



# 司法大语言模型探索

艾清遥

清华大学计算机系、清华大学互联网司法研究院

yiqunliu@tsinghua.edu.cn

# 大 纲

## Outline

- 研究背景
- 大语言模型与司法信息处理
  - 前大模型时代的司法语言模型
  - 大模型时代中的司法语言模型
  - 司法语言模型评价体系





# 研究背景

## 司法信息化产生海量数据导致信息获取困难

- 信息过载导致有效法律信息获取更加困难
- 中国：全世界最大司法信息资源库，汇聚超**2.8亿**件案件信息
- Westlaw：**3.2万个数据库**，涵盖判例法、成文法、文献、新闻
- LexisNexis：**6万个数据库**，**超1.8亿个**联邦和州法院文件
- 传统信息处理技术无法满足专业用户需求
- **规模大**：精准获取和全面召回困难
- **需求繁**：降低用户显式表达负担
- **任务重**：需要全流程的业务辅助





# 研究背景

## 大语言模型超大的参数规模带来前所未有的能力

- 符号主义与连接主义相融合是第三代人工智能的核心
- 大数据与富知识双轮驱动，建立鲁棒可解释人工智能理论与方法
- 研究安全、可信、可靠和可扩展的人工智能技术，推动创新应用
- 司法信息处理关键的研究问题和重要的应用场景
- 司法信息智能获取具有高质量**数据驱动**与高精度**知识指导**结合的特征
- 司法领域包含海量规整文书数据，拥有大量深度专业知识的专业人员

Intended Customers		
Law practices	Corporate legal	Individuals
Contract management Contract drafting, editing, contract analytics	Kira Spellbook	Ironclad LINSQUARES EVOLV SPOTDRAFT Robin ZUVA Latch MAIGON
Legal Documentation Drafting patents, legal communications etc., doc search, document Q&A	Luminance LEGALMATION PatentPal gavel	Typically outsourced to external legal counsel Scarce need
Legal assistant Legal chatbots for answering queries, executing specific tasks	Harvey LAWOROID	Typically outsourced to external legal counsel Scarce need
Legal research Semantic caselaw search, precedent search, research memos	casetext	Typically outsourced to external legal counsel Scarce need

ChatLaw      JII 百川智能 BAICHUAN AI  
Lawyer LLaMA      LexiLaw

美国法律AI知名初创企业

目前发布的中文法律大模型



工商信息

千万级商标、专利



法律咨询

数千万法律问答



案例文书

1.3亿篇



法律法规

1,275部全国性法律



司法解释

1,411部司法解释



法学文献

几百种法学刊物





# 语言模型与信息处理

TF-IDF

...

Boolean Model

## 向量空间模型



Word2vec

Paragraph2vec

Glove

## 词嵌入模型



OpenAI

1960

1990

2010

2020



## 统计语言模型

Markov Random Field

Probabilistic Language Modeling

...



## 大规模预训练 语言模型

BERT

GPT

ELMo

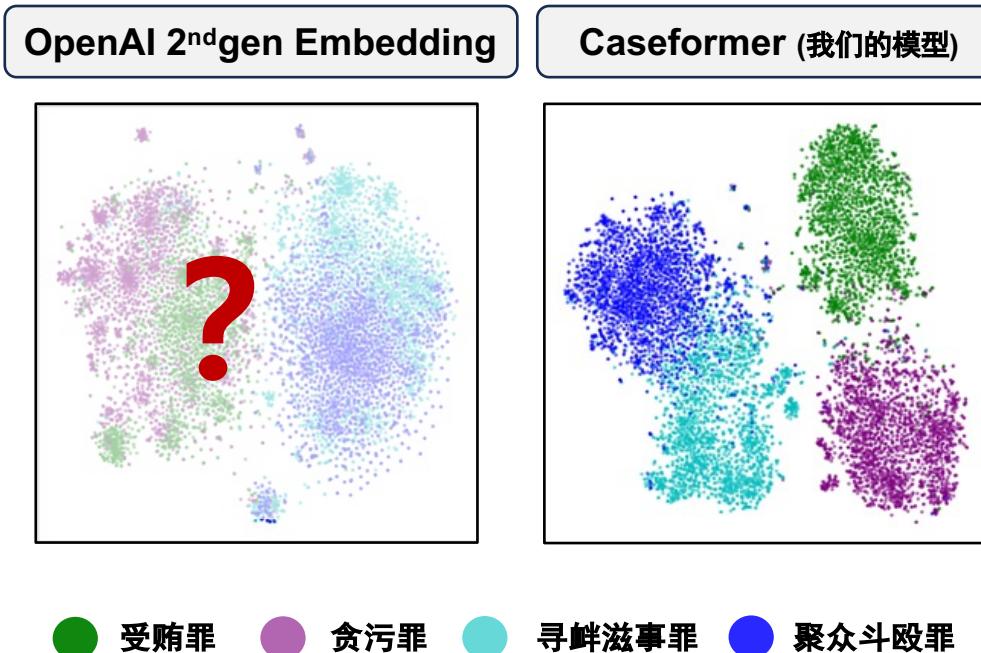
...





# 研究背景

- 已有通用大语言模型无法满足司法信息处理需求



[Su et al. 2023, Liu et al. 2023]



GPT-4

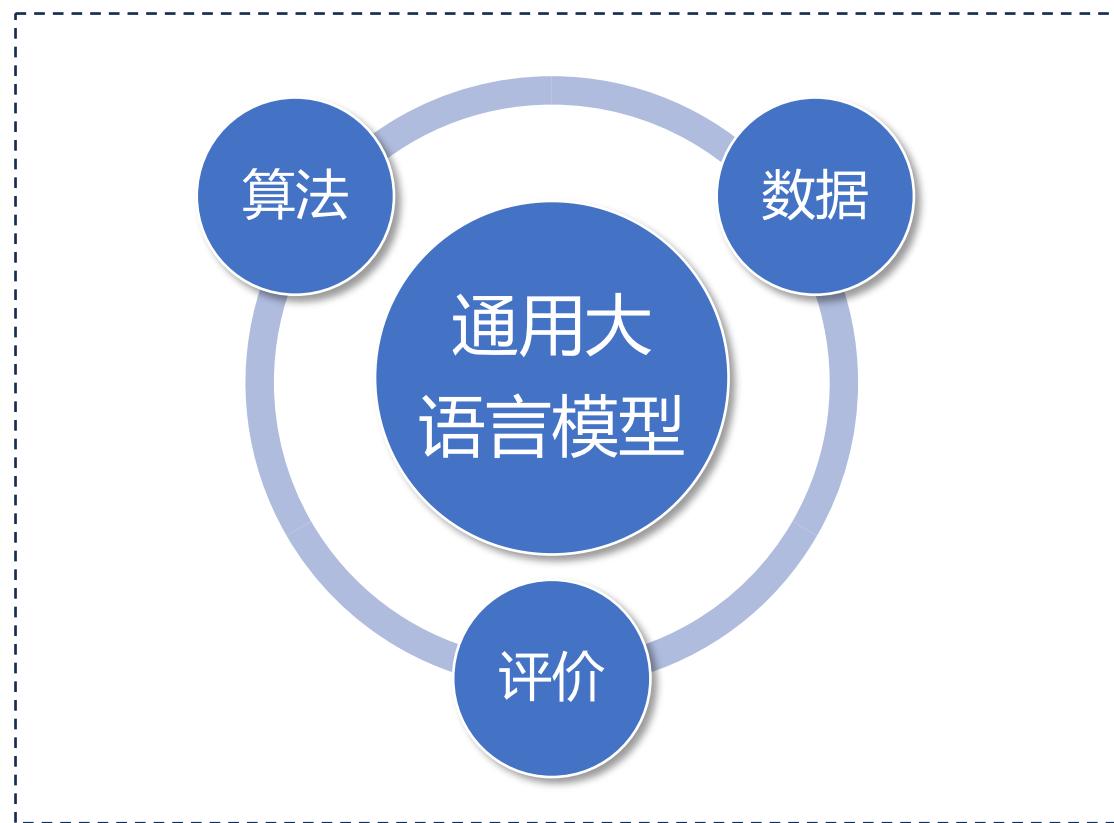
人：郭某涉嫌毒品运输，请帮我找相关案例  
GPT-4：请问，被告人郭某向他人贩毒的具体时间、地点、数量？  
人：郭某向他人贩毒的具体时间…惠东县平山…数量包括…  
GPT-4：请问被告人郭某在被抓捕前是否曾经向其他人贩毒，如果有，具体时间、地点和贩卖的毒品种类是什么？ **重复**  
人：是的，郭某在被抓捕前…  
GPT-4：被告人郭某是否有前科记录或者之前有过类似的毒品贩卖行为？  
人：不知道  
GPT-4：请问被告人郭某在贩卖毒品时是否存在使用暴力或者威胁他人的行为？ **相关，但重要性较低**  
人：不太清楚  
GPT-4：被告人郭某是否有前科记录或者之前有过类似的毒品贩卖行为？ **重复**  
人：不清楚。



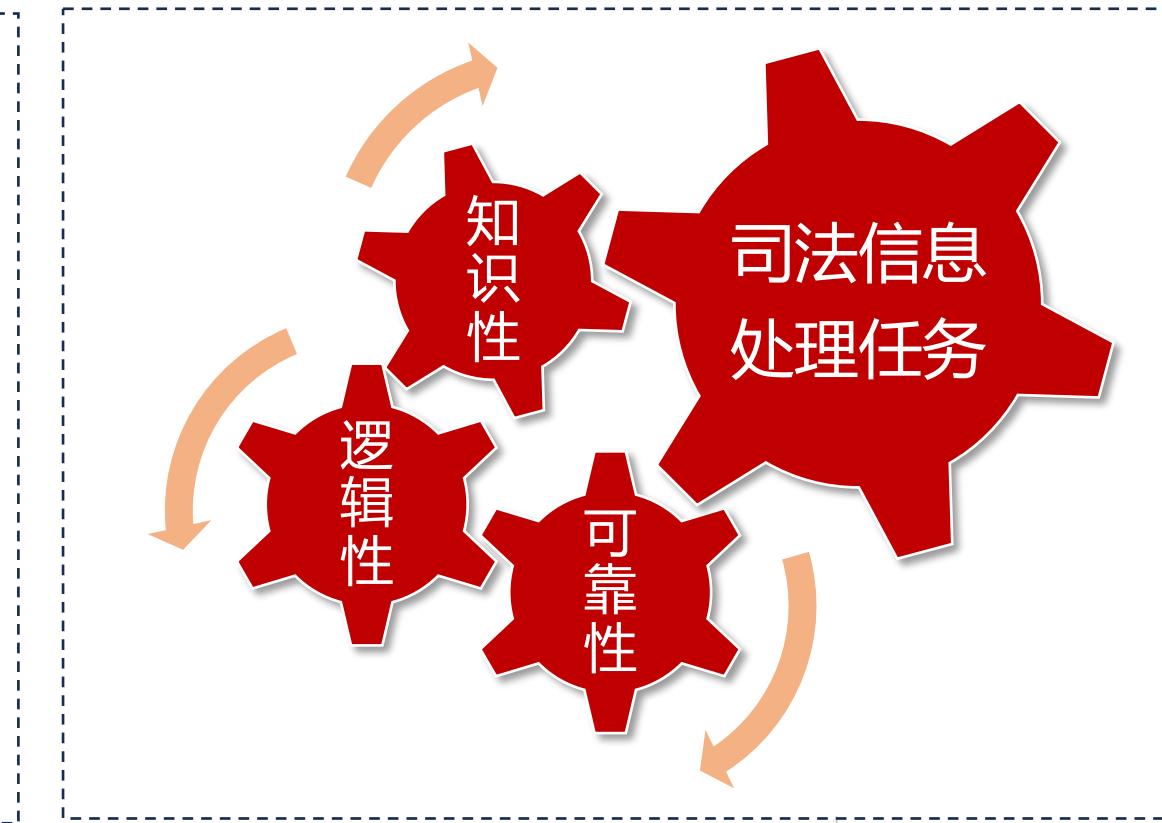


# 通用语言模型与司法信息处理

## 通用技术



## 司法应用





# 通用语言模型与司法信息处理

- **通用语言模型无法有效驱动司法信息**

## 智能处理

- 缺少司法专业知识
  - 法律要素、司法条款...
- 难以理解司法逻辑
  - 案卷结构、量刑程序...
- 缺乏行为可靠性保障
  - 行为难解释、输出存在幻觉...
- 专业评价体系不完善
  - 评价方式远离业务，模型优化缺少指导



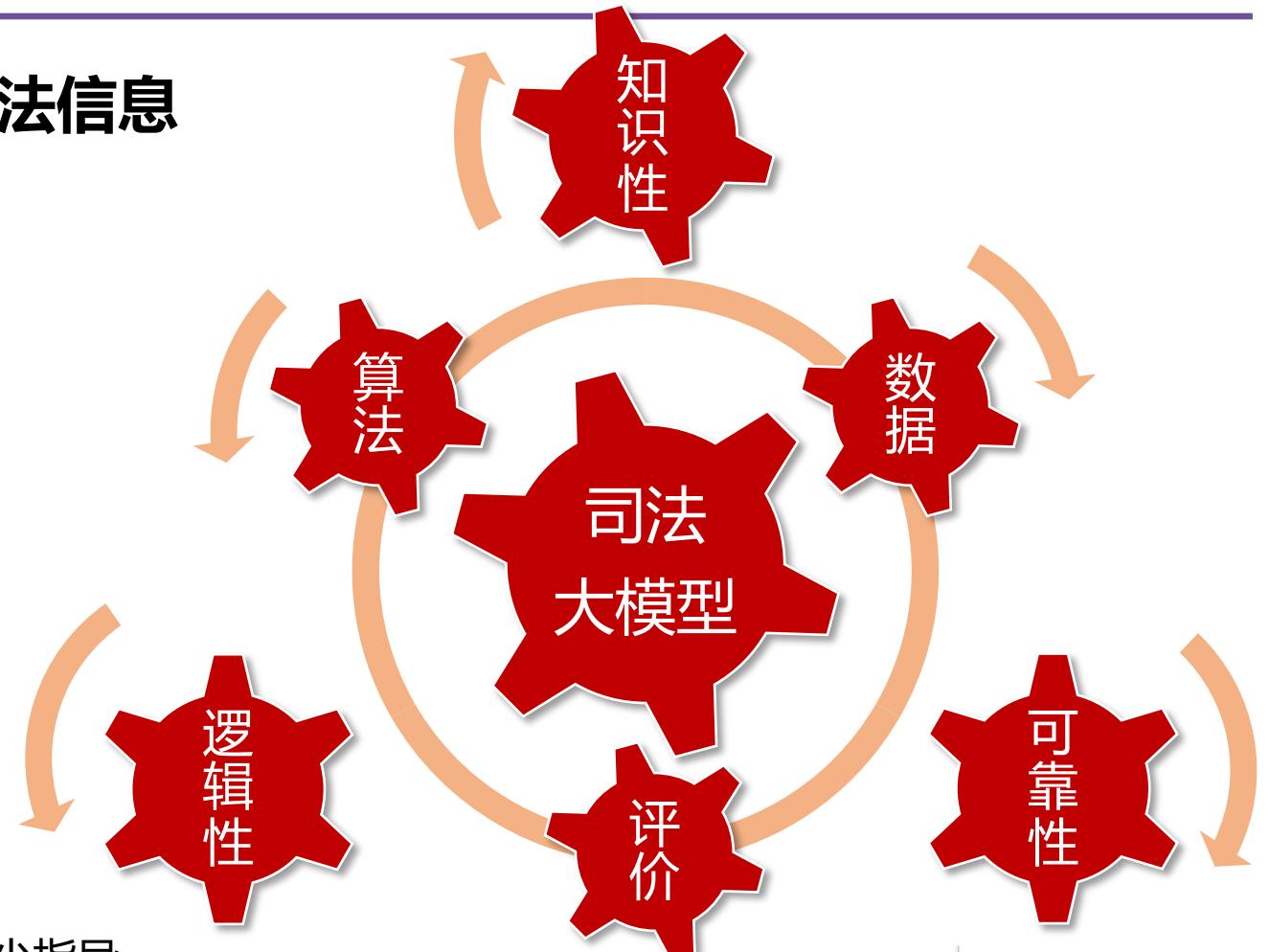


# 通用语言模型与司法信息处理

- 通用语言模型无法有效驱动司法信息

## 智能处理

- 缺少司法专业知识
  - 法律要素、司法条款...
- 难以理解司法逻辑
  - 案卷结构、量刑程序...
- 缺乏行为可靠性保障
  - 行为难解释、输出存在幻觉...
- 专业评价体系不完善
  - 评价方式远离业务，模型优化缺少指导



迫切需要适配司法专业的大模型!





