International Research Journal of Multidisciplinary Sciences ISSN:AWAITED
VOL-1 ISSUE-7 July 2025 PP:1-10

Delegating Legal Reasoning: An Agentic Approach to Judgment Prediction

[1]*BhawnaKaushik*,[2]*Yazdani hasan*

1.bhawna.kaushik@niu.edu.in, NoidaInternationalUniversity

2.yazhassid@gmail.com, NoidaInternationalUniversity

Abstract

Legal judgment prediction (LJP) has become increasingly important in the legal field.In this paper, we identify that existing large language models (LLMs) have significant problemsofinsufficientreasoningduetoalack of legal knowledge.Therefore, we introduce GLARE,anagenticlegalreasoningframework thatdynamicallyacquireskeylegalknowledge byinvokingdifferentmodules,therebyimproving the breadth and depth of reasoning.Experimentsconductedonthereal-worlddataset verifytheeffectivenessofourmethod. Furthermore,thereasoningchaingeneratedduringthe analysis process can increase interpretability and provide the possibility for practical applications.

1 Introduction

Legal judgment prediction (LJP) is an important task in legal natural language processing (NLP), aims to make correct judgment predictions based onthecase'sfactdescription(Liuetal.,2023).The judgment predictions include law articles, charges, and terms of penalty (Xu et al., 2024).This task not only provides judgment references to lawyers and judges, as well as providing legal consulting servicestothegeneralpublic(Luoetal.,2017;Shulayeva et al., 2017; McGinnis and Pearce, 2013).

Recently,largereasoningmodels(LRMs)have made remarkable progress across in reasoning-intensive tasks, including multi-hop question answeringandstrategicplanning(Wangetal.,2024b; Choietal.,2025).Thesemodelscanperformmulti-step reasoning that mimics human thinking (Fuetal.,2022).Intuitively,LJPappearstobeanideal fitforsuchmodels.Legaldecision-makingofteninvolvescomparingmultiplecandidatecharges,evaluatingwhethereachsatisfiesthelegalcriteria,and narrowingdowntothemostappropriateonebased onthecasefacts.Asaresult,itisnaturaltoexpect thatstrongreasoningmodelswouldleadtomajor improvements in LJP.

However, existing reasoning models fail to delivertheexpectedbreakthroughsinLJP.Inpractice, they tend to predict the most likely charges without comparing them to similar alternatives, and their reasoning chains are often short and lacking inmeaningfulintermediatesteps. Theseissuesbecome especially clear in cases involving rare or confusing charges, where accurate judgment dependsonsubtledistinctionsandcarefulreasoning. Although models may produce step-by-step outputsinsuchscenarios,thereasoningoftenstaysat asurfacelevel,focusingonpatternmatchingrather than legal principles.

We argue that the main reasonforthelimitedperformance of reasoning models in legal judgment tasks is not a lack of reasoning ability, but a lack of the specialized knowledge that legal reasoning depends on (Yuan et al., 2024).Effective legal analysisrequireslong-taillegalknowledge,suchas determiningtheapplicabilityofspecificstatutes.In somecases,thisknowledgeisevenabsentfromofficiallegaltexts.Whensuchinformationismissing, modelsstruggletoproducecompleteandtrustwor- thy reasoning chains as shown in Figure 1. These observationshighlighttheneedfordomain-specific knowledgeaugmentationmechanismsthatcandynamicallysupply essential information during the reasoning process.

To address the knowledge gaps in legal reasoning, we propose GLARE (*A**G**entic* *Leg**A**l* ***R**easoning* *Fram**E**work*), a modular system that enables language models to dynamically acquire key legal knowledge to improve the breadth and depth of reasoning. First, the **Charge Expansion Module** (CEM) expands a diverse set of confusingchargesbyleveragingmultiplesignals,suchas legal structure and historical co-occurrence. This helpsthemodelcompareawiderrangeofcandi-

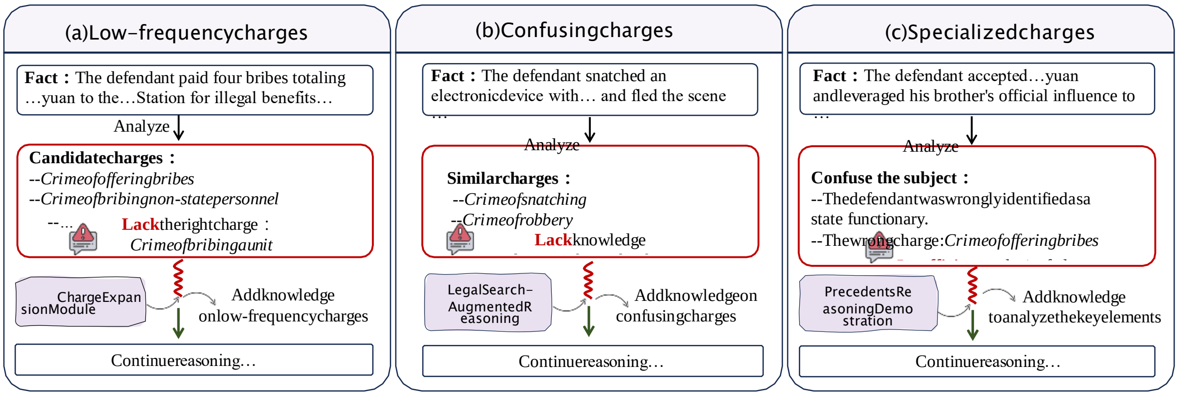

Figure1: Lackofknowledgeinthreeaspects: (a)Lackknowledgeoflow-frequencycharges. (b)Lackknowledge ofconfusingcharges.(c)Lackknowledgetoanalyzethekeyelementsofthechargeswithstrongprofessionalism.

