p. 216.

[43] H. W. Chung et al., "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.

[44] D. M. Ziegler et al., "Fine-tuning language models from human preferences," 2019, *arXiv:1909.08593*.

[45] J. Wu et al., “Recursively summarizing books with human feedback,” 2021, *arXiv:2109.10862*.

[46] T. Gao, A. Fisch, and D. Chen, “Making pre-trained language models better few-shot learners,” in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 3816–3830.

[47] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, “The natural language decathlon: Multitask learning as question answering,” 2018, *arXiv:1806.08730*.

[48] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, “Learning to predict charges for criminal cases with legal basis,” 2017, *arXiv:1707.09168*.

[49] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, “Legal judgment prediction via topological learning,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3540–3549.

[50] W. Yang, W. Jia, X. Zhou, and Y. Luo, “Legal judgment prediction via multi-perspective bi-feedback network,” 2019, *arXiv:1905.03969*.

[51] F. Yao et al., “LEVEN: A large-scale Chinese legal event detection dataset,” in *Proc. Findings Assoc. Comput. Linguistics (ACL)*, 2022, pp. 183–201.

[52] X. Duan et al., “CJRC: A reliable human-annotated benchmark dataset for Chinese judicial reading comprehension,” in *Proc. 18th China Nat. Conf. Chin. Comput. Linguistics (CCL)*. Kunming, China: Springer, Oct. 2019, pp. 439–451.

[53] Y. Zhou, L. Liu, Y. Chen, R. Huang, Y. Qin, and C. Lin, “A novel MRC framework for evidence extracts in judgment documents,” *Artif. Intell. Law*, vol. 30, pp. 1–17, Jan. 2023.

[54] H. Zhong et al., “JEC-QA: A legal-domain question answering dataset,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 9701–9708.

[55] L. Xu et al., “CLUE: A Chinese language understanding evaluation benchmark,” in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4762–4772.

[56] J. Bu et al., “ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2021, pp. 2069–2079.

[57] C. Han et al., “DuEE-Fin: A large-scale dataset for document-level event extraction,” in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.* Guilin, China: Springer, 2022, pp. 172–183.

[58] W. He et al., “DuReader: A Chinese machine reading comprehension dataset from real-world applications,” in *Proc. Workshop Mach. Reading Question Answering*, 2018, pp. 37–46.

[59] I.-H. Hsu et al., “DEGREE: A data-efficient generation-based event extraction model,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 1890–1908.

[60] J. Zhang et al., “Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence,” 2022, *arXiv:2209.02970*.

Zhuoyang Song received the B.S. degree in computer science and technology from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2020. He is currently pursuing the master’s degree with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing.

His current research interests include natural language processing, large language model, and controlled text generation.

Min Huang (Member, IEEE) received the Ph.D. degree in computer sciences from Wuhan University, Wuhan, China, in 2007.

From 2017 to 2018, she was a Visiting Scholar with the School of Informatics, University of Edinburgh, Edinburgh, U.K. She is currently an Associate Professor with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. Her current research interests include image processing, knowledge engineering, big data, and deep learning.

Qinghai Miao (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2007.

From 2017 to 2018, he was a Visiting Scholar with the School of Informatics, University of Edinburgh, Edinburgh, U.K. He is currently an Associate Professor with the School of Artificial Intelligence, University of Chinese Academy of Sciences. His current research interests include parallel intelligence, machine learning, computer vision, and intelligent transportation systems.

Fei-Yue Wang (Fellow, IEEE) received the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined The University of Arizona, Tucson, AZ, USA, in 1990, where he became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Chinese Talents Program from the State Planning Council, and in 2002, was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS, and the Vice President at the Institute of Automation, CAS in 2006. He found CAS Center for Social Computing and Parallel Management in 2008 and became the State Specially Appointed Expert and the Founding Director with the State Key Laboratory for Management and Control of Complex Systems in 2011. His current research interests include methods and applications for parallel intelligence, social computing, and knowledge automation.

Dr. Wang is a fellow of INCOSE, IFAC, ASME, and AAAS. In 2007, he received the National Prize in Natural Sciences of China, numerous best papers awards from the IEEE TRANSACTIONS, and became an Outstanding Scientist of ACM for his work in intelligent control and social computing. He received the IEEE ITS Outstanding Application and Research Awards in 2009, 2011, and 2015, respectively, the IEEE SMC Norbert Wiener Award in 2014, and became the IFAC Pavel J. Nowacki Distinguished Lecturer in 2021. Since 1997, he has been serving as the General or Program Chair for more than 30 IEEE, INFORMS, IFAC, ACM, and ASME conferences. He was the President of the IEEE ITS Society, from 2005 to 2007, the IEEE Council of RFID, from 2019 to 2021, the Chinese Association for Science and Technology, USA, in 2005, the American Zhu Kezhen Education Foundation, from 2007 to 2008, the Vice President of the ACM China Council, from 2010 to 2011, the Vice President and the Secretary General of the Chinese Association of Automation, from 2008 to 2018, the Vice President of IEEE Systems, Man, and Cybernetics Society, from 2019 to 2021. He was the Founding Editor-in-Chief (EiC) of the International Journal of Intelligent Control and Systems, from 1995 to 2000, IEEE ITS Magazine, from 2006 to 2007, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, from 2014 to 2017, *China’s Journal of Command and Control*, from 2015 to 2021, and *China’s Journal of Intelligent Science and Technology*, from 2019 to 2021. He was the EiC of the IEEE INTELLIGENT SYSTEMS, from 2009 to 2012, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, from 2009 to 2016, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, from 2017 to 2020. Currently, he is the President of CAA’s Supervision Council and the EiC of IEEE TRANSACTION ON INTELLIGENT VEHICLES.