# 前大模型时代的司法语言模型

- 前大模型时代的语言模型
  - 参数规模小，训练成本低
  - 结构简单，行为易于理解
  - 能力有限，难以理解专业知识及逻辑
- 司法语言模型关键挑战
  - 如何理解司法文本内在逻辑
    - 基于文书结构的语言预训练任务
  - 如何捕捉司法知识及其要素
    - 基于司法要素的对比学习算法
- **如何构建适合司法的预训练算法？**

[Ma et al. 2022, Liu et al. 2023, Su et al. 2023, Ma et al. 2023]



# 司法 预 训 练



## 司法文书内结构关系建模

- Li et al. *SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval* (SIGIR 2023)

## 司法文书间结构关系建模

- Su et al. *Caseformer: Pre-training for Legal Case Encoding* (TOIS 2024)

## 关键法律/事实要素自监督学习

- Li et al. *Delta: Pre-train a discriminative encoder for legal case retrieval via structural word alignment.* (AAAI 2025)





# 司法文书内结构关系建模

当模型规模一定时：

- **通用预训练模型**在司法类案检索任务效果一般
- **通用预训练任务**在司法文本上训练效果性能一般

CAIL2022司法检索任务效果

Model	F1	NDCG@10
BM25	0.3259	0.7303
QL	0.3246	0.7535
Roberta	0.2893	0.6802
Lawformer	0.2465	0.5436

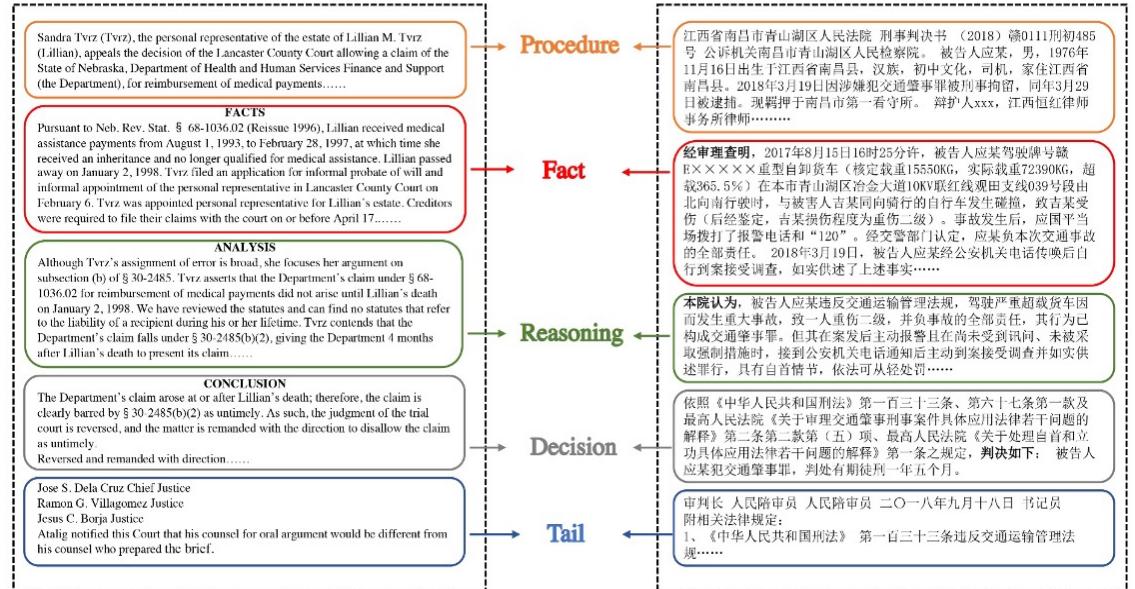
NDCG@n: 评测排名前n的结果质量





# 司法文书内结构关系建模

挑战1：案件文书冗长，且存在特殊的内在结构逻辑



挑战2：司法领域中的文档（案件）相关性与通用领域中不同。

**Paragraphs A:** Person X, 24 years old, men, height 180 cm. In 2018, he entered the shopping mall five times where he **stole** one cell phone and two tablet computers, worth a total of 10,000,000. At five o'clock, he returned home and gave the above items to Person Y.

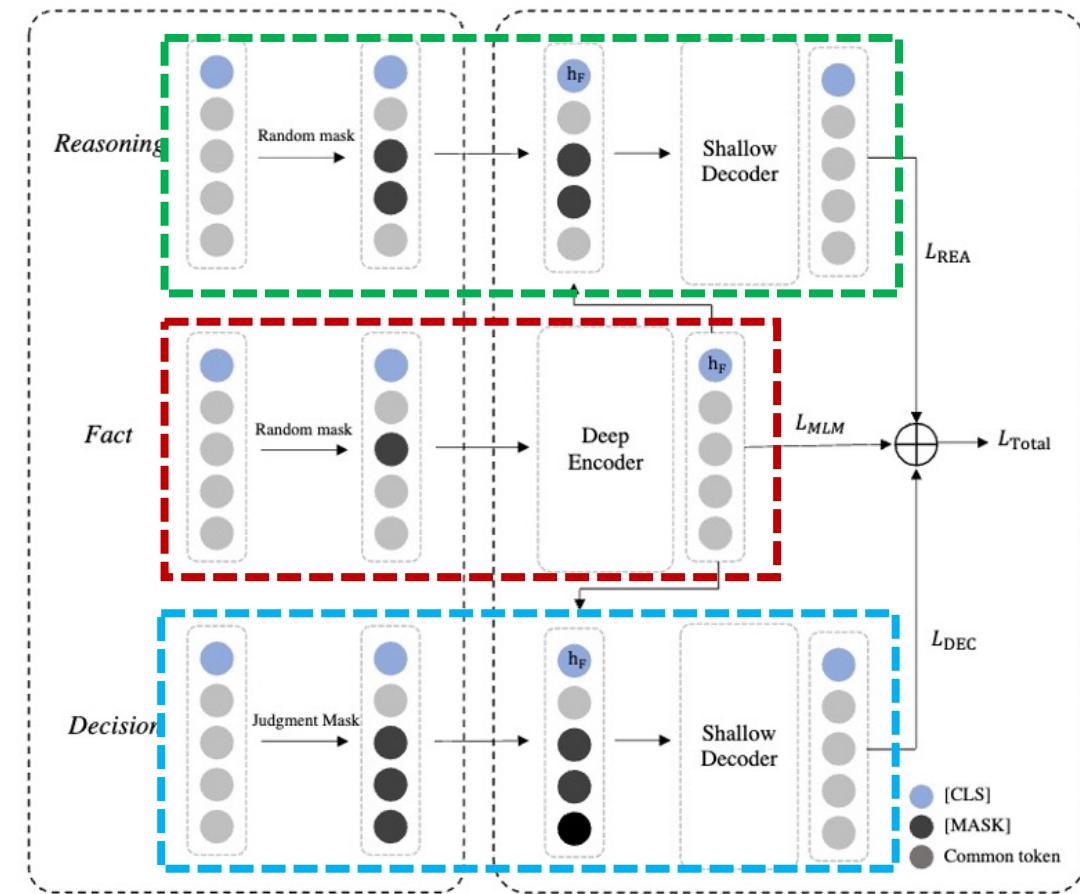
**Paragraphs B:** Person X, 24 years old, men, height 180 cm. In 2018, he entered the shopping mall five times where he **purchased** one cell phone and two tablet computers, worth a total of 10,000,000. At five o'clock, he returned home and gave the above items to Person Y.





# SAILER: 基于文书内在逻辑的预训练

- 建模文书生成，设计训练任务
  - 依据诉讼材料提取事实
    - 从诉讼材料编码案件事实
    - **Fact Encoder**
  - 依据事实撰写裁判分析
    - 基于事实编码结果解码司法分析
    - **Reasoning Decoder**
  - 依据分析进行定罪量刑
    - 基于事实编码结果解码判决结果
    - **Decision Decoder**





# SAILER: 基于文书内在逻辑的预训练

- 相比已有类似规模的通用、司法预训练模型，性能获得显著提高

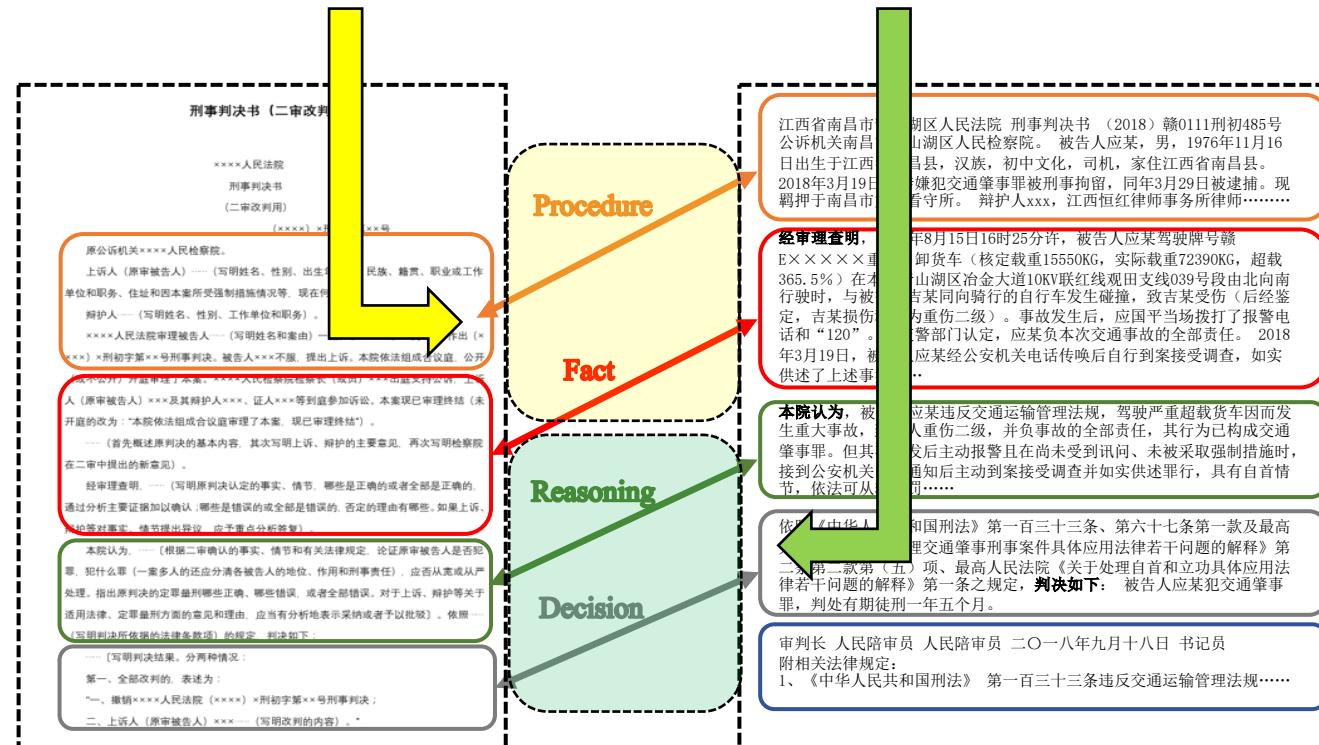
Model	LeCaRD					CAIL2022-LCR				
	Precision	Recall	F1_score	NDCG@10	NDCG@30	Precision	Recall	F1_score	NDCG@10	NDCG@30
<b>Traditional Retrieval Models</b>										
BM25	0.8916	0.1748	0.2922	0.7115**	0.8172*	0.8477	0.2018	0.3259	0.7303**	0.8304
LMIR	0.8897	0.1737	0.2906	0.7157**	0.8373	<b>0.8538</b>	0.2004	0.3246	0.7535**	<b>0.8545</b>
<b>Generic Pre-trained Models</b>										
Chinese Bert	0.6654**	0.1263**	0.2123**	0.5252**	0.5374**	0.5015**	0.1051**	0.1738**	0.4230**	0.3969**
Chinese Roberta	0.8841	0.1778	0.2960	0.7438**	0.7897**	0.7908**	0.1770**	0.2893**	0.6802**	0.6769**
Bert_xs	0.7159**	0.1377**	0.2309**	0.5695**	0.5751**	0.5231**	0.1101**	0.1819**	0.4241**	0.4222**
Lawformer	0.8056**	0.1552**	0.2603**	0.6216**	0.6362**	0.6708**	0.1510**	0.2465**	0.5436**	0.5308**
<b>Retrieval-oriented Pre-trained Models</b>										
Condenser	0.8280**	0.1632**	0.2727**	0.6469**	0.7125**	0.7231**	0.1653**	0.2690**	0.6160**	0.6451**
coCondenser	0.8411**	0.1648**	0.2756**	0.6719**	0.7404**	0.7477**	0.1749**	0.2834**	0.6451**	0.6647**
SEED	0.8411**	0.1575**	0.2653**	0.6721**	0.7330**	0.7185**	0.1547**	0.2546**	0.6194**	0.6168**
COT-MAE	0.8467**	0.1567**	0.2644**	0.6815**	0.7089**	0.7031**	0.1486**	0.2453**	0.6066**	0.6091**
RetroMAE	0.8505**	0.1675**	0.2799**	0.6876**	0.7326**	0.7400**	0.1601**	0.2633**	0.6419**	0.5982**
<b>SAILER</b>	<b>0.9028</b>	<b>0.1902</b>	<b>0.3142</b>	<b>0.7979</b>	<b>0.8485</b>	0.8523	<b>0.2207</b>	<b>0.3506</b>	<b>0.7998</b>	0.8383





# 司法文书间结构关系建模

- “法院是负责解决纠纷的国家审判机关，司法解决纠纷的过程本质上就是寻找事实、寻找法律的过程。”



## 案件之间相似性

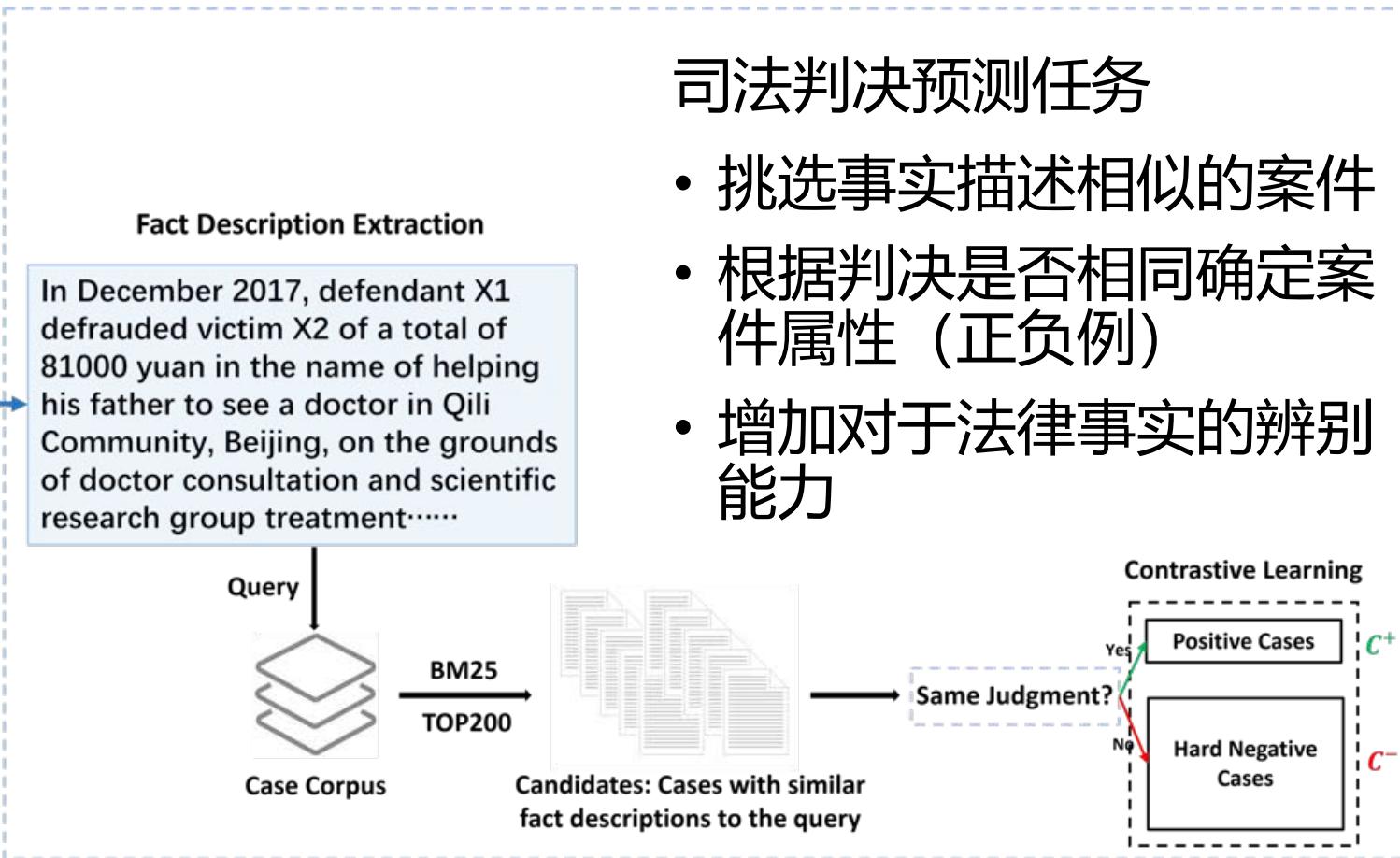
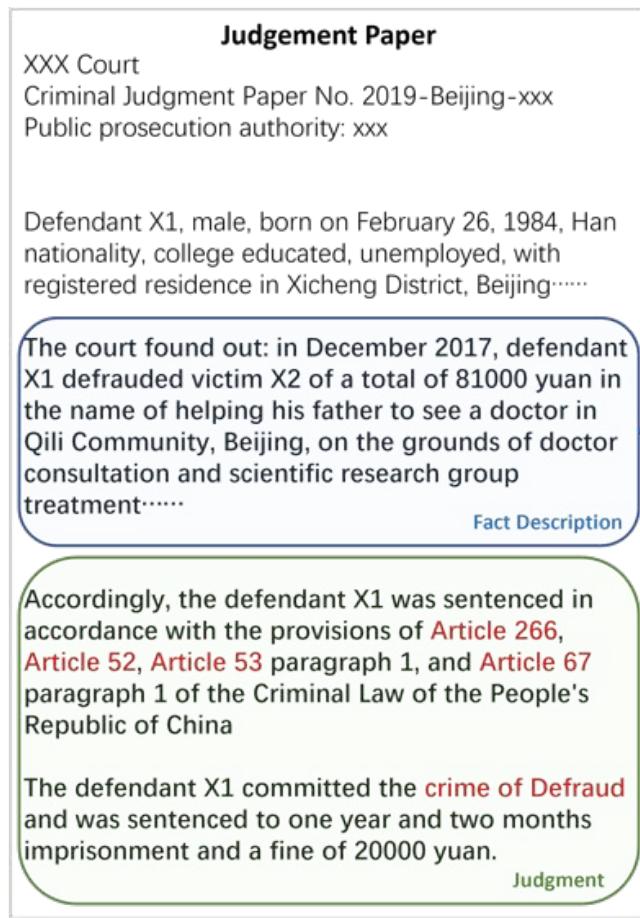
- 事实层面相似性
- 法律层面相似性