dates and avoid premature conclusions.Second, the**PrecedentsReasoningDemonstration**(PRD) module is built on reasoning paths that are constructedofflinefromreallegalcases.Duringinference,themodelretrievesthemostrelevantprecedentsthroughsemanticsearchandlearnsfromtheir reasoning chains via in-context learning. Finally, the**LegalSearch-AugmentedReasoning**(LSAR) moduleallowsthemodeltodetectknowledgegaps and retrieve supporting legal information when needed.We guide the model to focus its search ondifferencesbetweensimilarchargesanddetails ofhowspecificlawsapply,ratherthangeneralcase facts. Retrieved content is structured and injected intothereasoningprocesstosupportmoreaccurate conclusions. Byintegratingessentiallegalknowledge, the model achieves more trustworthy and transparent judgment prediction.

Followingpriorworkinlegaljudgmentprediction,weconductexperimentsontwopubliclyavailablereal-worldlegaldatasets.Experimentalresults show that our method consistently outperforms a rangeofstrongbaselines. Notably,itachievessubstantialimprovementsonchallengingcasesinvolvingconfusinganddifficultcharges,wherelong-tail legalknowledgeiscrucial. Thesegainsstemfrom ourapproach’sabilitytoeffectivelyenrichandincorporate relevant legal knowledge.

Insummary,ourcontributionsareasfollows:

(1) WeintroduceGLARE,anagenticframework for legal judgment prediction that enhances reasoningbydynamicallyintegratinglegalknowledge throughout the decision-making process.

(2) Wedesignthreecomplementarymodulesto enrich the model’s reasoning process by expandingcandidatecharges,leveragingreal-worldprecedents,andinjectingretrievedlegalknowledge.

(3) Extensive experiments on two real-world datasets show that GLARE significantly outperforms strong baselines, with especially notable gains on cases requiring crucial legal knowledge.

2 RelatedWork

LegaljudgmentpredictionLegaljudgmentprediction has experienced significant development andbecomeanincreasinglycrucialNLPtask.Earlierresearch(Segal,1984)reliedonartificiallydesigned features to capture information from legal texts.Suleaetal.,2017appliedtraditionalmachine learningmethodstopredictthelegaljudgment.Recent advances in deep learning (Xu et al., 2020; ZhangandDou,2023)havemotivatedresearchers toleverageneuralnetworksforautomatedtextrepresentation learning. Recently, LLMs has further promotedtheprogressofLJP(Dengetal.,2024a), andseveralstudies(Wuetal.,2023;PengandChen, 2024) employ Retrieval-Augmented Generation (RAG) technology (Zhao et al., 2024) to enhance LLMsbyincorporatingexternallegalknowledge. However,existingLLM-basedmethodsstruggleto utilizecomprehensivelegalknowledge(Feietal., 2023)andrefertothewayofprecedentreasoning toanalyzecases. Inthiscontext,wemakefulluse of external knowledge and precedents.

Reasoning skills in language models Recent workhasimprovedLLMs’reasoningthroughbet- ter prompting techniques (Sahoo et al., 2024).Wei et al.(2022) showed that chain-of-thought prompting can explicitly guide LLMs to reason step by step.In the legal domain specifically,LoT (Jiang and Yang, 2023) proposed legal syl- logism reasoning to improve performance on LJP

