

# 面向法律领域的大模型微调与应用

沈晨晨<sup>1</sup>, 岳圣斌<sup>2</sup>, 刘书隽<sup>1</sup>, 周宇轩<sup>1</sup>, 王思远<sup>3</sup>, 陈伟<sup>4</sup>, 萧尧<sup>5</sup>, 李秉轩<sup>1</sup>, 宋鋈<sup>6</sup>, 沈晓宇<sup>7</sup>, 黄萱菁<sup>8</sup>, 魏忠钰<sup>1</sup>

1. 复旦大学大数据学院, 上海 200433;
2. 复旦大学上海数学中心, 上海 200433;
3. 南加州大学INK实验室, 加利福尼亚州 洛杉矶 CA 90089-0915;
4. 华中科技大学软件学院, 湖北 武汉 430074;
5. 上海纽约大学文理学院, 上海 200124;
6. 西北政法大学法治学院, 陕西 西安 710122;
7. 宁波东方理工大学(暂名)数字孪生研究院, 浙江 宁波 315200;
8. 复旦大学计算机科学技术学院, 上海 200433

## 摘要

近年来, 大语言模型在多个自然语言处理任务上展现出了出色的能力, 为智慧法律系统的发展带来巨大的帮助。现有法律领域的大模型, 通过微调通用大模型能够实现利用法律知识进行简单的问题回答, 即大多以法律咨询问答为主, 没有考虑到法律领域的其他使用场景, 如法律信息抽取、判决预测等, 而真实世界中的法律服务要比对话服务复杂得多。提出中文法律智慧大模型LawLLM, 该模型可以面向不同用户群体, 提供多样的法律服务。同时, 探究了针对法律领域裁判文书的长文本信息抽取的应用。LawLLM在Lawbench上的Zero-shot的平均表现超过了所有对比的大模型, 均值比具有 $175 \times 10^9$ 个参数的GPT-3.5-Turbo高0.19%, LawLLM在Lawbench上的Few-shot的平均表现仅次于GPT-3.5-Turbo, 相比其低0.02%。

## 关键词

大语言模型; 智慧法律; 自然语言处理

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2024057

## *Fine-tuning and application of large language model in law domain*

SHEN Chenchen<sup>1</sup>, YUE Shengbin<sup>2</sup>, LIU Shujun<sup>1</sup>, ZHOU Yuxuan<sup>1</sup>, WANG Siyuan<sup>3</sup>, CHEN Wei<sup>4</sup>, XIAO Yao<sup>5</sup>, LI Bingxuan<sup>1</sup>, SONG Yun<sup>6</sup>, SHEN Xiaoyu<sup>7</sup>, HUANG Xuanjing<sup>8</sup>, WEI Zhongyu<sup>1</sup>

1. School of Data Science, Fudan University, Shanghai 200433, China
2. Shanghai Center for Mathematical Sciences, Fudan University, Shanghai 200433, China
3. INK Lab, University of Southern California, Los Angeles CA 90089-0915, America
4. School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
5. School of Arts and Science, New York University Shanghai, Shanghai 200124, China
6. Law School, Northwest University of Political and Law, Xi'an 710122, China
7. Institute of Digital Twin, Eastern Institute of Technology, Ningbo 315200, China
8. School of Computer Science, Fudan University, Shanghai 200433, China

## Abstract

In recent years, large language model have demonstrated outstanding capabilities across multiple natural language

processing tasks, greatly aiding the development of intelligent law systems. Existing large language model in the law domain, primarily focuses on fine-tuning foundation model for simple question answering tasks in legal consultation, have not considered other application scenarios in the legal field, such as legal information extraction, judgment prediction, etc. Real world law services are far more complex than dialogues. LawLLM was proposed, a Chinese law intelligent large language model, capable of providing diverse law services for different user groups. Additionally, the application of information extraction on long-context judicial documents in the law domain was explored. The average performance of LawLLM on Lawbench outperforms all the compared large language models, being 0.19% higher than the 175 billion parameter GPT-3.5-Turbo model. LawLLM's average performance on few-shot tasks on Lawbench is slightly lower than GPT-3.5-Turbo, by 0.02%.