# CaseFormer: 司法文书对比预训练

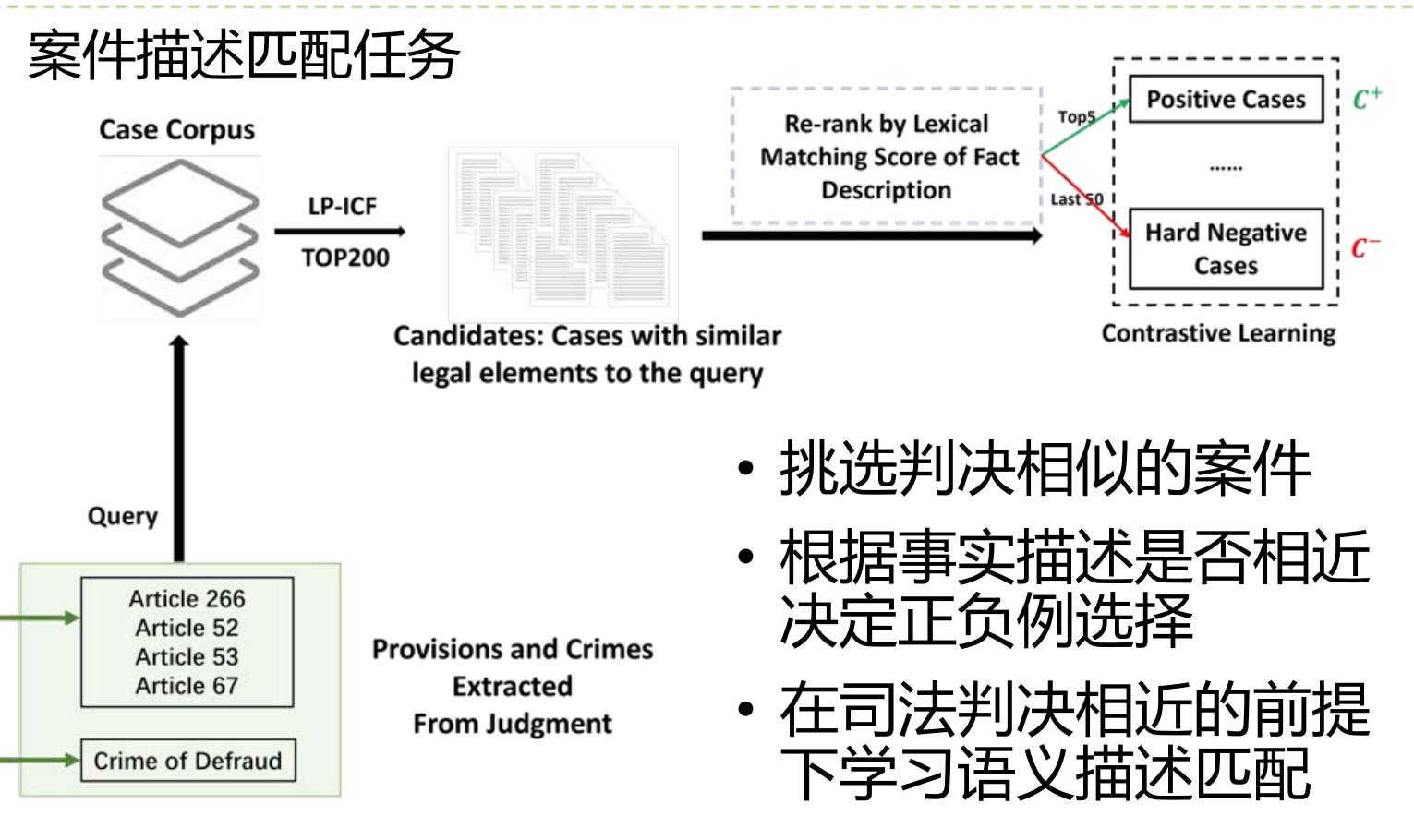
- 基于事实与法条的对比学习





# CaseFormer: 司法文书对比预训练

- 基于事实与法条的对比学习



- 挑选判决相似的案件
- 根据事实描述是否相近决定正负例选择
- 在司法判决相近的前提下学习语义描述匹配

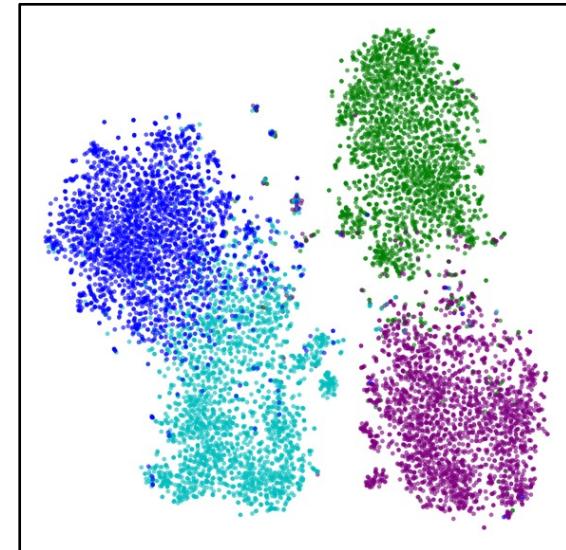


# CaseFormer: 司法文书对比预训练

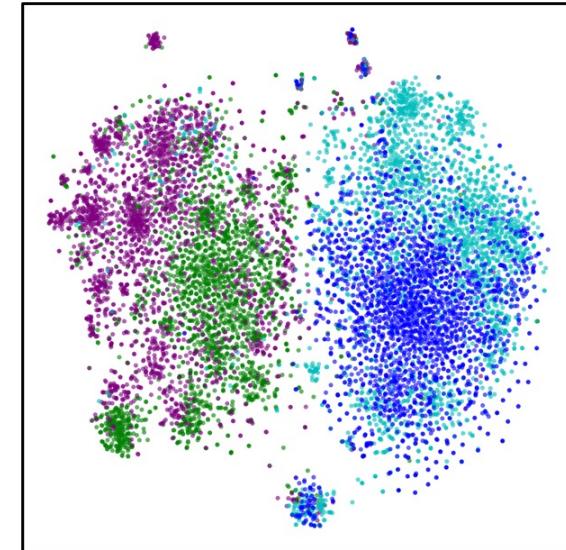
CAIL2022司法检索任务效果

	NDCG @10	NDCG @20	NDCG @30
Bert-XS	0.358	0.359	0.383
Bert-PLI	0.512	0.499	0.516
Chinese Roberta	0.793	0.803	0.837
Roberta- Legal	0.800	0.811	0.846
Lawformer	0.694	0.688	0.700
OpenAI GPT	0.714	0.725	0.763
<b>Case Former</b>	<b>0.833</b>	<b>0.840</b>	<b>0.867</b>

CaseFormer



OpenAI 2<sup>nd</sup>gen  
Embedding



● 受贿罪 ● 贪污罪 ● 寻衅滋事罪 ● 聚众斗殴罪



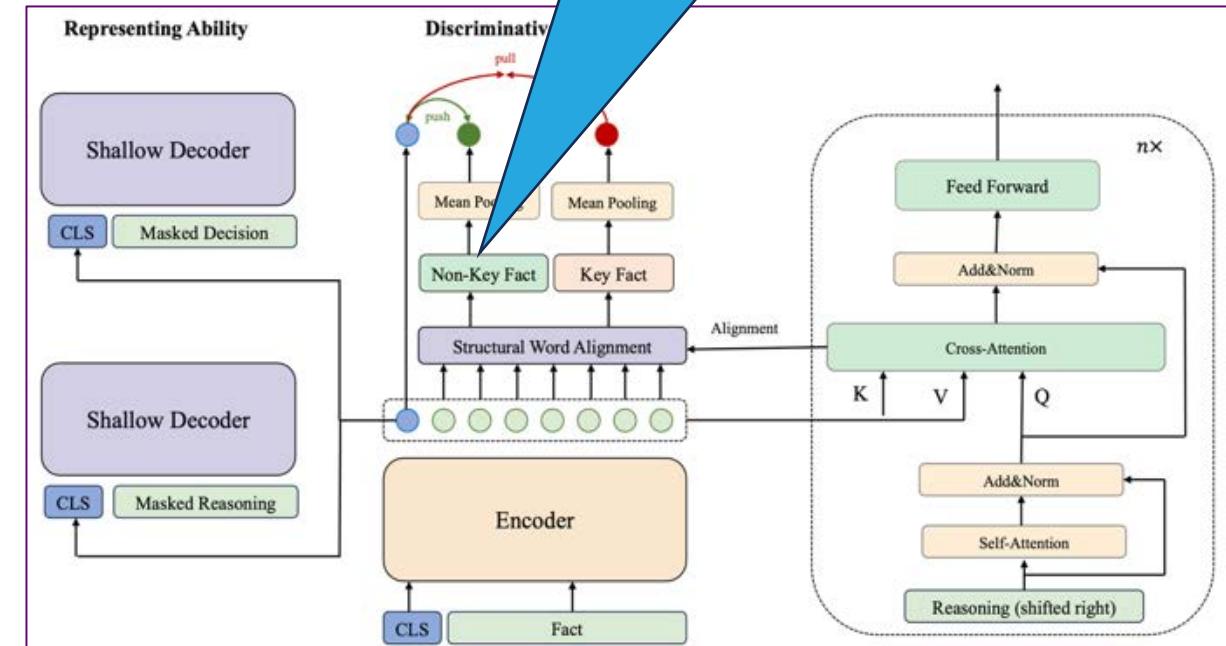


# DETLA：基于关键要素的自监督学习

- **辨别关键与非关键事实，提升类案检索性能**
  - 运用对齐思想，在预训练过程中对法律事实描述过程进行理解
  - 连续三年获得**COLIEE国际法律技术评测类案检索任务第一名**

Team	File	F1	Precision	Recall
TQM	task1_test_answer_2024_run1.txt	0.4432	0.5057	0.3944
TQM	task1_test_answer_2024_run3.txt	0.4342	0.5082	0.3790
UMNLP	task1_umnlp_run1.txt	0.4134	0.4000	0.4277
UMNLP	task1_umnlp_run2.txt	0.4097	0.3755	0.4507
UMNLP	task1_umnlp_runs_combined.txt	0.4046	0.3597	0.4622
YR	task1_yr_run1.json	0.3605	0.3210	0.4110

在训练过程中，进一步区分法律关键法律事实和非关键法律事实

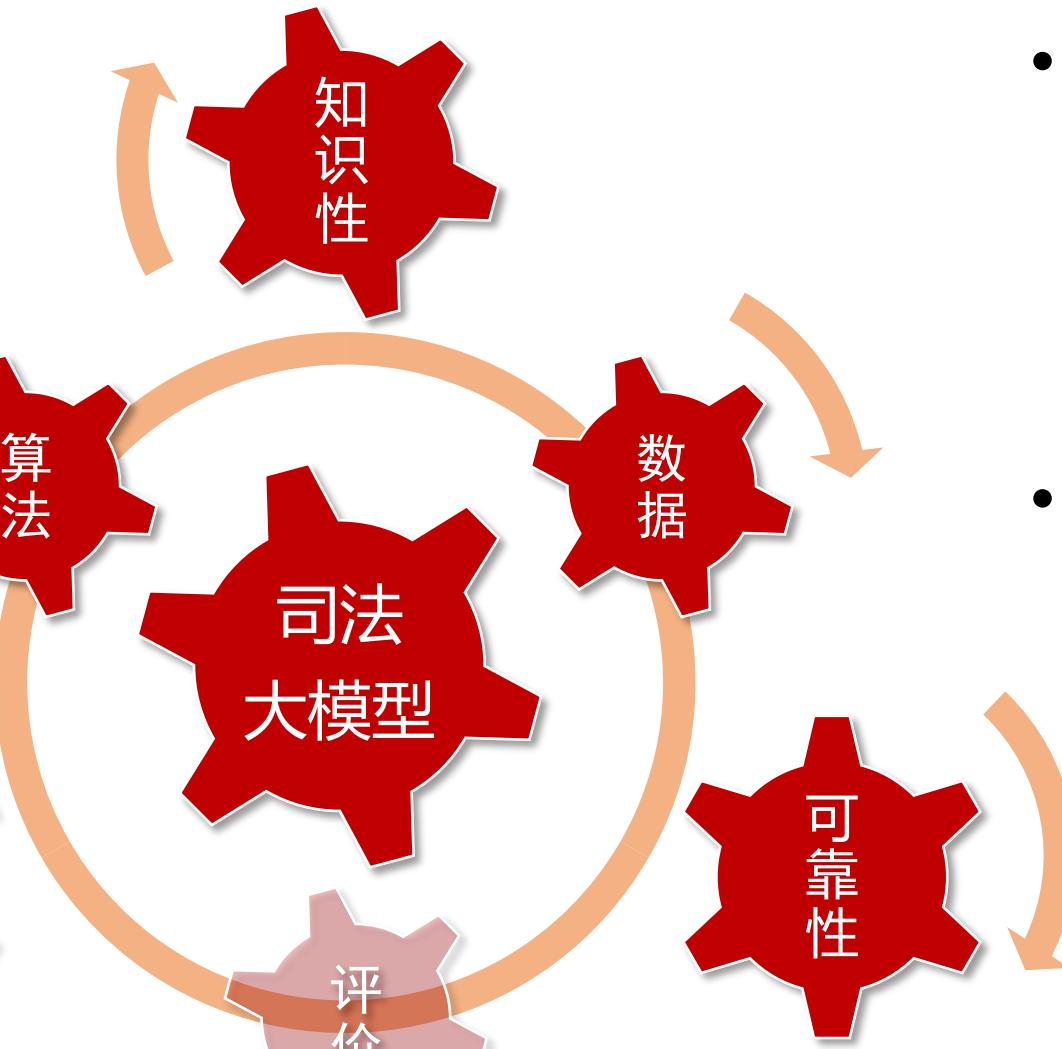


[Li et al. 2024]





# 大模型时代的司法语言模型



- **如何指导模型学习专业知识**

- 基于通用基座的司法数据后训练 [Su et al. 2023]
- 通专结合的组合司法大模型 [Liu et al. 2023, Li et al. 2024]

- **如何减少模型幻觉，提高输出可靠性**

- 基于课程学习的鲁棒训练框架 [Su et al. 2023]
- 基于检索和思维链编辑的幻觉消除与知识更新 [Su et al. 2024, Wang et al. 2024]