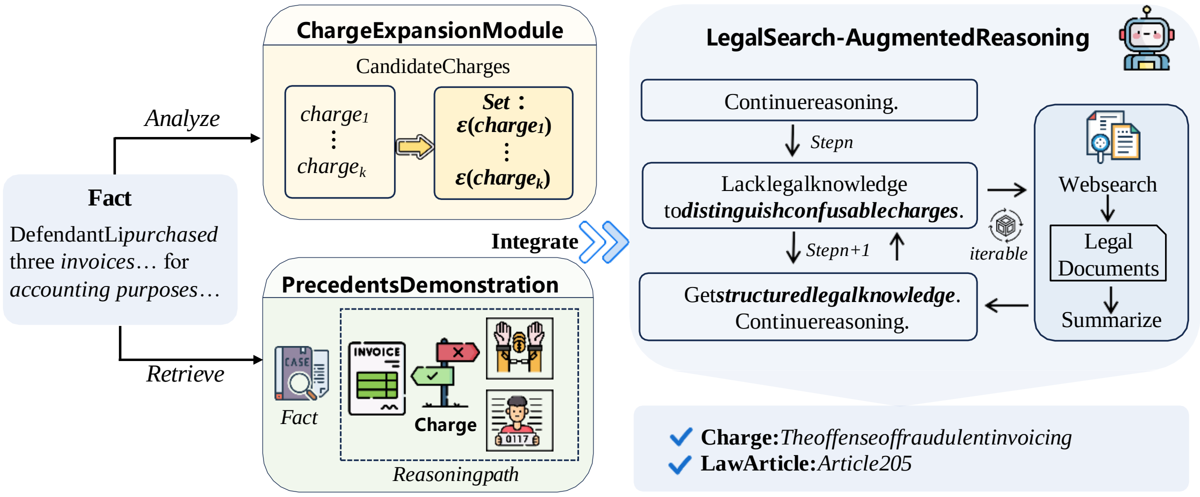

Figure2: Overviewofouragenticlegalreasoningframework. LLMscanutilizethreeexternalmodulestoacquire knowledge:ChargeExpandModuleexpandsadiversesetofcharges,precedentsretrievedfromofflinebuiltdatabase canprovidein-contextlearning,LegalSearch-AugmentedReasoningallowsthemodeltodetectknowledgegaps and retrieve supporting legal information.

task.ADAPT (Deng et al., 2024b) further establishedacomprehensiveworkflowforLJPthatenablesdiscriminativereasoninginLLMs. However, these approaches primarily rely on the LLMs intrinsiccapabilities,whichinherentlyconstrainthe reasoningbreadthandthedepthofanalysis(Zhang, 2024; Ke et al., 2025). Therefore, we propose an agenticlegalreasoningframeworktodynamically acquirekeylegalknowledgetoimprovethebreadth and depth of reasoning.

3 Methodology

Preliminaries

Wefirstformallydefinelegaljudgmentprediction. Givenacasefactdescription$f$,themodelwillanalyzeandpredictthefinaljudgmentresultsincludingtherelevantlawarticles,theconvictedcharges and the term of imprisonment for the defendent. Following previous works (Shui et al., 2023; Weiet al., 2025), we exclude the task of sentencing prediction from our scope as its subjective nature bringschallengesthatarenotwellalignedwiththe current capabilities of large language models.

Inthiswork,wetreatlargelanguagemodelsas agentic legal reasoners that can dynamically acquireandincorporateexternallegalknowledgeto enhance their analysis. Rather than relying solely onparametricknowledge,ourapproachequipsthe model with access to external modules, enabling it to enrich its reasoning with case-specific legal context.Givenacasefactdescription$f$anda setofexternal modules M, the model performs step-by-stepanalysistoconstructacoherentreasoning chain R and arrive at a final judgment predictionp. Weformalizethisprocessasamapping: $(f, \mathrm{M}) \to (\mathrm{R}, p)$.

AgenticLegalReasoningFramework

We propose GLARE, an agentic legal reason-ing framework that autonomously invokes exter- nal modules to support comprehensive and informedjudgmentprediction.AsshowninFigure2, GLAREfollowsastructuredthree-stagereasoning pipeline:

1. **Charge Expansion:** The model begins by analyzingthecasefactsandgeneratingpreliminary candidatecharges.Topreventprematurenarrowing of the decision space, it triggers the Charge ExpansionModuletosupplementtheinitialcandidates with legally similar charges.

2. **Precedent-Enhanced Reasoning:** The model retrieves relevant precedents from an offline-constructed database that includes fact descriptionsandsynthesizedreasoningchains.Thereasoning chains were constructed in advance to illustratethekeydistinctionsbetweenconfusing charges.Theseprecedentsserveascase-specific reasoningdemonstrations,helpingthemodelbetter understand how similar legal criteria apply and guiding it through more precise reasoning via in-context learning.

3. **Iterative Search-augmented Reasoning:** As themodelreasonsthrougheachcandidatecharge, it dynamically identifies knowledge gaps such asmissinglegaldefinitionsandcharge-specific thresholds.Ratherthantreatingretrievalasaone-time step, the model interleaves reasoning and retrievalinaloop.Retrievedresultsareinjected back into the reasoning context, enabling the modeltorefineitscurrentanalysis. Thisiterative process continues until the model has collected sufficient knowledge to complete its reasoning and reach a final judgment.

The three modules collaboratively supplement legal knowledge and extend the legal reasoning chain. Next,wewillintroducethesethreemodules in detail.

ChargeExpansionModule

Toenablechargecomparisonandavoidpremature conclusions,weexpandeachcandidatechargeby retrievingrelatedcharges.Theexpansionisbased ontwocomplementaryperspectives:legalstructure and historical co-occurrence.

LegalStructure-basedExpansion. TheCriminalLawisorganizedintochapters,eachrepresenting a specific legal interest or domain.Charges withinthesamechaptertypicallydifferinsubtlelegalcriteria,whilechargesacrossdifferentchapters may involve similar actions or consequences but fallunderdistinctlegalcategories.Tocaptureboth fine-grained intra-domain distinctions and cross-domainconceptualsimilarities,weretrieverelated charges from both within the same chapter and across different chapters.

Specifically, for a given charge c, we use the pretraineddenseretrieverBGE(Xiaoetal.,2024) tofindthetop-k mostsimilarchargesfrom(a)the same chapter and (b) other chapters:

$$E_1(c)=\mathrm{topk}_{\mathrm{same}}(c)\cup\mathrm{topk}_{\mathrm{diff}}(c),\quad(1)$$

where top-$k_{\mathrm{same}}(c)$and top-$k_{\mathrm{diff}}(c)$represent the most similar charges from the same and different chapters,respectively.Thisdual-sourceexpansion helps the model compare similar alternatives, reducing the risk of overlooking relevant charges.

History-basedExpansion. Certainchargestend to appear together in real-world cases, reflecting practical legal dependencies or common joint indictments.WeleveragetheMultiLJP(Lyuetal., 2023) dataset, where each case may involve multiple defendants and multiple charges. By analyzingthesecases,weconstructaco-occurrencedictionary that records how frequently each pair of charges appears together.For a given charge c, we select the top-k most frequently co-occurring charges as the expansion set $E_2(c)$.

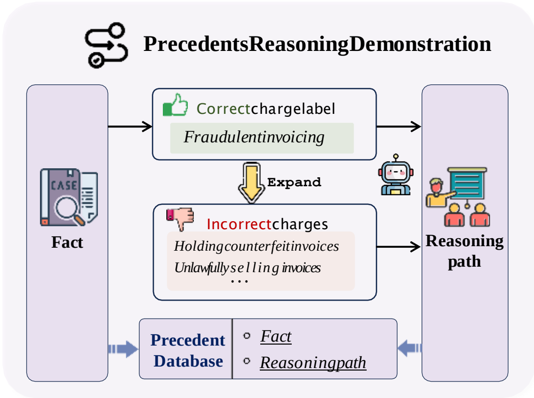

Figure3:ThemoduleofPrecedentsReasoningDemonstration: LLManalyzesthereasonsfortheselectionor exclusionofeachchargebasedonfacts,therebygenerating the reasoning path of precedents.

FinalExpansionSet. Givenaninitialsetofcandidate charges $\{c_1,c_2,\ldots\}$predicted by the languagemodel,weapplythetwostrategiesaboveto expand each charge:

$$E(c_i)=E_1(c_i)\cup E_2(c_i)\quad(2)$$

PrecedentsReasoningDemonstration

Previous precedent-based approaches (Wu et al., 2023;ChenandZhang,2023;Santoshetal.,2024) typically retrieve the fact description and final judgment of prior cases, then insert them directly into the prompt.However, such methods offer little insightintothe*reasoningprocess*behindthosedecisions.As a result, they tend to rely on shallow fact matching rather than learning how to distinguish between legally similar charges.

To address this issue, we construct reasoning-augmentedprecedentsthatmakethedecisionlogic explicit. AsshowninFigure3,wefirstexpandits original charge c into a set of similar charges C.

Giventhecasefact$f$,thecorrectcharge$c$,and thesetofalternatives$C$,wepromptLLMtogenerateareasoningpath$r$ thatexplainswhycisappropriateandwhytheothercandidatesin$C\setminus\{c\}$

shouldbeexcluded[1].Thisreasoningisgenerated **offline**andstoredtogetherwiththecasefacts.

LegalSearch-AugmentedReasoning

While recent retrieval-augmented generation (RAG)approaches(Wuetal.,2023;PengandChen, 2024;Fengetal.,2024)enhancelegalmodelsbyretrieving precedents, statutes, and charge definitions, they remain limited in key aspects.Specifically, they often fail to resolve fine-grained distinctions betweensimilarchargesorprovidedetailedrulesto determinefacts. Moreover,thesemethodsrelyon static retrieval from fixed knowledge bases, making them inflexible and unable to accommodate evolving judicial practices.

Toaddresstheselimitations,weintroducea**dynamicanditerativelegalsearch-augmentedreasoningmechanism**. Ratherthanpassivelyinjecting generic legal content, our method allows the LLM to actively identify knowledge gaps during reasoningandgeneratetargetedqueries.These queriesfocuson*subtledifferencesbetweencandidate charges* and *fact-specific questions*.We exclusivelysourceauthoritativelegalinterpretations from official channels, thereby minimizing noise. The system retrieves relevant legal texts from the webinrealtime,enablingup-to-dateandcontext-related augmentation.

We further ground the model's reasoning in a **syllogistic structure**: the retrieved legal context servesasthemajorpremise,thecasefactastheminorpremise,andtheconclusionisderivedthrough logicalalignment(JiangandYang,2023;Heetal., 2025).This structure helps the model remain groundedinfactualevidenceandreducehallucinations. Theoverallreasoningprocessisformalized as an iterative function:

$$R_t = f_\theta(R_{<t}, q_t, d_t, f), \tag{3}$$

where$R_t$denotesthecurrentreasoningstate, $R_{<t}$arethehistoricalreasoningpaths,$q_t$and$d_t$are thequeryandcorrespondingretrieveddocuments of this step, and $f$is the case fact.

Thisdesignenablesthemodeltoincrementally constructalegallygroundedreasoningchain,adaptively integrating external knowledge as needed. By decoupling retrieval from static knowledge basesandaligningitwiththemodel'sevolving needs, our framework offers greater flexibility to real-world legal dynamics.