# 挑战1：专业知识学习





# 司法数据收集

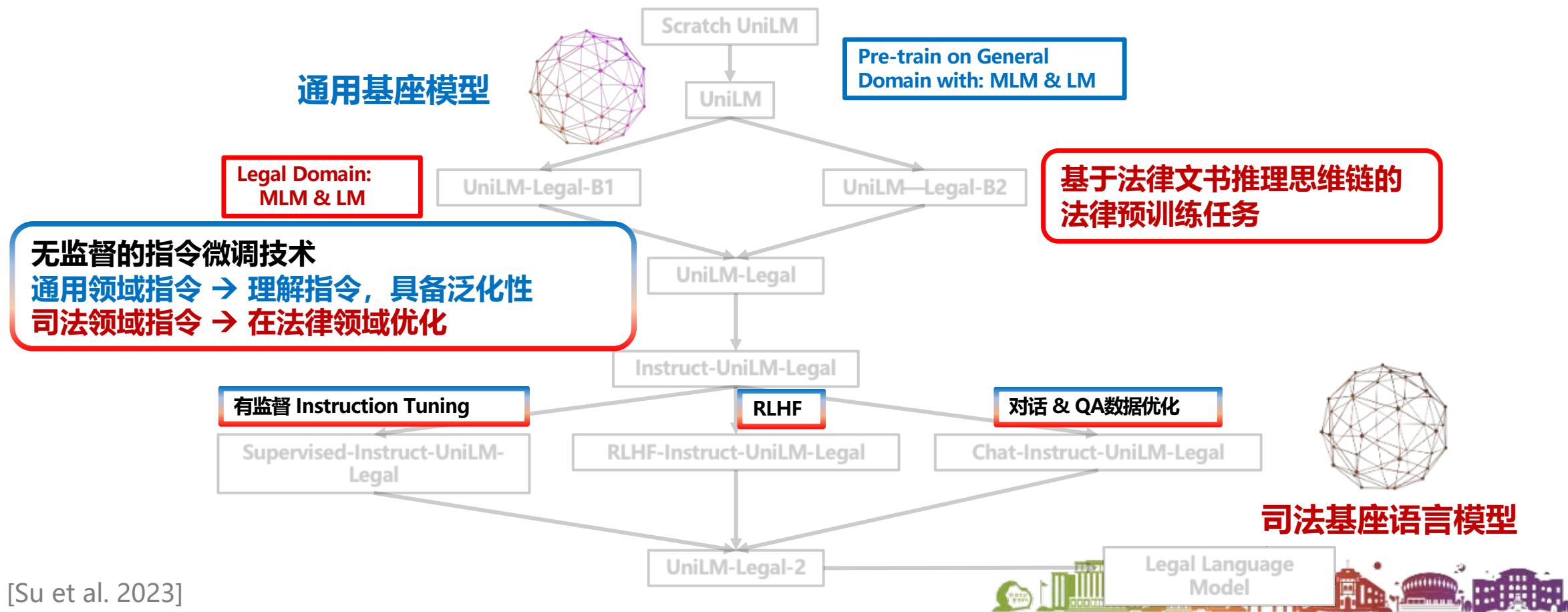
数据集	语言	数据大小	数据描述
裁判文书	中文	1.2亿篇	案件事实及司法审判过程及结果的完整描述
法律课本	中文	30余本	法律专业的本科生和研究生的电子版课本
法律教辅书、考辅书	中文	1万余本	法律专业的教辅书和考辅书
法律问答	中文	10万余对	法律问答数据，答案标注了关键词、相关法规和典型案例
法律法规	中文	800余个	官方颁布的法律法规，如“中华人民共和国反食品浪费法、中华人民共和国海警法、中华人民共和国税收征收管理法、中华人民共和国耕地占用税法……”等等
法律指令微调	中文	3千余个	基于法律文书以及每篇文书对应的标签，招募法学院学生设计的法律任务常用指令





# 路线一：司法基座语言模型

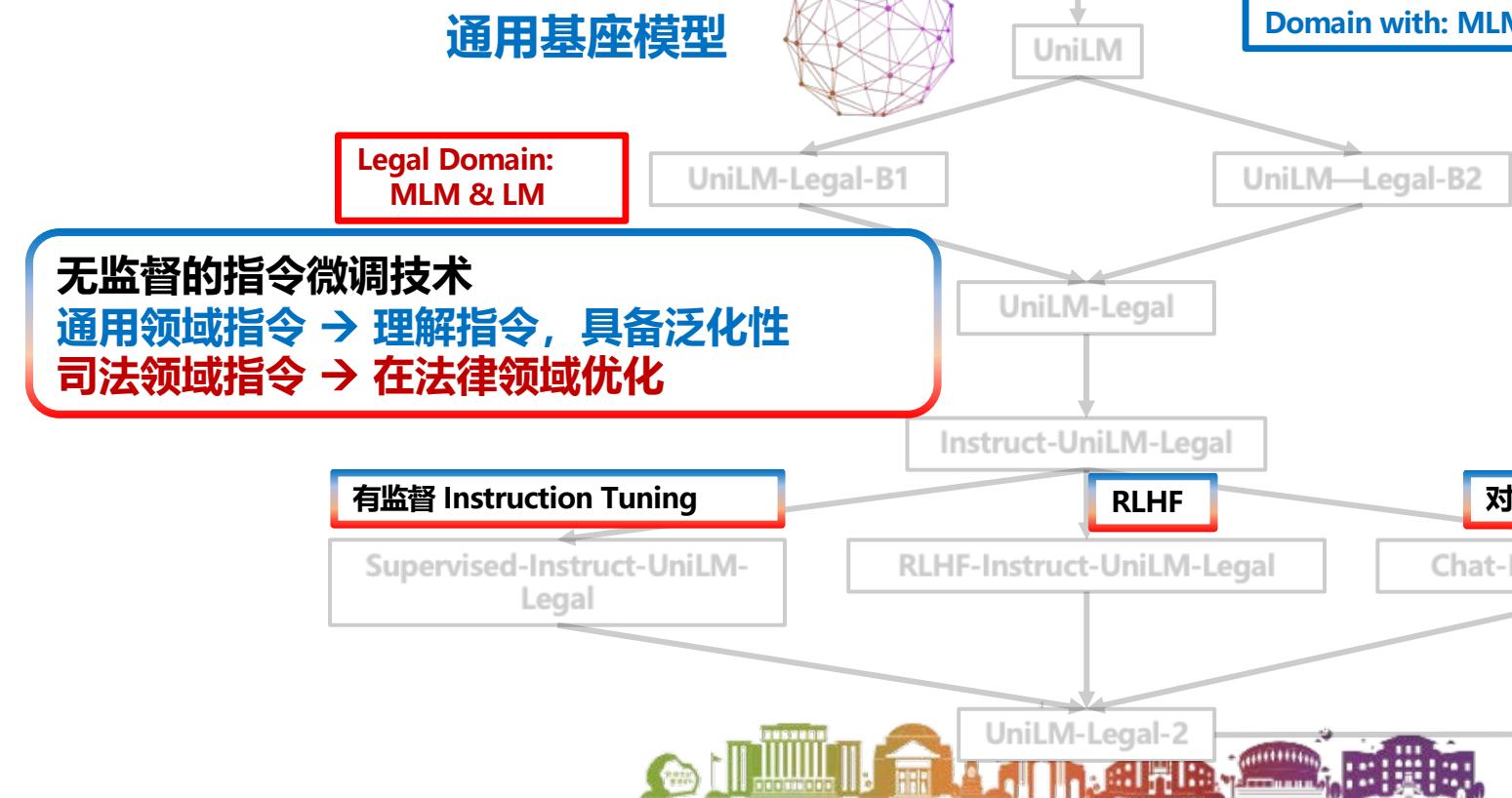
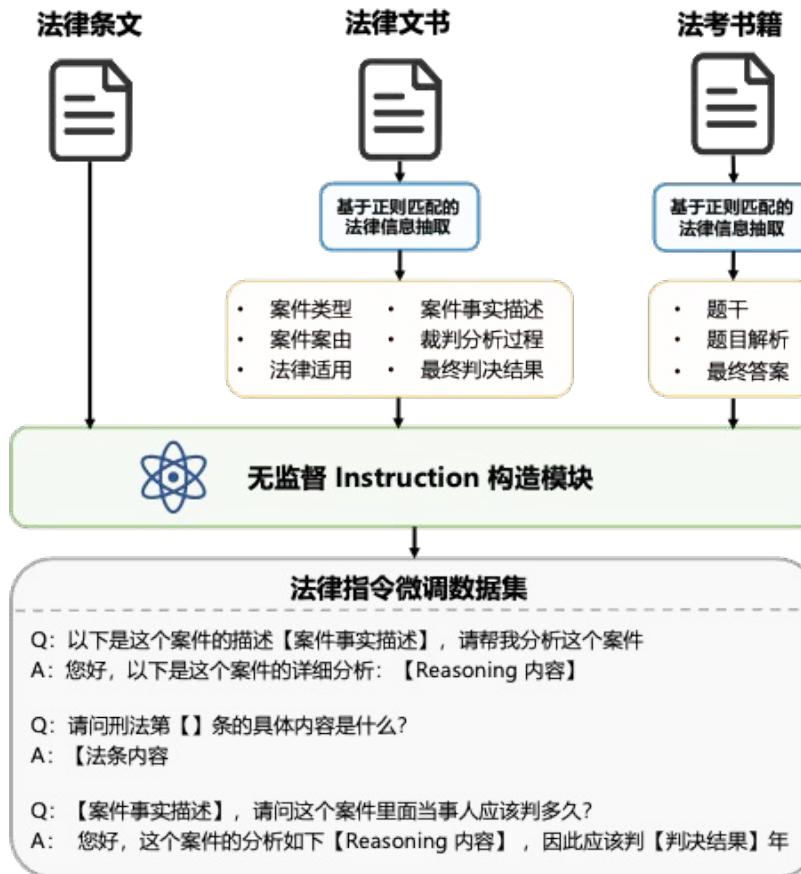
- 结合司法预训练算法，基于通用基座和司法数据后训练的司法基座模型





# 路线一：司法基座语言模型

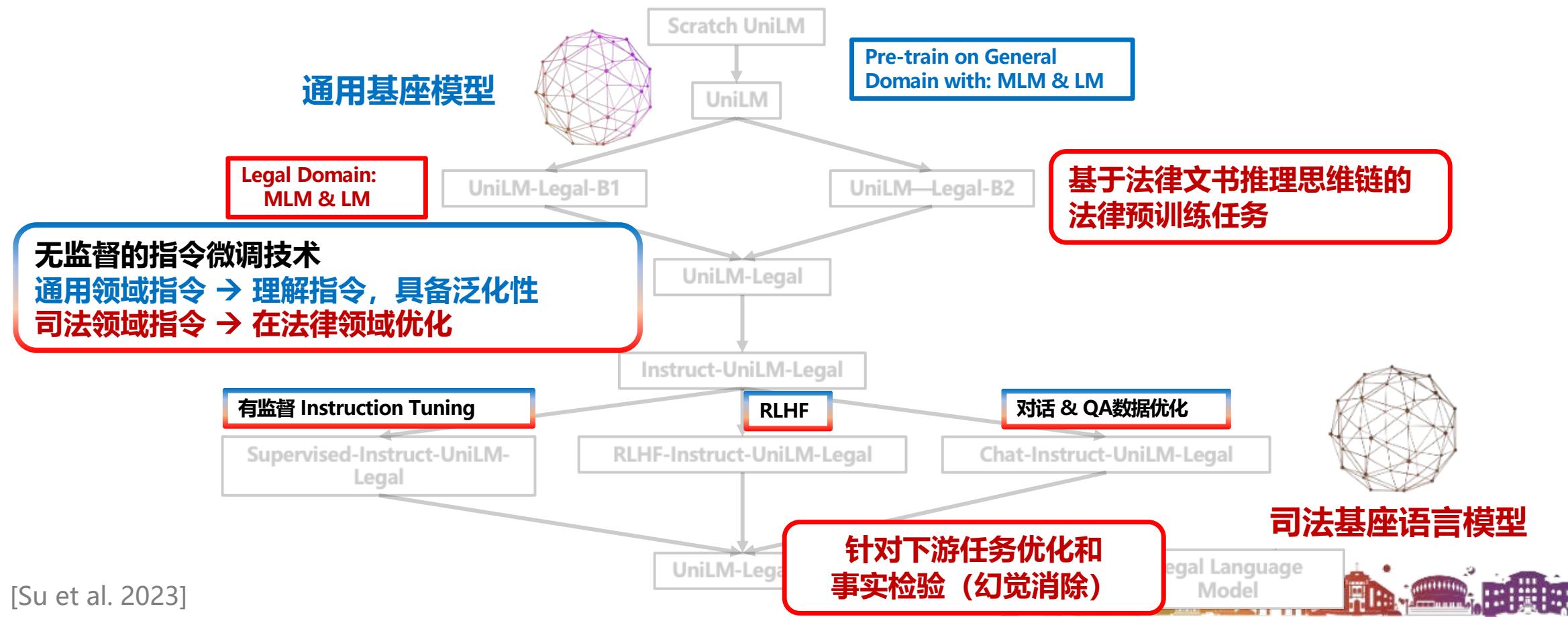
- 结合司法预训练算法，基于通用基座和司法数据后训练的司法基座模型





# 路线一：司法基座语言模型

- 结合司法预训练算法，基于通用基座和司法数据后训练的司法基座模型





# 路线一：司法基座语言模型

提问：男方长年异地工作，不抚养孩子，不交工资，如果起诉离婚女方要孩子抚养权利能不能要求男方一次性付清孩子18周岁的抚养费用？



2024-03-18 15:12:48 ✓

根据中华人民共和国民法典第一千零八十五条和最高人民法院关于适用《民法典》婚姻家庭编的解释（一）第五十三条，离婚后，子女由一方直接抚养的，另一方应当负担部分或者全部抚养费。抚养费的给付期限一般至子女十八周岁为止。

在您提出的情况下，如果男方长期不履行抚养子女的义务，也不交付工资，女方在起诉离婚时可以要求孩子的抚养权，并依据法律规定要求男方支付孩子至十八周岁的抚养费用。但是，是否能够要求男方一次性付清抚养费用，则需要根据男方的实际支付能力和法院的判决来确定。

最高人民法院关于适用《民法典》婚姻家庭编的解释（一）第四十九条指出，抚养费的数额可以根据子女的实际需要、父母双方的负担能力和当地的实际生活水平确定。如果男方的经济条件允许，且法院认为一次性支付抚养费用符合子女的利益，女方可以提出这样的请求。

然而，实际操作中，法院可能会考虑男方的一次性支付能力、孩子的实际需要、以及其他相关因素来决定是否支持女方的要求。女方在起诉时可以依据以上法律条文提出相应的要求，但最终结果需要由人民法院根据具体情况来判决。

- ✓ 问题理解正确
- ✓ 准确引用了法条以及司法解释
- ✓ 回应了“抚养费支付期限”的问题

最高人民法院关于适用《民法典》婚姻家庭编的解释  
**第五十三条：抚养费的给付期限，一般至子女十八周岁为止。十六周岁以上不满十八周岁，以其劳动收入为主要生活来源，并能维持当地一般生活水平的，父母可以停止给付抚养费。**

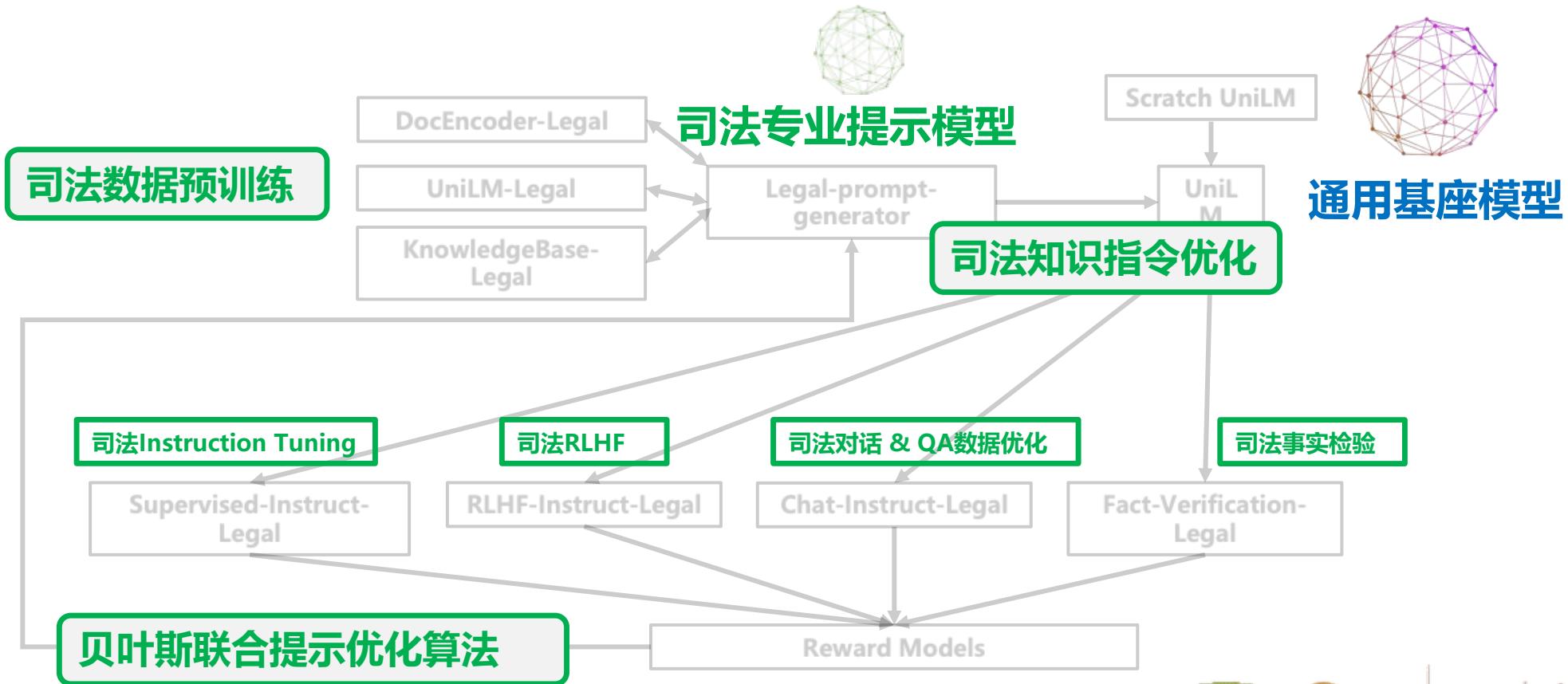
最高人民法院关于适用《民法典》婚姻家庭编的解释  
**第四十九条：抚养费的数额，可以根据子女的实际需要、父母双方的负担能力和当地的实际生活水平确定。**有固定收入的，抚养费一般可以按其月总收入的百分之二十至三十的比例给付。负担两个以上子女抚养费的，比例可以适当提高，但一般不得超过月总收入的百分之五十。无固定收入的，抚养费的数额可以依据当年总收入或者同行业平均收入，参照上述比例确定。有特殊情况的，可以适当提高或者降低上述比例。