[1]Weprovidethedetailedpromptandexamplesforsynthesizing reasoning paths in Appendix C

4 Experiments

DatasetsandEvaluation

We conducted experiments in both single-defendant and multi-defendant scenarios to verify the effectiveness of our method in practical applications.For the single-defendant case, we usetheCAIL2018dataset(Xiaoetal.,2018). For the multi-defendant case, we adopt the CMDL dataset(Huangetal.,2024).Weuniformlysampled across all charges to construct a balanced test set. ThedetailsareshowninTable1.ForthePRDmod- ule, We employ the training set from both dataset asourprecedentdatabase. Forevaluationmetrics, we adopt the same measures used in prior work: Accuracy(Acc.),MacroPrecision(Ma-P),Macro Recall (Ma-R),and Macro F1 (Ma-F).

Dataset	CAIL2018	CMDL
#Traincases	100,531	63,032
#Testcases	1,000	834
# Charges	190	164
# Articles	175	147
#Averagecriminalpercase	1	3.79
Averagelengthpercase	409.6	1124.94

Table1:Statisticsofdataset.

Baselines

Wecompareourmethodagainsttwocategoriesof baseline approaches:

ClassificationMethods: Thesemethodstakelegaljudgmentpredictionasaclassificationtask,relyingonsupervisedlearningwithlabeleddatasets. **TopJudge** (Zhong et al., 2018) employs a graph structure to model the topological dependency among the three subtasks: charge prediction, law article prediction, and sentence term prediction. **NeurJudge** (Yue et al., 2021) integrates a legal knowledgegraphintotheneuralarchitecture,capturing explicit relationships among legal entities and improving reasoning over structured legal knowledge.**BERT** (Devlin et al., 2019), a standard pre-trained transformer model, is adapted to legaltextsviasupervisedtraining.Itservesas a strong baseline for judgment prediction tasks. **Lawformer**(Xiaoetal.,2021)isbuiltuponLongformer(Beltagyetal.,2020)andfurtherpretrained

Methods	Charge				LawArticle			
	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F
ClassificationMethods								
TopJudge	52.1	50.9	45.7	43.5	52.8	47.7	43.8	41.2
LADAN	76.7	73.4	71.0	69.5	77.5	71.0	69.2	67.5
NeurJudge	74.7	77.7	71.5	71.5	77.4	80.7	74.6	74.3
BERT	85.8	83.4	86.6	83.3	85.8	80.4	82.8	79.9
Lawformer	71.3	58.2	62.7	57.8	72.9	58.1	61.4	56.9
DirectReasoning								
Qwen2.5-32B	74.5	75.3	69.3	69.1	77.1	73.3	66.6	67.1
QwQ-32B	82.5	86.9	80.5	80.9	84.0	83.1	76.1	77.0
Qwen2.5-72B	76.6	78.9	72.2	72.3	77.7	73.4	66.8	67.3
DeepSeek-R1-671B	84.8	86.3	81.3	81.7	87.2	86.8	81.8	82.6
Retrieval-augmentedReasoning								
Precedent-based-RAG-Qwen2.5-32B	88.5	88.2	85.8	85.7	89.4	87.2	83.7	84.5
Precedent-based-RAG-QwQ-32B	89.4	89.9	87.3	87.1	90.4	88.4	85.2	85.4
Precedent-based-RAG-Qwen2.5-72B	88.1	87.5	85.1	84.9	89.4	86.8	83.9	84.0
Search-o1-QwQ-32B	81.8	85.3	78.8	79.3	83.9	83.3	76.4	77.4
AgenticRetrieval-augmentedReasoning								
GLARE-Qwen2.5-32B(ours)	**89.8**	89.8	87.8	87.8	90.4	89.2	87.3	87.5
GLARE-QwQ-32B(ours)	89.7	**90.7**	**88.6**	**88.6**	**91.3**	**90.6**	**88.3**	**88.5**

Table2:PerformancecomparisononCAIL2018dataset.Thebestresultsareinbold.

on large-scale Chinese legal corpora, which enhancesitsabilitytoprocesslongerlegaldocuments and capture complex contextual semantics.

LLM-based Methods: These methods utilize LLMs to perform legal reasoning in zero-shot or few-shot settings (Brown et al., 2020).**Direct Reasoning** directly feeds the case facts into the LLM to predict the applicable law articles and charges,withoutrelyingonanyretrievalaugmentation or additional external context.The modelsusedinthissettingincludeQwen2.5-32B/72B-Instruct (Yang et al., 2024a), QwQ-32B (Team, 2025),andDeepSeek-R1-671B(Guoetal.,2025). **Retrieval-augmentedReasoning:***(1)Precedent-basedRAG*enhancesreasoningbyretrievingtop-5 precedents including their facts and labels, which are appended to the prompt.The models used in this setting include Qwen2.5-32B/72B-Instruct (Yang et al., 2024a), QwQ-32B (Team, 2025). *(2)Search-o1*(Lietal.,2025)dynamically retrieves external knowledge when it encounters uncertain or ambiguous knowledge in the general domain.WeusereasoningmodelQwQ-32B(Team, 2025) in this setting.