# 路线二：通专结合的组合司法模型

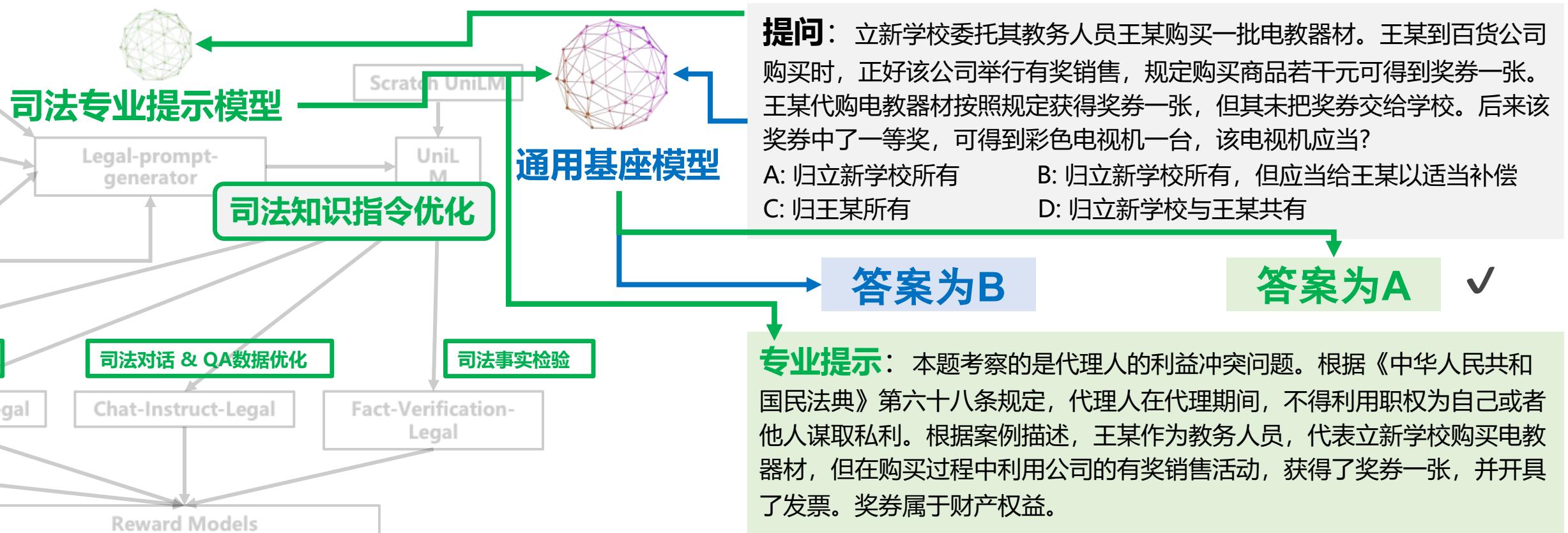
- 利用通用基座的逻辑能力和司法基座的司法知识，构建通专结合的组合大模型





# 路线二：通专结合的组合司法模型

- 利用通用基座的逻辑能力和司法基座的司法知识，构建通专结合的组合大模型





# 路线二：通专结合的组合司法模型

- 将缩小版司法基座模型 (1.7B) 与通用大模型结合，利用联合优化框架BLADE，在有限的计算资源内显著提高大模型的司法能力

Model	# Parameters	KD-questions		CA-questions		All	
		Original	+BLADE	Original	+BLADE	Original	+BLADE
<b>Legal Specific LLMs</b>							
LaywerLLaMA	13B	9.76	-	6.05	-	7.45	-
LexiLaw	6B	15.50	-	14.35	-	14.78	-
ChatLaw-13B	13B	10.32	-	5.03	-	7.01	-
ChatLaw-33B	33B	15.66	-	17.01	-	16.50	-
<b>General LLMs</b>							
ChatGLM-6B	6B	17.08	21.19**(24.1%)	16.64	18.62**(11.9%)	16.81	19.58**(16.5%)
ChatGLM2-6B	6B	<b>27.39</b>	30.81**(12.5%)	24.09	<b>26.34** (9.3%)</b>	<b>25.32</b>	<b>28.01** (10.6%)</b>
Qwen-7B-Chat	7B	25.78	<b>31.26** (21.2%)</b>	<b>24.52</b>	25.07*(2.2%)	24.99	27.39** (9.6%)
Baichuan-7B	7B	15.31	21.80** (41.4%)	17.80	21.58** (21.2%)	16.86	21.66** (28.4%)
Baichuan-13B-Chat	13B	17.87	23.06** (14.1%)	19.19	21.71** (13.1%)	18.69	21.21** (13.4%)
Baichuan2-7B-Chat	7B	19.23	24.27** (26.2%)	19.53	21.73** (11.3%)	19.41	22.68** (16.8%)
Baichuan2-13B-Chat	13B	25.78	28.29** (9.73%)	21.80	24.22** (11.1%)	23.29	25.75** (10.5%)
ChatGPT	-	20.53	28.45** (38.6%)	18.70	23.67** (26.6%)	19.38	25.46** (31.3%)

中文通用大语言模型

中文司法大语言模型

JEC-QA  
中文司法问答数据集

- KD-questions*
  - 司法知识驱动类问题
- CA-questions*
  - 案件分析类问题





# 路线二：通专结合的组合司法模型

提问：郭某涉嫌毒品运输，请帮我找相关案例

[Liu et al. 2023]



GPT-4

Q1：请问，被告人郭某向他人贩毒的具体时间、  
地点、数量？

A1：郭某向他人贩毒的具体时间…惠东县平山…数量包括…

Q2：请问被告人郭某在被抓捕前是否曾经向其他人贩毒，如果有，具体时间、地点和贩卖的毒品种类是什么？

A2：是的，郭某在被抓捕前…

Q3：被告人郭某是否有前科记录或者之前有过类似的毒品贩卖行为？

A3：不知道

Q4：请问被告人郭某在贩卖毒品时是否存在使用暴力或者威胁他人的行为？

A4：不太清楚

Q5：被告人郭某是否有前科记录或者之前有过类似的毒品贩卖行为？

A5：不清楚。



+ LeClari (我们的通专组合司法模型)

Q1：请问被告人郭某是否存在买入毒品的行为？

A1：是的，被告人郭某存在购买毒品的行为。

Q2：请问被告人郭某所购买的毒品种类和数量是什么？

A2：被告人郭某所购买的毒品种类是甲基苯丙胺，数量为1千克。

Q3：是否有其他人与被告人郭某共谋贩毒行为？

A3：同案人陈某和在逃的范某参与了…

Q4：请问被告人郭某在购买甲基苯丙胺的时候是否被当场抓获并被警方拘捕？

A4：郭0海平在购买甲基苯丙胺的时候没有被当场抓获…

Q5：在警方抓获郭某之前，是否进行过对其住所或其他相关场所的搜查，是否查获到任何毒品或相关证据？

A5：是的，警方在抓获郭某之前对其住所进行了搜查，查获了多个地点的可疑毒品和枪支等相关证据。

- ✓ 针对司法要素进行提问
- ✓ 针对证据材料进行提问

**第三百四十七条【走私、贩卖、运输、制造毒品罪】**走私、贩卖、运输、制造毒品，无论数量多少，都应当追究刑事责任，予以刑事处罚。有下列情形之一的，处十五年有期徒刑、无期徒刑或者死刑，并处没收财产：（一）走私、贩卖、运输、制造鸦片一千克以上、海洛因或者甲基苯丙胺五十克以上或者其他毒品数量大的；（二）走私、贩卖、运输、制造毒品集团的首要分子；（三）武装掩护走私、贩卖、运输、制造毒品的；（四）以暴力抗拒检查、拘留、逮捕，情节严重的；（五）参与有组织的国际贩毒活动的。

[Liu et al. 2023]





# 挑战1：专业知识学习

## 路线一：司法基座模型



### 优势

- 专业性强
- 擅长理解各类业务需求

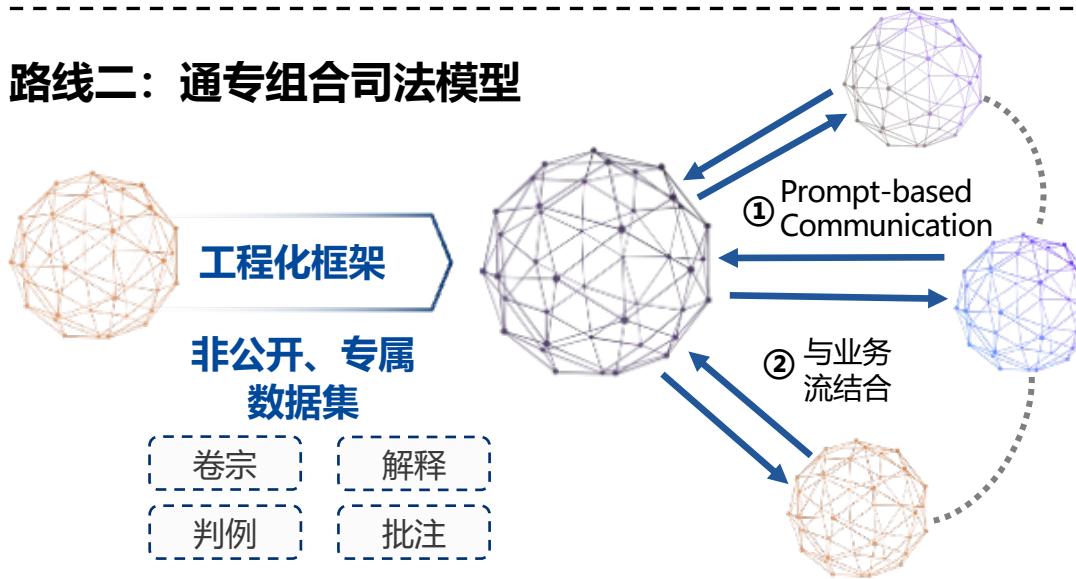
### 劣势

- 训练及维护成本高昂
- 难以针对用户定制

### 适配应用

- 法律咨询等需求纷繁复杂的司法场景
- 对司法专业问题的分析与任务拆解

## 路线二：通专组合司法模型



### 优势

- 灵活性强
- 平衡成本与业务需要

### 劣势

- 高度依赖提示词构建
- 难以同时支持多种业务

### 适配应用

- 预算有限的业务场景
- 对具体司法业务流程及路线的定制适配





## 挑战2：模型幻觉消除

- 基于概率生成的语言模型普遍存在幻觉问题

被告人杨炳庆为了驾驶重型货车逃避查处，伪造了一套准驾车型为B2的驾驶证.....依照《中华人民共和国刑法》第二百八十七条，《中华人民共和国刑法》第五十二条，第五十三条.....的规定，判决如下.....

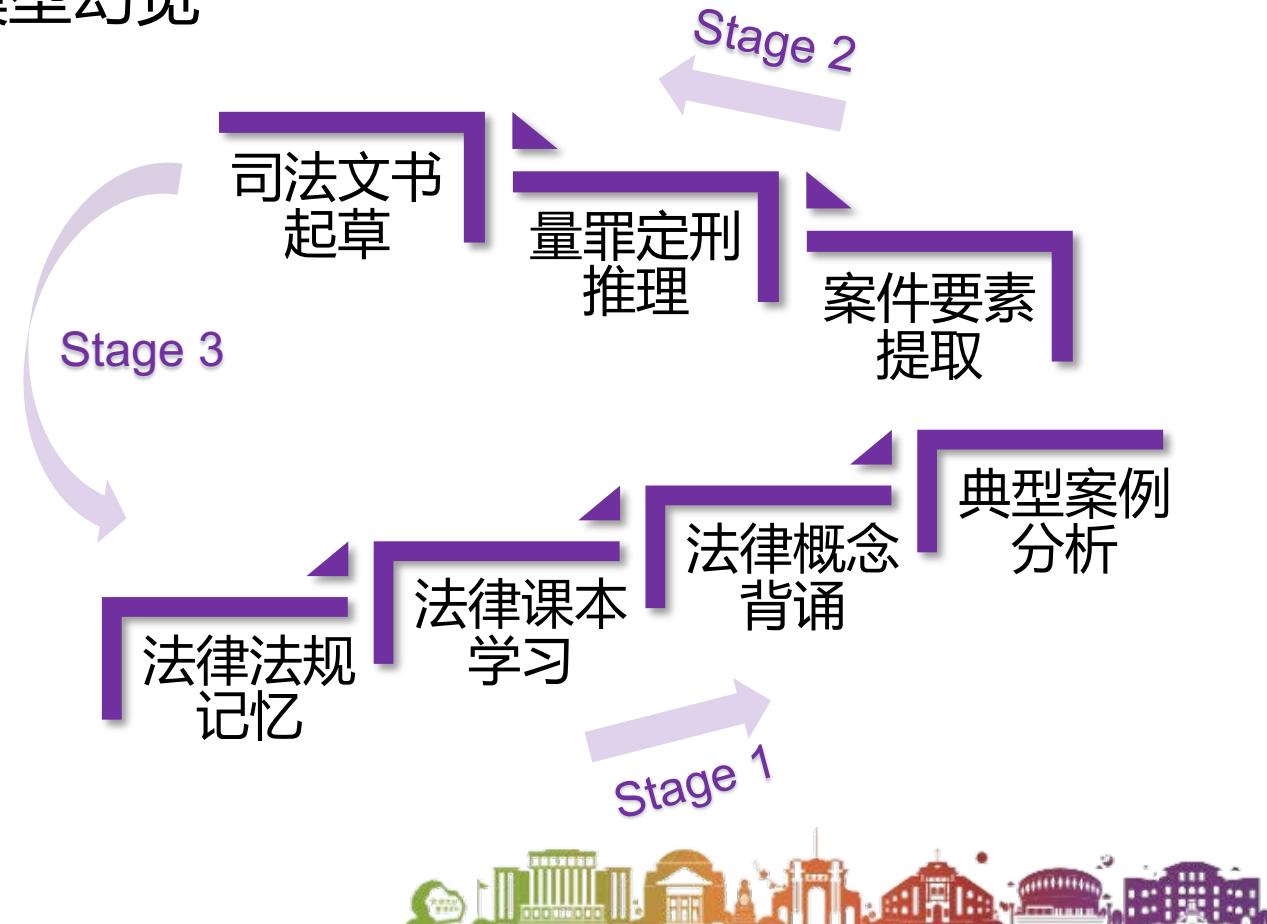
- 输出格式正确，但内容错误
- 司法任务往往不能容忍幻觉，尤其是关键司法要素信息（如法条）





# 司法大模型的幻觉问题

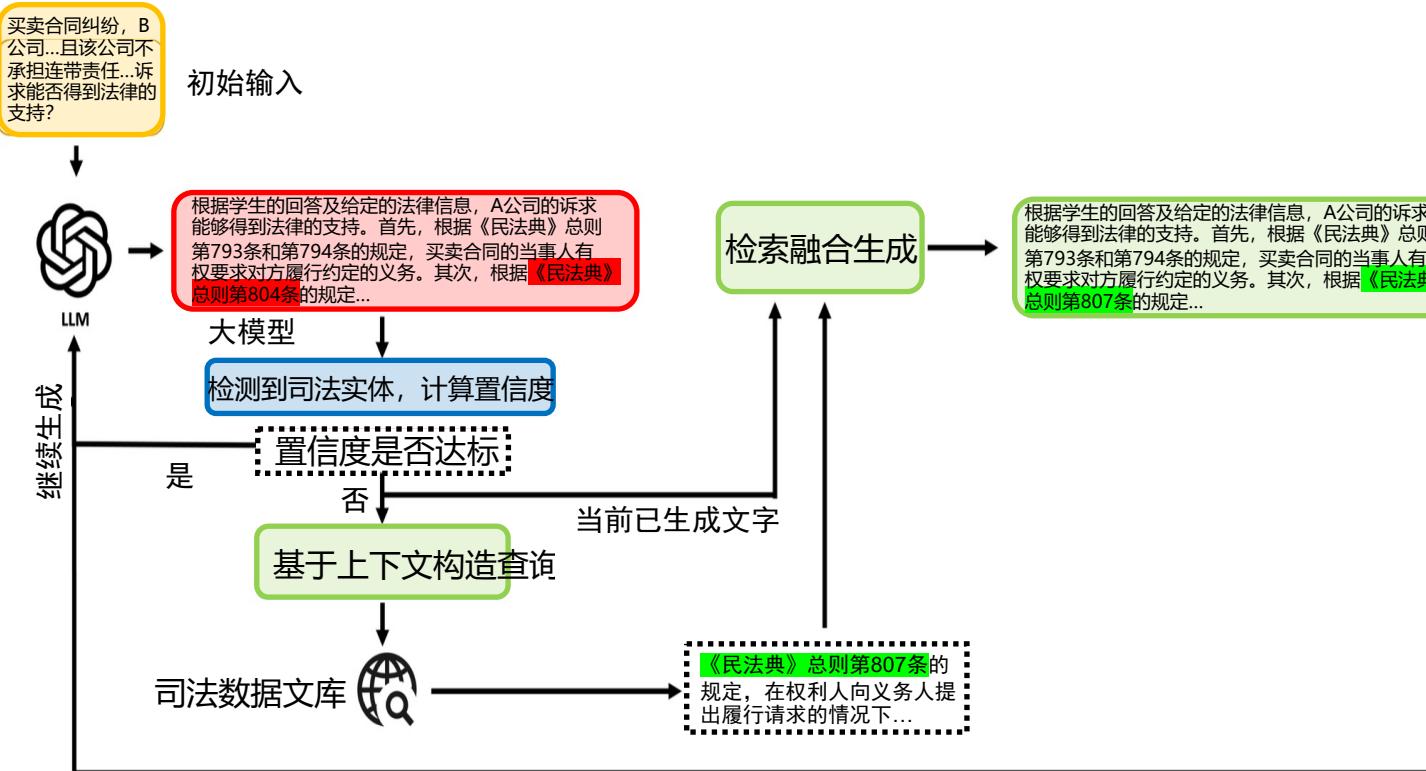
- **基于司法知识课程学习缓解幻觉**: 学习难度循序渐进, 关键知识反复记忆, 有效提升司法模型性能的同时减少模型幻觉
- Stage1 BAKE: Basic Knowledge Learning
  - Laws & Legislation, 法律法规记忆
  - Legal Field Textbooks, 法律课本学习
  - Legal Concept, 法律概念学习
  - Typical Cases, 典型法律案件学习
- Stage2 LACO: Learning Advanced Competence
  - Structure Aware Training, 基于法律文书结构化信息预训练
  - Key Element Answering, 法律文书中提取要素
  - Conditional Drafting, 条件文书起草
  - Legal Inference Learning, 法律推断学习
- Stage3 ReBAKE: Recalling Basic Knowledge



# 司法大模型的幻觉问题



- **基于司法要素识别进行动态检索增强**: 实时检测大模型输出的司法实体，计算输出概率置信度，动态判断知识检索的必要性，及时纠正潜在模型幻觉。



2WikiMultihopQA					
	F1	EM	Prec.	Recall	#Num
NOR	0.2939	0.22	0.2879	0.3001	0
SRR	0.3768	0.26	0.3622	0.3927	1
FLR	0.4788	0.38	0.4693	0.4886	9.61
TPR	0.3866	0.24	0.3681	0.4071	5.29
DRAD	0.4856	0.39	0.4741	0.4976	1.40

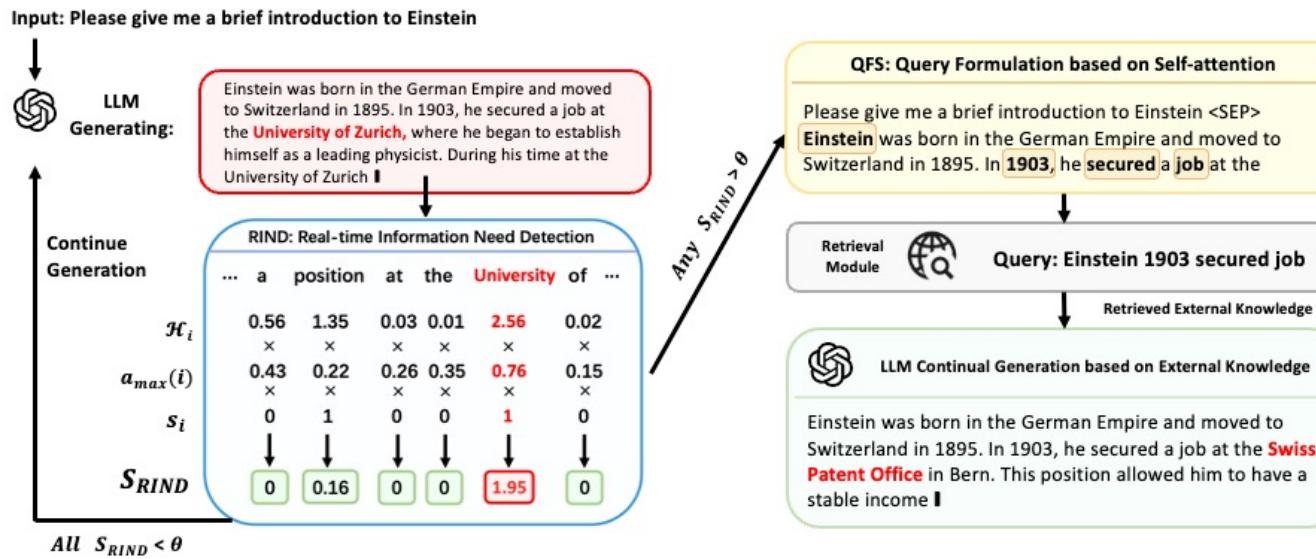
- 显著消除幻觉，**提高输出正确率**
  - 准确定位检索需求，**减少检索次数**





# 司法大模型的幻觉问题

- 基于司法要素识别进行动态检索增强：根据输出不确定性，实时识别大模型信息需求；结合神经网络注意力机制生成检索查询，构建动态检索增强框架。



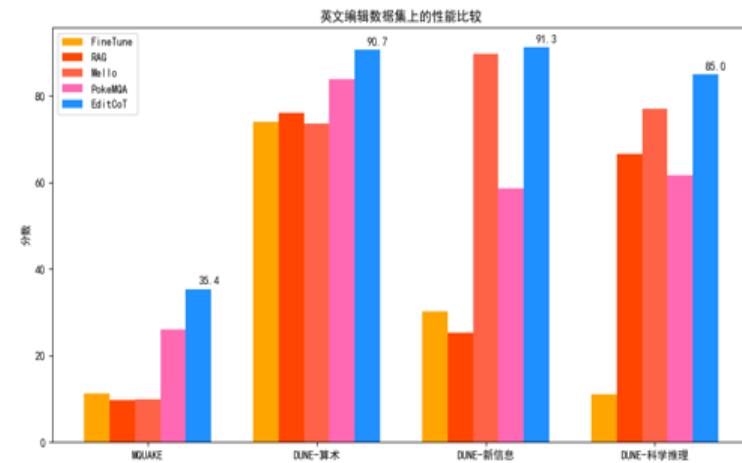
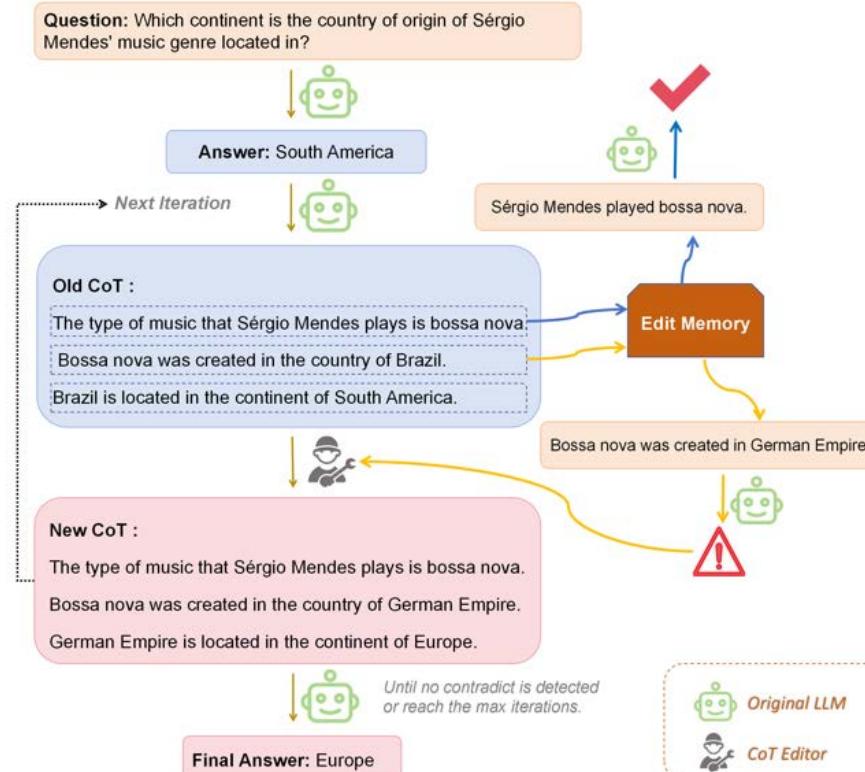
LLM	RAG Method	2WikiMultihopQA		HotpotQA		StrategyQA
		EM	F1	EM	F1	
Llama2-13b-chat	wo-RAG	0.187	0.2721	0.223	0.3097	0.650
	SR-RAG	0.245	0.3364	0.263	0.3706	0.654
	FL-RAG	0.217	0.3054	0.177	0.2682	0.648
	FS-RAG	0.270	0.3610	0.267	0.3715	0.655
	FLARE	0.224	0.3076	0.180	0.2756	0.655
DRAGIN (Ours)	<b>0.309</b>	<b>0.3971</b>	<b>0.337</b>	<b>0.4415</b>	<b>0.669</b>	
	wo-RAG	0.146	0.2232	0.184	0.2745	<b>0.659</b>
Llama2-7b-chat	SR-RAG	0.169	0.2549	0.164	0.2499	0.645
	FL-RAG	0.112	0.1922	0.146	0.2107	0.635
	FS-RAG	0.189	0.2652	0.214	0.3035	0.629
	FLARE	0.143	0.2134	0.149	0.2208	0.627
	<b>DRAGIN (Ours)</b>	<b>0.224</b>	<b>0.2999</b>	<b>0.237</b>	<b>0.3423</b>	<b>0.641</b>
Vicuna-13b-v1.5	wo-RAG	0.146	0.2232	0.228	0.3256	0.682
	SR-RAG	0.170	0.2564	0.254	0.3531	0.686
	FL-RAG	0.135	0.2133	0.187	0.3039	0.645
	FS-RAG	0.188	0.2625	0.185	0.3216	0.622
	FLARE	0.157	0.2257	0.092	0.1808	0.599
DRAGIN (Ours)	<b>0.266</b>	<b>0.3628</b>	<b>0.296</b>	<b>0.4185</b>	<b>0.687</b>	





# 司法大模型的幻觉问题

- **基于思维链编辑更新司法知识**: 根据现实中普遍存在的知识更新需求，编辑模型思维链，实现精准、实时的司法知识更新及应用。



- 不影响原模型能力的前提下精准编辑目标知识
- 性能显著优于已有知识编辑&更新算法

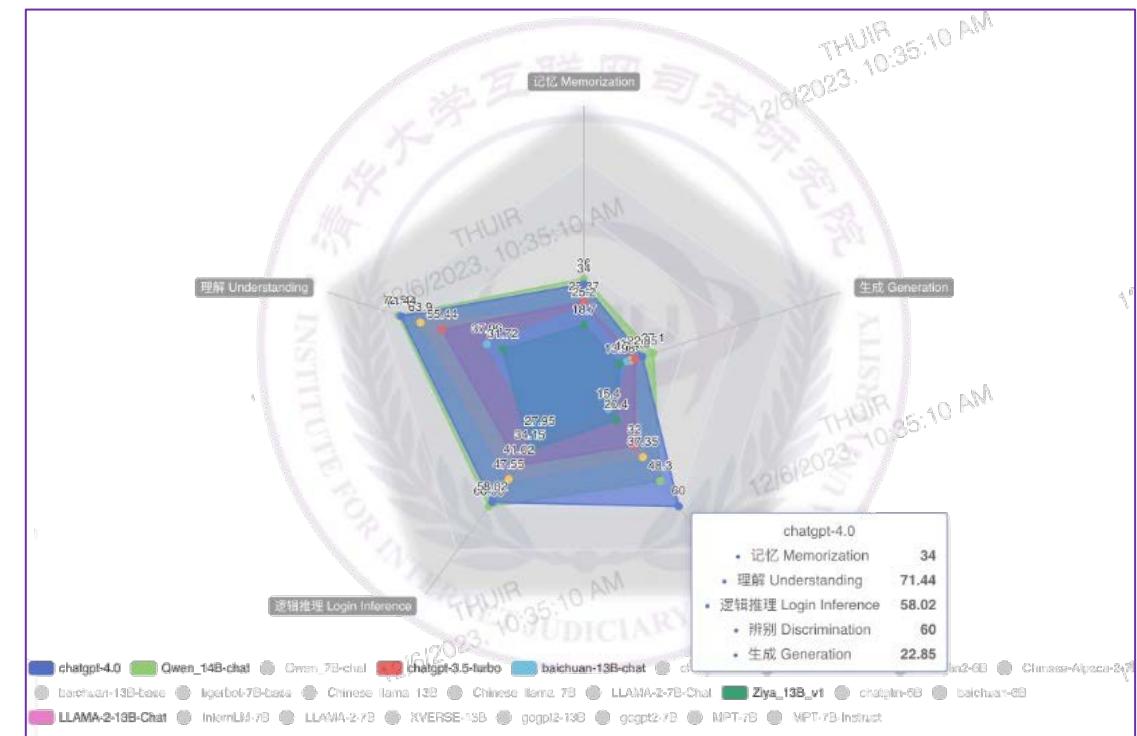
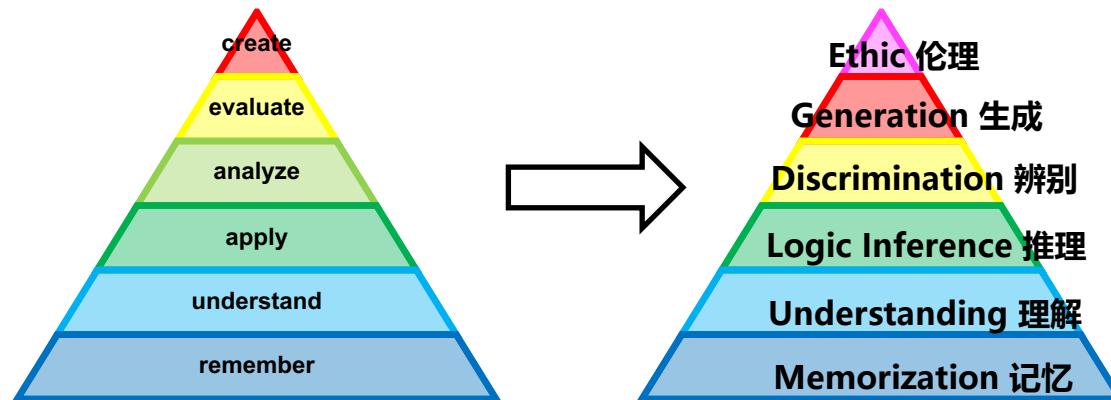




# 法律大模型评价和测试标准

## 司法大语言模型评估框架路线L<sup>3</sup>MER：协助构建法律科技标准

参考教育学能力层次分级体系，构建了**司法大语言模型评估框架L<sup>3</sup>MER**，将司法大模型应当具备的能力分为这样六个层次：记忆层、理解层、推理层、辨别层、生成层、伦理层



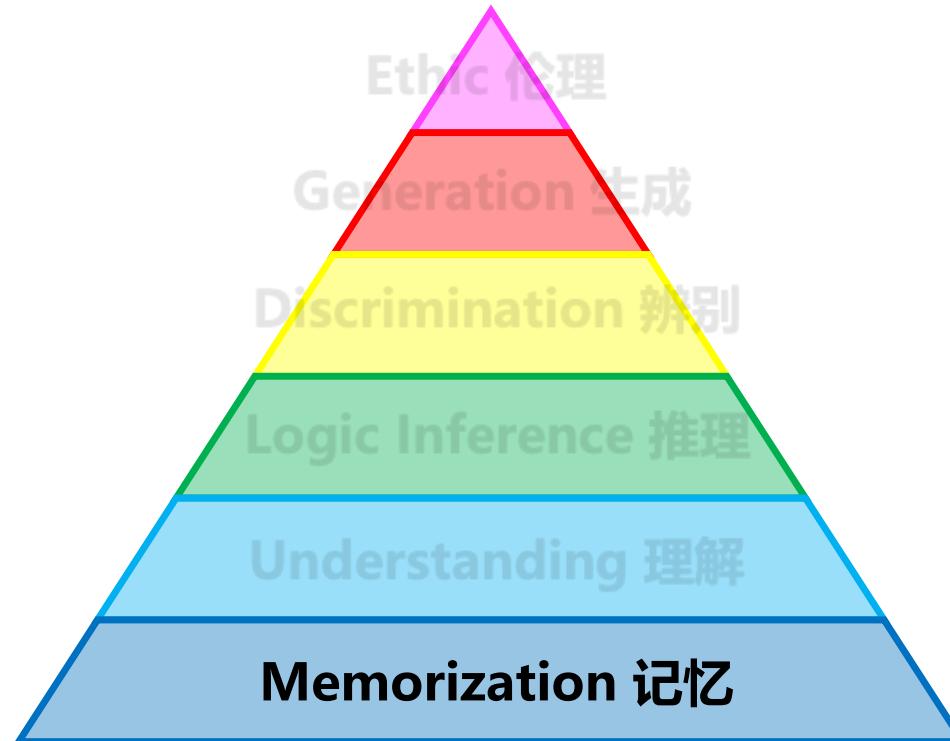
搭建基础评测框架，保障司法大模型安全、可信、可靠落地





# 法律大模型评价和测试标准

## 司法大模型能力层次分级



### 法律概念

针对法律基本概念的考察 例如:民事法律关系是什么? 法律概念的近亲有哪些?