ExperimentSettings

Inourexperiments,weadoptQwen2.5-32B(Yanget al., 2024a) and QwQ-32B (Team, 2025) as the basemodelstorunthefullreasoningpipeline. For generation,we set thefollowing parameters:a maximum of 32,768 tokens and temperature of 0.6. For charge expansion, we set the top-k expanded chargesto3ineachexpansionmethod. Forprecedent retrieval, we use SAILER (Li et al., 2023) to encodecasefactsandsetthetop-$k$retrievedprecedents to 5.In the legal search module, we utilize SerperAPI[2]withtheregionconfiguredforChina and the number of returned results limited to the top 10. For charges that are not in the predefined label set, we map them to the most similar charge withinthelabelsetusingBGE(Xiaoetal.,2024).

ExperimentResults

TheresultsarereportedinTable2andAppendixD, andnextwewillanalyzetheexperimentalresults:

1. Our method has demonstrated consistent performanceimprovementsinbothchargeprediction and law article prediction tasks, highlightingtheeffectivenessofouragenticreasoning approachforLJP.ComparedtotheDirectReasoningsetting,ourmethodimproveschargeprediction by7.7%andlawarticlepredictionby11.5%inF1 score. WhencomparedwithRetrieval-augmented Reasoning,itachievesanimprovementof1.5%on charge prediction and 3.1% on law article prediction in F1 score. In addition, our method not only performs well on the large reasoning models, but also effectively promotes the reasoning ability of theinstructmodels,indicatingthatourthreemod-

[2]https://serper.dev

Methods	CAIL2018				CMDL			
	Charge		Law Article		Charge		Law Article	
	Acc.	Ma-F	Acc.	Ma-F	Acc.	Ma-F	Acc.	Ma-F
DirectReasoning								
Qwen2.5-32B	60.2	39.3	63.7	41.2	57.4	64.7	57.9	63.7
QwQ-32B	78.4	57.0	79.2	58.2	67.9	72.9	69.8	74.2
Retrieval-augmentedReasoning								
Precedent-based-RAG-Qwen2.5-32B	82.6	62.7	83.0	62.3	65.7	69.5	65.3	67.6
Precedent-based-RAG-QwQ-32B	84.6	65.5	84.6	67.6	72.8	74.8	71.3	73.2
AgenticRetrieval-augmentedReasoning								
GLARE-Qwen2.5-32B(ours)	86.9	68.6	86.5	68.3	73.5	75.5	71.9	73.4
GLARE-QwQ-32B(ours)	**90.7**	**75.7**	**91.1**	**75.4**	**76.0**	**79.5**	**74.0**	**76.7**

Table3:Performancecomparisonondifficultcharges.

ules effectively supplement legal knowledge and thereby enhance reasoning performance.

2. Incontrasttodirectreasoning,precedent-based RAG enhances prediction performance through precedent retrieval. Large reasoning models like QwQ-32B and DeepSeek-R1-671B outperformotherinstructmodelsindirectreason-ing, indicating that LJP inherently requires multi-stepreasoningandslowthinking.Precedent-based RAGimprovesperformanceacrossmodelsofvari-ousscalesbyincorporatingprecedentretrieval.For example, QwQ-32B sees an 8.39% F1 improve-mentinlawarticleprediction.However,precedent-basedRAGonlyprovidesthecasefactsandlabels of precedents, leading models to rely on similar-itymatchingandcopyjudgmentpredictionsrather thantrulyreason.Additionally,Search-o1retrieves casefactswhichmayintroducenoises,ratherthan specific legal knowledge, thus underperforming compared to direct reasoning.

3. LLM-based methods outperform classification methods. However, BERT demonstrates superior performance compared to LLMs such as Qwen2.5-72B-Instructinthedirectreasoningset-ting, although the latter have significantly larger parametersizes.Thekeyreasonsareasfollows:
(1) BERT frames charge and article predictionas multi-class classification tasks, enabling direct mapping from facts to fixed labels, which aligns well with the task.In contrast, LLMs take a generative approach and without legal-specific train-ingtheyoftenfailtomakeaccuratepredictions.
(2) Our dataset includes many rare and confusing charges.Fine-tuned BERT models trained on legal corpora can better distinguish these nuanced charges, while even large LLMs lack the domain knowledge needed for such difficult charges.

AblationStudy

Methods	Charge		LawArticle	
	Acc.	Ma-P	Acc.	Ma-F
w/oCEM	89.6	87.7	90.3	85.2
w/oPRD	80.0	78.1	81.6	75.4
w/o LSAR	89.6	87.9	90.4	86.5
GLARE(ours)	**89.7**	**88.6**	**91.3**	**88.5**

Table4:AblationStudy.Thebestresultsareinbold.

To evaluate the effectiveness of each compo-nentintheGLAREframework,weconductedabla-tionexperimentswiththefollowingstrategies: **(1) w/o CEM:** The Charge Expansion Module is re-moved,sothemodelcannotexpandadiversesetof candidate charges. **(2) w/o PRD:** The Precedents Reasoning Demonstration module is removed, so the model cannot refer to reasoning path from precedents.**(3) w/o LSAR:** The Legal Search-Augmented Reasoning module is removed, dis-ablingthemodel'sabilitytosupplementitsknowl-edge via external legal search when faced with ambiguous or unfamiliar charges.