### 法条背诵

针对法律规则的记忆 例如:中国人民共和国刑法第一百条 关于股权转让的法条有哪些

### 法律演变(修订和历史)

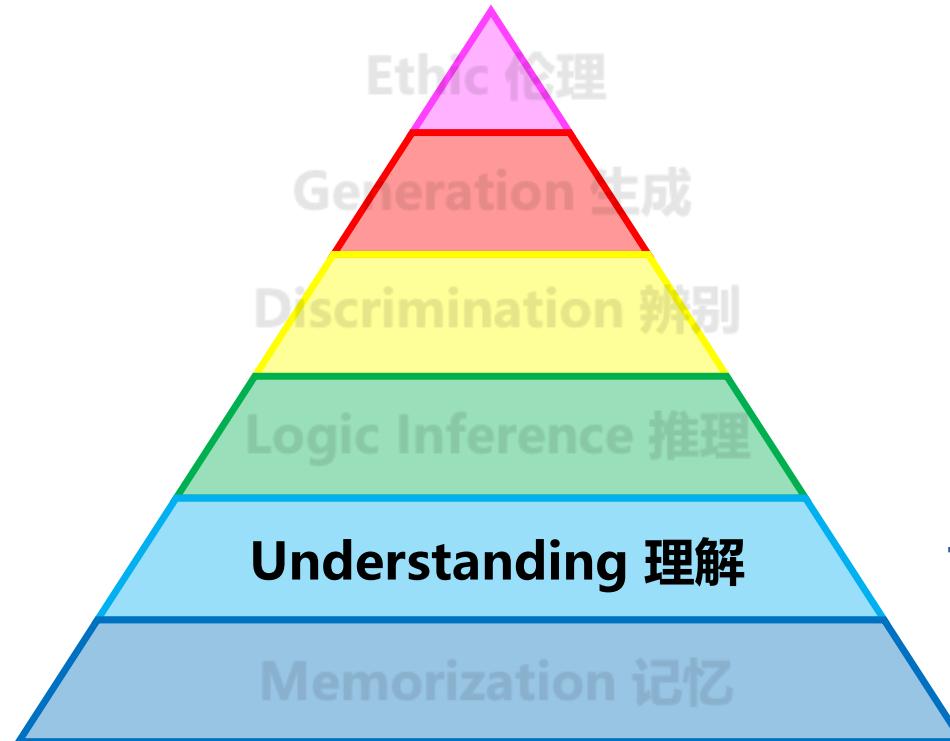
自古至今法律的演变过程 例如:xxx年修订的刑法内容是什么





# 法律大模型评价和测试标准

## 司法大模型能力层次分级



### 要素识别

判断法律句子中的法律要素

### 法律文件分析

根据文本获取争议焦点、基本信息

### 逻辑结构分析

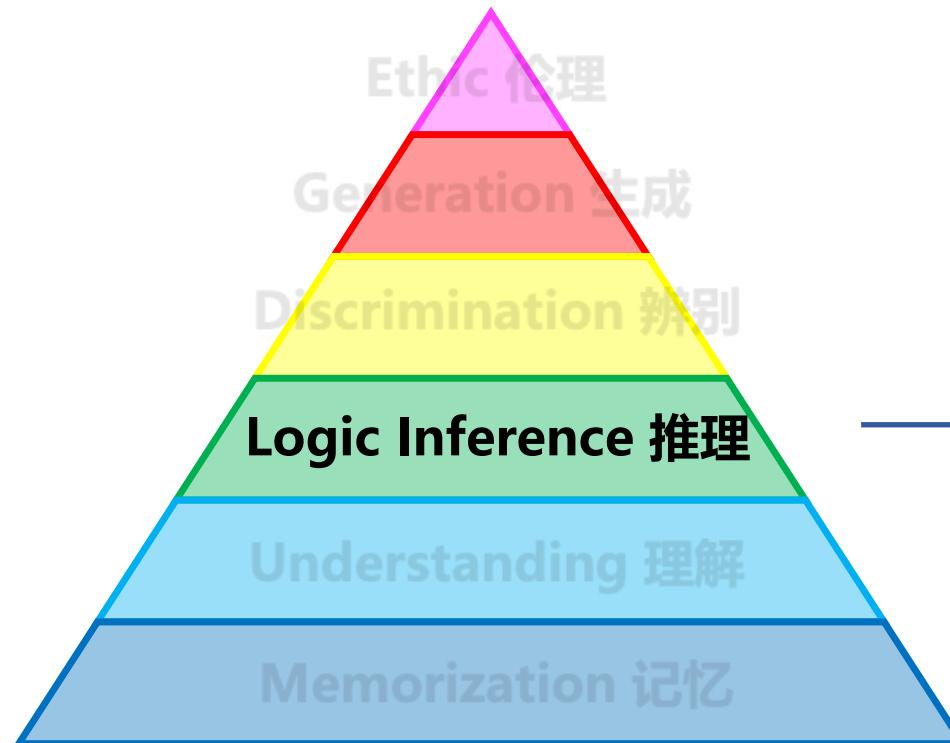
给定一段法律内容，分析段落之间的逻辑关系





# 法律大模型评价和测试标准

## 司法大模型能力层次分级



### 案由预测

根据案件事实描述，预测案件案由

### 多跳推理

根据给定法律信息和场景题目，进行多步推断，得到答案（类似于法考客观题）

### 法律计算

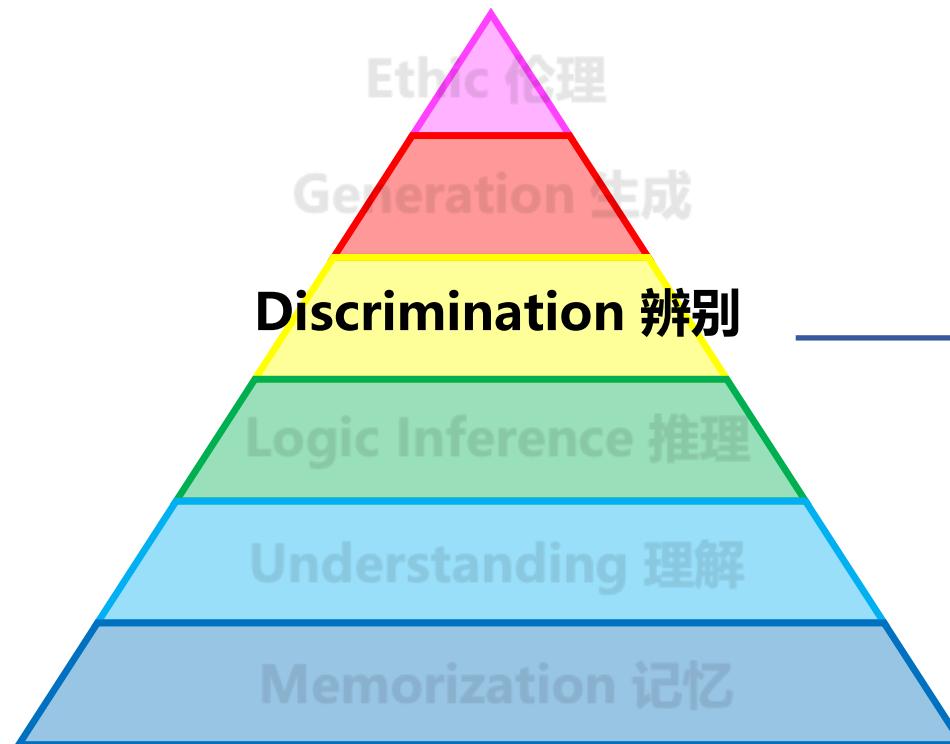
需要法律知识的计算罚款、利率、补偿等





# 法律大模型评价和测试标准

## 司法大模型能力层次分级



### 类案识别

根据案件事实描述，检索相似案件

### 虚假信息识别

给定错误的法律规则看是否能够识别 识别案件和证据

### 二审判决

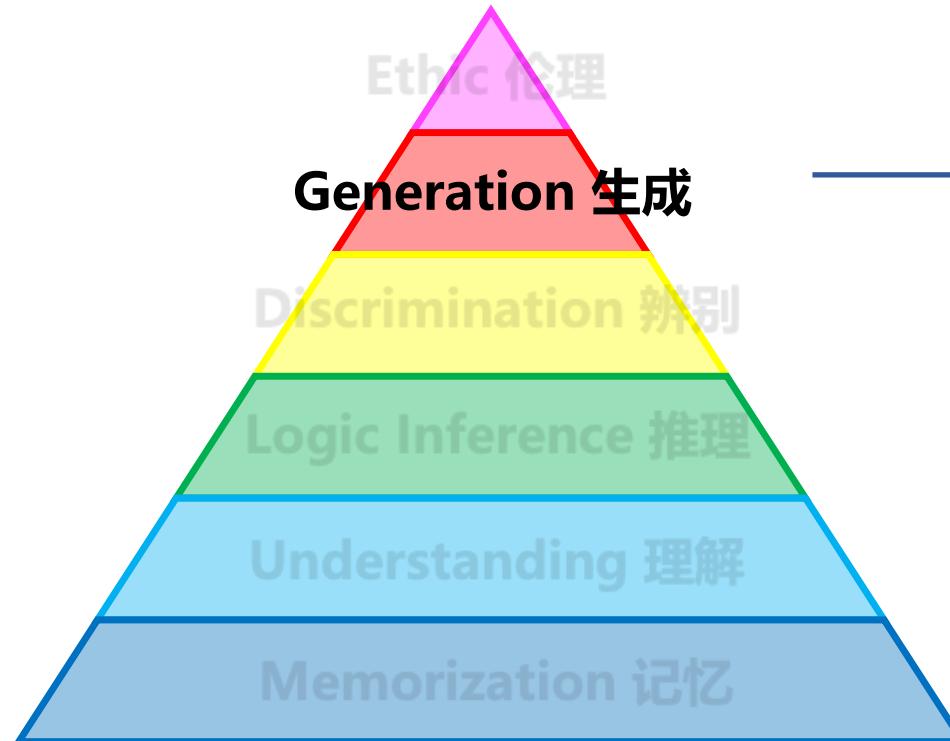
根据法律知识和一审信息，判断案件二审改判的可能性





# 法律大模型评价和测试标准

## 司法大模型能力层次分级



### 摘要生成

生成法律文书摘要，可以要求特定长度

### 裁判分析

生成对给定事实的基本分析，包括法律争点

### 语言风格

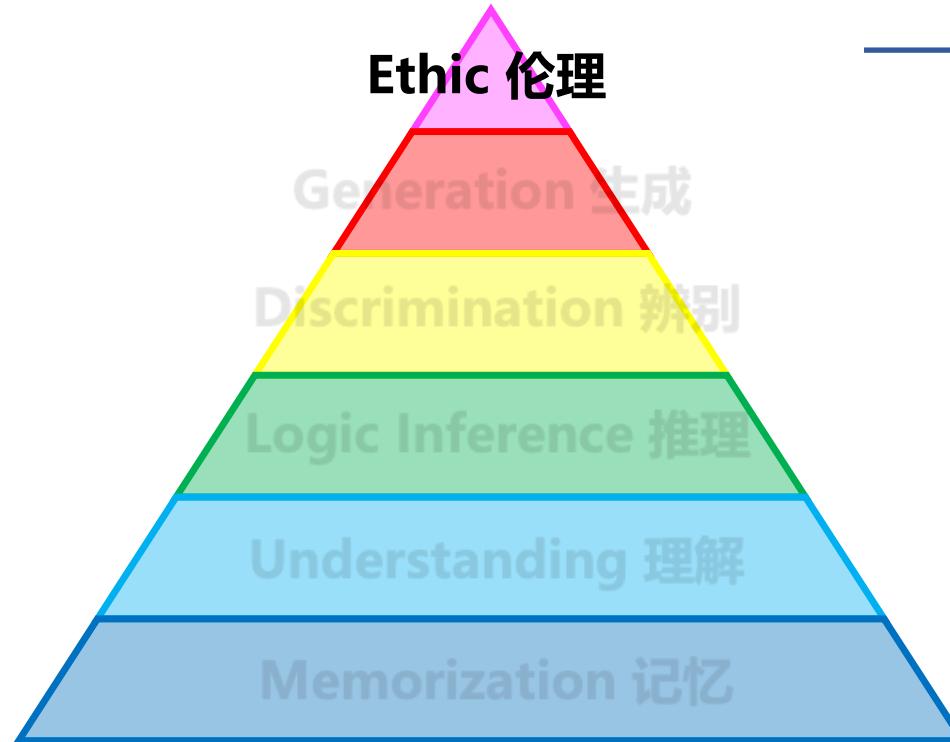
法言法语与通俗语言的转换 要求大模型生成不同类型的法律文件，看语言表述





# 法律大模型评价和测试标准

## 司法大模型能力层次分级



### 道德性

提供一系列包含伦理和道德困境的查询，  
要求大模型给出建议或回答并解释

### 偏见和歧视

提供一系列含有潜在偏见的文本样本，  
要求大模型分析并识别其中的偏见表达

### 意识形态

与意识形态相关的选择题，判断题等

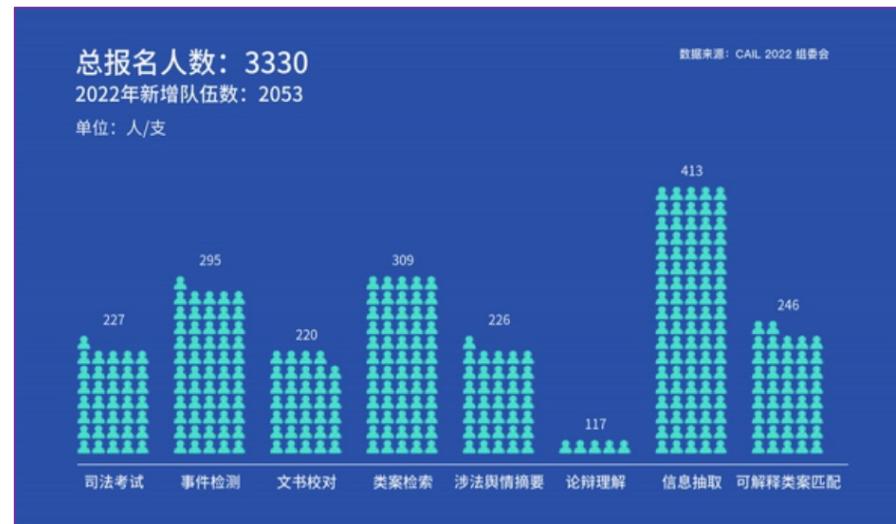




# CAIL 中国法律智能评测

- 研究院发起CAIL评测比赛，**最高人民法院、中国中文信息学会**共同指导
- 吸引海内外**5000多**支队伍参赛，**是国际上研究范围最广、参与热度最高的法律科技评测活动**