As shown in Table 4, the removal of any single module results in degraded performance.In particular,removingPRDcausesthemostsignifi-cantdegradation:theaccuracyofchargeprediction drops from 89.7% to 80%.This highlights the crucialroleofprecedentreasoningpathinenhanc- ing legal judgment prediction.Removing CEM weakensthemodel'sabilitytorecognizeambigu- ous or low-frequency charges, while LSAR helps the model fill knowledge gaps by retrieving au-thoritativelegalinformation. Overall,theGLARE framework performs best across all metrics, val-idating the strength of agentic reasoning in legal judgment prediction.

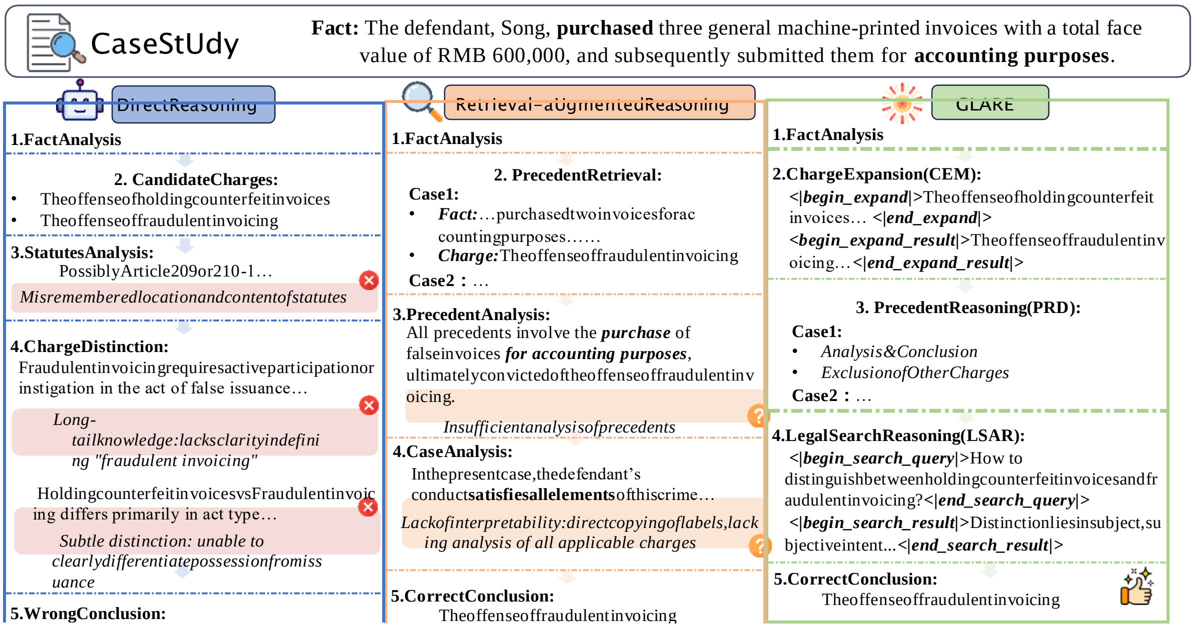

Figure4: CaseStudy. Theredparthighlightsthemodel'slimitationsduetoinsufficientinternalknowledge,while the yellow part demonstrates the lack of interpretability in vanilla precedent-based RAG reasoning.

EfficiencyAnalysis

Inthiswork,wefocusonmulti-stepreasoningand slowthinkingforlegaljudgmentprediction,sothe latency is less important.Nevertheless, we still conducted an analysis to further understand each module. BasedontheanalysisofFigure5(a),we can draw the following conclusions:

(1) The overall inference efficiency is relativelyhigh. Theaveragereasoningroundsforeach caseis5.17andtheaveragecallnumbersforeach module is between 1.7 and 1.8 times, indicating thatthemoduleschedulingiswell-balanced,withoutobviousredundancyorrepeatedinvocation. So the overall delay is within our acceptable range.

(2) TheCEMmoduleisthemostefficient. In bothexpansionmethods,thechargestructuresare established offline in advance, so its computation costislowandruntimeisminimal.Asshown in the figure 5(b), compared with direct reasoningandRAGapproaches,ourmethodconsidersa comprehensivesetofchargesandperformsamore thorough analysis.

(3) The PRD module has the highest latency butwithinanacceptablerange. Sincethismoduleneedstoencodetheentirecasedescriptionand the case text is usually long,the reasoning timeis relatively long. However, the PPR module can provide the reasoning path of precedents and has significant reasoning interpretability.

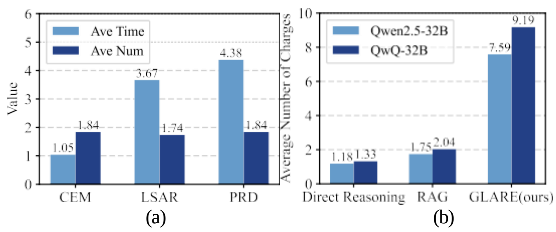

Figure 5: (a) Efficiency analysis of each module.(b) Average charge numbers of different methods.

CaseStudy

AsshownintheFigure4,weconductedcasestudy onthreeLLM-basedmethodstofurtherverifythe effectivenessandinterpretabilityofourmethod.DirectReasoningreliesontheLLM'sinternalknowledge,whichmaybeinaccurateorinsufficient,leading to incorrect judgments.RAG methods often lack explicit links between retrieved cases and finaldecisions,makingithardtotracehowexternal knowledgeaffectsreasoning.However,ourmethod ensuresthateachreasoningstephasaclearknowledgebasisthroughtheexplicitinvocationofthree modules, thereby extending the reasoning chain.

PerformanceonDifficult Charges

ToevaluateGLARE'sabilitytohandlechallenging charges requiring long-tail knowledge, We conductedexperimentsonlow-frequencychargeswith lessthan100cases(e.g.,thecrimeofbribingaunit) and confusing charges (e.g., robbery vs.snatching).The results are reported in Table 3, which reveal two key insights: (1) Our method dynamically acquires critical legal knowledge, outperforming Direct Reasoning by over 10% and Retrieval-augmentedReasoningbyover5%. (2)RAG-based methodsstruggletoretrieverelevantprecedentsfor such charges, leading to poor performance, while direct reasoning fall short due to limited long-tail knowledge. Theseresultshighlightthestrengthof our external modules in supplementing legal reasoning with critical knowledge.

5 Conclusion

In this study, we propose a novel framework, GLARE, to address the legal gaps in legal reasoning.GLARE dynamically acquires key legal knowledge to improve the breadth and depth of reasoning.Experimental results demonstrate the effectivenessofourapproach,whichnotonlyimproves prediction performance but also generates complete reasoning chains that enhance the interpretability of LJP tasks. We believe that GLARE holds great potential for real-world legal applications and will contribute meaningfully to the advancement of intelligent judicial systems.

Limitations

GeneralizabilityWe adopted the legal dataset fromChinaJudgmentsOnlinetoverifytheapplicabilityofthemethodintheChina'sjudicialsystem. However,theGLAREframeworkisapplicableto countries following both common law and civil law systems.When applied to the actual judicial practiceofaspecificcountry,weneedtoinjectthe specificlegalknowledgebaseofeachcountryand adapt to the local judicial culture.

EfficiencyOur method promotes the reasoning ability of the model through multiple rounds of reasoningandtheinvocationofthreemodules.Althoughthisprocesshasanincreasedtimecostcompared to the traditional direct reasoning method, thetaskoflegaljudgmentpredictionitselfisatask thatrequiresmulti-stepreasoningandslowthinking. Moreover,thistimecostismuchlessthanthe time needed for humans to analyze cases in real life. Therefore, such a time cost is acceptable.

EthicalDiscussion

Potential Bias in Legal Data Large language modelsmaylearnhistoricalbiasfromlegaljudgments in training data.In practice, judicial decisionsmaybeinfluencedbymanyexternalfactors, suchaspublicopinion,regionaldifferencesorthe personalinclinationsofjudges.Weneedtoidentify possible biases before deploying such models in real-world scenarios.

Human-CentricDeployment Oursystemisdesigned to assist judges by providing supplementaryrecommendationsratherthanreplacinghuman decision-making.Weadviseuserstocriticallyevaluatethemodel'spredictionsandmakeindependent decisionsabouttheiradoption,ratherthanuncritically accepting the model's reasoning.

References:

1. Ashley, K. D. (2017). Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age . Cambridge University Press. .

2. Susskind, R. (2019). Online Courts and the Future of Justice . Oxford University Press.

3. Zeng, J., & Wang, T. (2023). A Survey of Legal Judgment Prediction: Datasets, Metrics, Models, and Challenges. AI Open , 4, 1-12.
Why: A recent survey that provides a comprehensive overview of the LJP field, its state-of-the-art, and open problems.

4. Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law , 28(2), 237-266.

5. 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).
6. Chalkidis, I., Kampas, D., & Androutsopoulos, I. (2019). Neural legal judgment prediction in English. In
7. 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).

8. Katz, D. M., Bommarito II, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the supreme court of the united states. PloS one , 12(4), e0174698.

9. Savelka, J., Ashley, K. D., Gray, M. A., & Walker, V. R. (2023). Explaining Legal Concepts with Augmented Large Language Models (GPT-4). arXiv preprint arXiv:2306.09525 .

10. Wooldridge, M. (2009). An Introduction to MultiAgent Systems . John Wiley & Sons.
11. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... & Wen, J. (2023). A Survey on Large Language Model based Autonomous Agents. arXiv preprint arXiv:2308.11432 .
Why: An excellent and recent survey on the very hot topic of using LLMs as the "brains" of autonomous agents, directly relevant to your approach.

12. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, K., ... & Tang, J. (2023). The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv preprint arXiv:2309.07864 .

13. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems , 35, 24824-24837.
14. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. arXiv preprint arXiv:2210.03629 .

15. Long, J. (2023). Large Language Model Guided Tree-of-Thought. arXiv preprint arXiv:2305.08291 .

16. Surden, H. (2019). Artificial Intelligence and Law: An Overview. Georgia State University Law Review , 35(4).

17. Citron, D. K., & Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. Washington Law Review , 89, 1.
Why: A classic legal article discussing the due process requirements for automated decision-making systems, crucial for any work on "judgment prediction."

18. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine , 38(3), 50-57.

19. Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society , 7(1).
20. Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. Information, Communication & Society , 20(1), 118-136.