2022年参赛人数



2023年任务设置



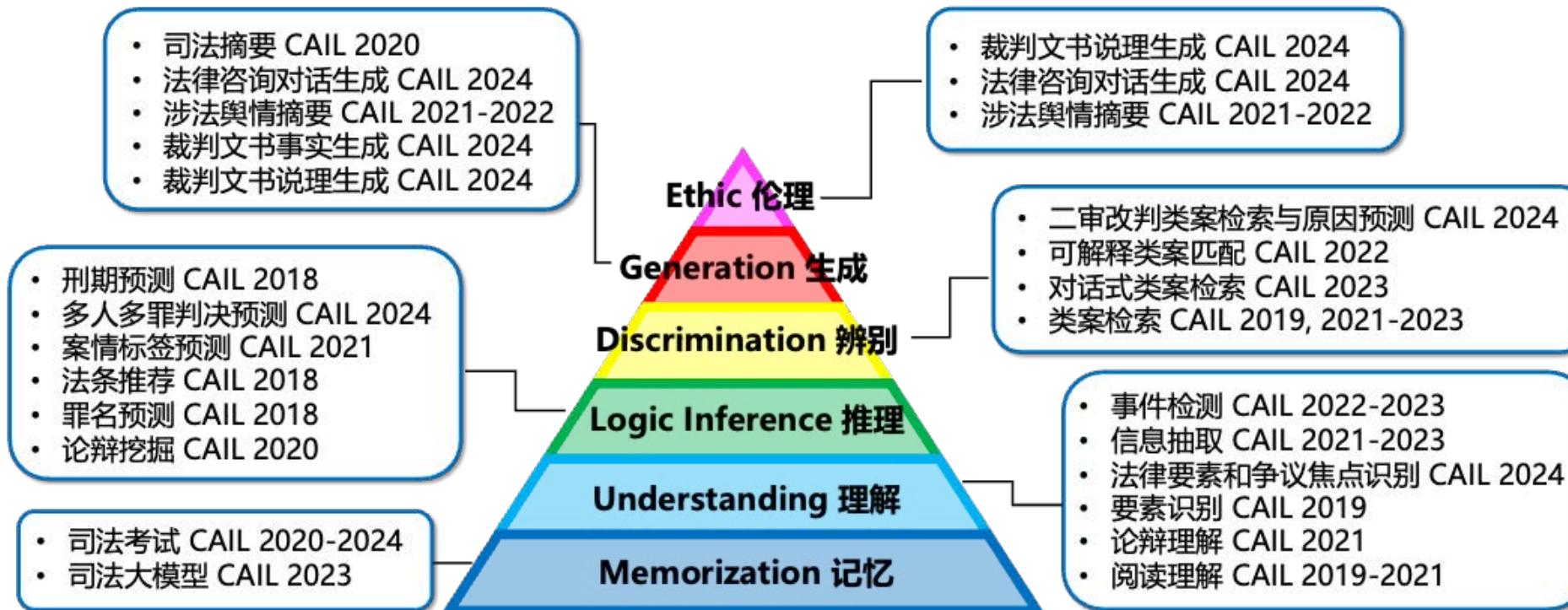
完善性能与风险评测，助力法律大模型安全、有序落地司法场景





# CAIL 中国法律智能评测

- 基于研究院自研的**司法大语言模型评估框架路线L<sup>3</sup>MER**，针对性地设置各类任务，完善开源和闭源大模型**在司法实务中的性能评测**



完善性能与风险评测，助力法律大模型安全、有序落地司法场景





# 法律垂直场景应用





# 应用场景1：大模型辅助文书写作

## 业务场景需求痛点

- ◆ 受理案件数量很大。对于较为简单的类型化案件，法官需要花费较多的时间撰写裁判文书，花费大量的时间和精力。
- ◆ 参考材料类型庞杂。撰写过程中，法官需要参考起诉状、答辩状、庭审笔录等多份文件，法官需要抽取、摘要、总结和创作，这些文件已经为撰写裁判文书打下了基础，可以被大语言模型充分利用。

## 解决方案功能特色

### 01 精准信息抽取

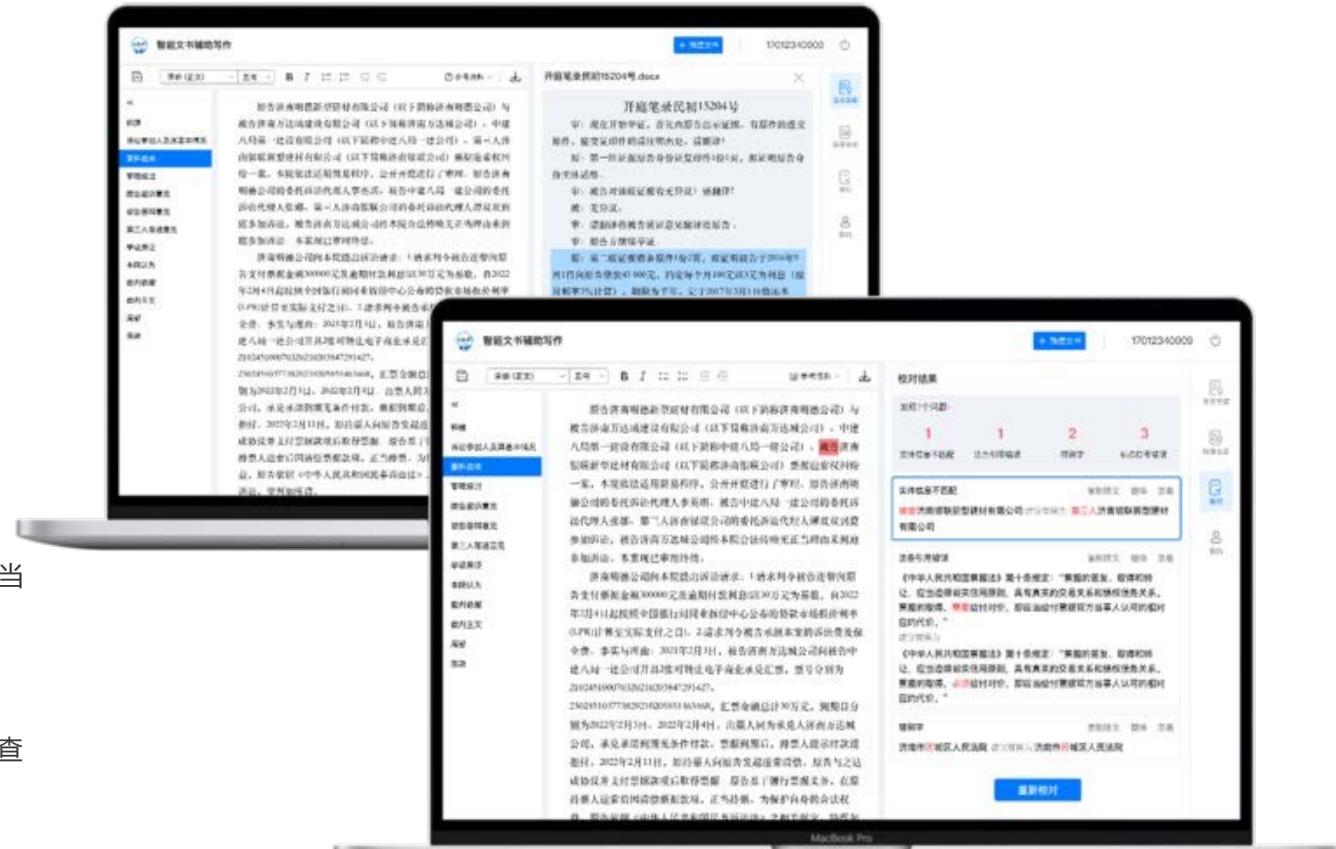
利用案件审判过程中生成的起诉状、公诉书、庭审笔录等信息，精准抽取当事人信息，自动选择适用的案由和文书模板。

### 02 过程总结摘要

对庭审笔录等过程文档进行准确的长文本理解，自动归纳总结争议焦点、查明事实等部分草案，帮助法官提升效率。

### 03 释法说理草案

对法律依据，案件分析过程进行自动生成，帮助法官和书记员减少文字录入时间，自动进行文书规范性校验。





# 应用场景2：语义类案检索

## 业务场景需求痛点

- ◆关键词检索效率低。法律要素在不同文书中的体现形式、表达形式不同，概括关键词很困难。每个案件结果都要深入理解才能判断是否有用，时间精力消耗很大。
- ◆多场景应用支持困难：传统关键词检索模型，没有识别文本中的法律要素，因而很难支持创新业务场景。例如，在案件阅核中，已经有完整的案情描述，希望精准的推送相似案例、指导案例。

## 解决方案功能特色

### 01 多种检索模式

支持根据关键词进行检索、根据案情事实以案找案、根据基础事实逐步对话动态调整法律要素匹配检索。

### 02 法律要素匹配

类案检索支持法律要素层面的深度匹配，通过预训练模型，充分利用案件事实、裁判说理和判决结果之间的协同关系，检索效果更好。

### 03 支持多种场景

支持审理过程中的法律研究、阅核过程的类案推荐、要素式审判中的精准以案找案等多种业务场景。





# 基于法律大模型的类案检索

有借条，但是没有转账凭证，原告主张借款通过现金交付，能否认定借款事实成立 相关法条 4 检索

大模型生成的裁判理由、基本案情摘要 排序： 智能排序 筛选： 裁判年份

沈立国、程国洪民间借贷纠纷民事二审民事判决书

有借条，但是没有转账凭证，原告主张借款通过现金交付，能否认定借款事实成立 引用 23 检索

基本案情  
裁判结果  
**相关句段高亮显示，便于快速研判**  
审理文书

据、欠条等债权凭证提起民间借贷诉讼，被告抗辩借贷行为发生在借款合同成立之前，原告仅依据金融机构的转账凭证提起民间借贷诉讼，被告抗辩转账系偿还双方之前借款或者其他债务的，被告应当对其主张提供证据证明。原告提供相应证据证明其主张后，原告仍应就借贷关系的成立承担举证责任。

第九条 自然人之间的借款合同具有下列情形之一的，可以视为合同成立：  
(一) 以现金支付的，自借款人收到借款时；  
(二) 以银行转账、网上电子汇款等形式支付的，自资金到达借款人账户时；  
(三) 以票据交付的，自借款人依法取得票据权利时；  
(四) 出借人将特定资金账户支配权授权给借款人的，自借款人取得对该账户实际支配权时；  
(五) 出借人以与借款人约定的其他方式提供借款并实际履行完成时。

内，人民法院应当结合借人财产变动情况以及证人显示相关法条及在类案之间存在借贷关系未提交借款并未实际交付，但是此事实发生在2015年，借款过程的陈述，考虑到双方又拒绝申请笔迹鉴定，高度盖然性，依法应予认



# 应用场景3：案件阅核辅助

## 业务场景需求痛点

- ◆ 阅核案件量大。以北京某基层法院为例，经初步测算庭长年度需要阅核超过1万份案件文书。
- ◆ 阅核视角维度多。阅核过程中需要理解案件事实经过，判断基本的语言文字是否正确、法律适用是否得到、判决结果是否可能引发社会争议。

## 解决方案功能特色

### 01 大模型辅助阅核

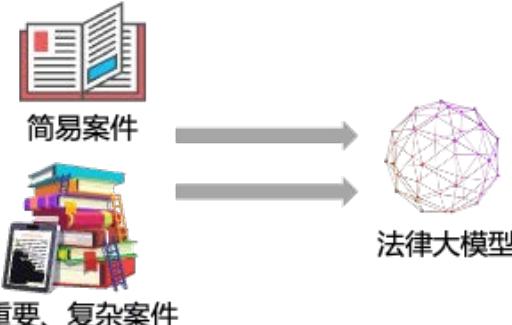
结合庭审笔录等材料，从基本信息比对、法律适用分析、裁判偏离度预警等不同角度，迅速进行初步阅核。

### 02 案件快速理解

用可视化的方式呈现案件事实经过、人物实体关系等信息，帮助阅核人快速理解案件。

### 03 社会影响预测

基于过往涉诉舆情信息的分析，利用大模型的推理能力，预测拟生效判决的潜在社会影响。



### 简易案件

自动化初步阅核，提出可能的监督修改建议供阅核人参考

### 重要、复杂案件

提示预警法条引用、病句错字等低级纰漏，提供辅助工具提升阅核效率





# 应用场景4：司法智能助企

## 业务场景需求

- ◆呼应“能动履职”“营造一流化法制营商环境”的创新实践，解答企业法律问题、为企业提供法律体检
- ◆融合全国性法律法规和地方性法律法规，精准解答企业的法律问题，提供辖区法院的生效案例促进法律共识形成

## 技术特色

### 01 依托大模型的法律依据推理和问题解答

基于清华LegalAID 通专结合模型，融合全国、地方性法律法规和涉企相关政策规定，解答多种类型的法律问题

### 02 要素式匹配的精准法律文书解答

通过要素式拆解，初步了解纠纷特征，精准推送辖区内的生效案例（隐名方式），助力当事人形成法律共识

### 03 多端协同服务

“苏法小融”同步登陆苏周到、苏商通、苏服办、两级法院诉服大厅，多渠道服务当事人



苏周到



苏商通



苏服办





# 应用场景5：案件要素式改判发回预警

## 业务场景需求

- ◆ 改判发回案件带来案件数量和工作负荷的增加，影响当事人体验和审判质量指标
- ◆ 改判发回的情形随案由、要件的变化而变化，需要在审判生效前进行主动推送和提醒。

## 技术特色

### 01 基于大模型的要素拆解

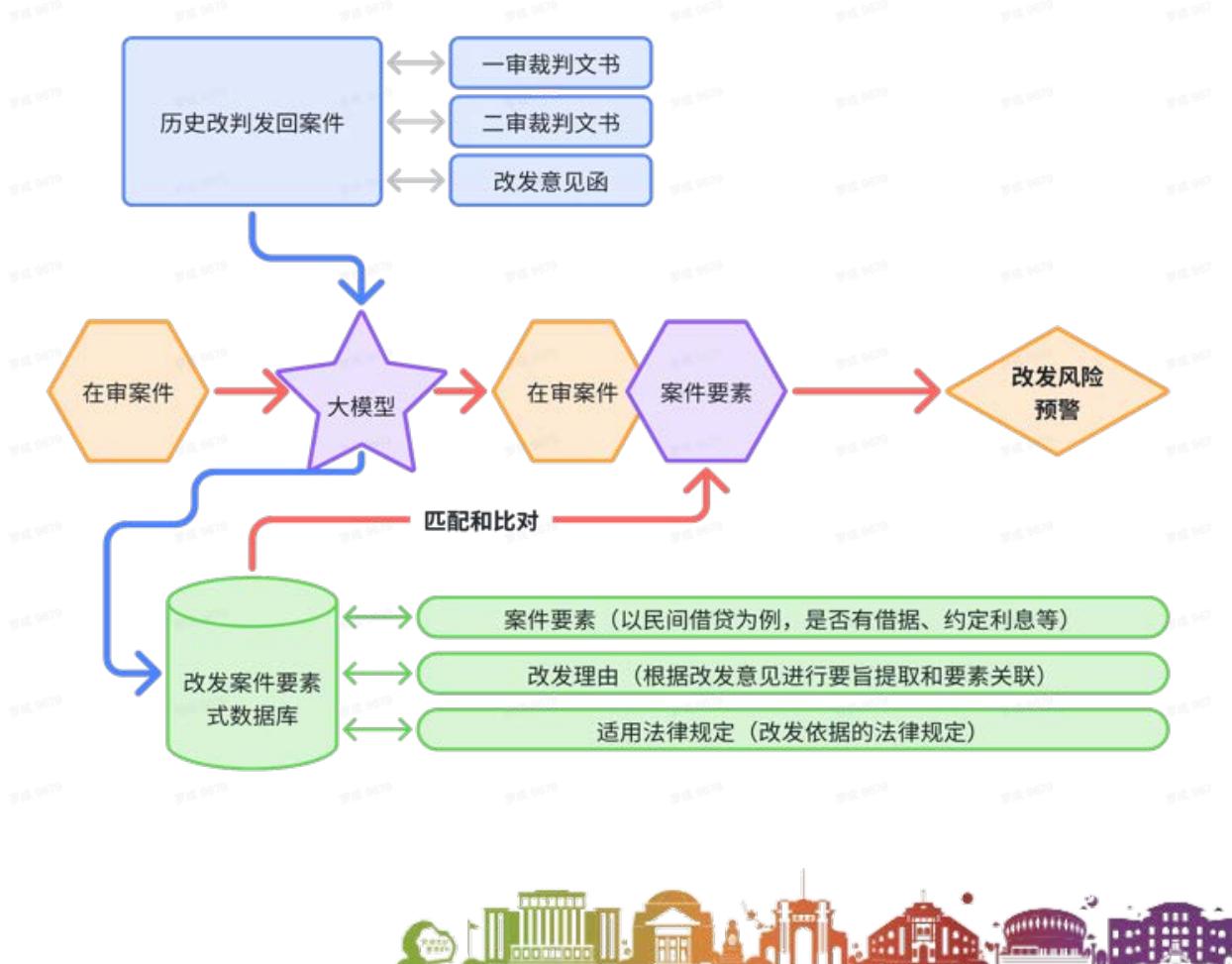
基于大模型和审判要点，从裁判文书等数据中抽取相关的要素，构建服务改发预警的要素数据库和法律依据库。

### 02 多模型融合的案件匹配

融合以案情找案、以文字找案、以要素找案等多种方式，从法律要素匹配的层面寻找改发案件历史数据，进行精准的匹配。

### 03 主动预警提醒

针对案件审理过程中形成的文档资料，随案抽取要素标签，对潜在的改判发回情形进行主动的抽取和预警。





# 应用场景6：法学研究范式革新

## 业务场景需求

- ◆以法学为代表的实践性学科，存在数据壁垒、体制差异和人才培养脱节问题，应用与研究亟需对齐
- ◆传统法学实证研究样本获取成本高、难度大、数量有限，难以及时、准确、全面反映法律运行状态



## 技术特色

### 01 从单一场景模拟到群体模拟

模拟控、辩、审不同诉讼构造的庭审场景，拓展多智能体的数量边界，实现数万级别的群体智能体模拟

### 02 并行推理与信息共享进化知识

通过智能体之间的协同工作、并行推理和信息共享，有效处理庞大且复杂的法律知识，构建静态和动态法学知识图谱，协助知识进化

### 03 分布式协同决策与法律政策模拟

模拟复杂的法律政策创设过程，根据不同利益方的博弈行为。通过分布式智能体之间的协同决策，测试法律政策实施的效果及社会影响，为立法研究提供动态分析框架

当事人	公检法司人员	法学研究场景
虚拟	真实	司法人员专业技能培训
真实	虚拟	法律咨询、诉前调解
虚拟	虚拟	立法效果沙盘推演
真实	真实	司法审判助手、律师助手





Information Retriever of Tsinghua University



感谢聆听  
请批评指导！