### Key words

large language model, intelligent law, natural language processing

## 0 引言

随着法律人工智能的兴起,自动化的法律任务,例如法律信息抽取<sup>[1]</sup>、互动论点对提取<sup>[2-3]</sup>、类似案例检索<sup>[4]</sup>、法律问答<sup>[5]</sup>、判决预测<sup>[6-7]</sup>等任务,可以减轻法律从业者繁重的文书工作负担,提高工作效率和准确性。

由于法律知识的专业性和法律任务的多样性,在先前的智慧法律系统中,研究人员通常为每个特定的法律任务设计专门的算法或模型,面临研发成本高、集成难度大等困难。近年来,大语言模型(large language model, LLM, 以下简称大模型)在多个传统自然语言处理任务上展现出了出色的表现,为实现统一框架的智慧法律系统提供了技术支持、提高了可能性。

开发一个能同时处理多种法律任务的智慧法律系统具有非常重要的应用价值。一方面,它可以提高法律相关专业人士的工作效率。法律信息抽取、类似案例检索等任务是法官、律师的常见任务,它们涉及对大量文档或案例的分析,通常需要耗费大量的人力和时间来完成,智慧法律模型通过自动化完成这些过程,可以帮助专业

人士快速准确地获取所需信息,节省时间和精力。另一方面,它可以简化普通人获取法律服务的程序。通常情况下,人们在获取法律咨询服务时,需要先通过烦琐的流程来寻找合适的法律专业人员,然后再安排会面或电话咨询。智慧法律模型可以简化上述流程,直接为用户提供相关信息和建议。

除此之外,研究智慧法律模型的长文本处理能力也非常具有实际应用的价值。在实际应用中,法律文本(如裁判文书),通常是长文本甚至超长文本,法官等法律专业人士需要从这些文本中抽取和归类关键信息。尽管现有的大模型在各种自然语言处理任务中有非常好的表现,但通常仅限于处理上下文窗口大小范围内的文本。

## 1 研究现状

### 1.1 自然语言处理在智慧法律领域的应用

Zhong等<sup>[8]</sup>研究了自然语言处理(natural language processing, NLP)在智慧法律系统中的应用,讨论了智慧法律的历史、现况和未来的研究方向。该篇文章介绍了5种在司法领域的NLP任务和4个司法

数据集,任务包括司法信息抽取、司法要素识别、司法判决预测、司法类案检索、司法考试问答,并总结出了智慧法律系统应用的两大类NLP方法,分别是知识驱动的基于符号的方法(symbol-based-method)、数据驱动的基于词向量的方法(embedding-based-method)。

知识驱动的基于符号的方法主要用于关系抽取、事件时间线抽取、要素识别等任务;数据驱动的基于词向量的方法主要用于通用法律概念的词向量建模、法律知识图谱建模、预训练语言模型建模等。这两类NLP方法应用到法律文本摘要、类似案例匹配、法律知识咨询、罪名刑期预测等下游任务,对于每个任务需要分别设计和训练一个专有的模型。

## 1.2 法律大语言模型的研究现状

自OpenAI的ChatGPT<sup>[9]</sup>模型发布以来,研究人员开发了很多通用基座模型,在英文领域中能力较强的有LLaMa<sup>[10-11]</sup>系列,在中文领域中能力较强的有Baichuan系列<sup>[12]</sup>、ChatGLM系列<sup>[13-14]</sup>等。截至2024年3月,依然有新的基座模型被发布,这些基座模型基本以Transformer<sup>[15]</sup>结构为主,结合指令微调(instruction tuning)<sup>[16]</sup>,基于人类反馈的强化学习(reinforcement learning from human feedback, RLHF)<sup>[17]</sup>等技术,在通用领域具备了非常强大的能力。

考虑到领域内知识的专业性,基于这些基座模型,研究人员又开发了垂直领域的大模型,在医疗、金融、教育、法律领域都有涉及。现有法律领域的大模型,通过微调通用大模型能够利用法律知识进行简单的问题回答,即大多以法律咨询问答为主,没有考虑到法律领域的其他使用场景,如法律信息抽取、判决预测等,然而真实世界中的法律服务要比对话复杂得

多。现有法律领域的大模型有LawGPT、Lawyer-LLaMa<sup>[18]</sup>、LexiLaw、ChatLaw<sup>[19]</sup>等。LawGPT是一系列基于中文法律知识的开源大模型,该系列模型在通用中文基座模型(如Chinese-LLaMA<sup>[20]</sup>、ChatGLM等)的基础上做法律领域专有词表扩充、大规模中文法律语料预训练,并在此基础上,构造法律领域对话问答数据集、中国司法考试数据集进行指令精调,提升模型对法律内容的理解和执行能力。Lawyer LLaMA在大规模法律语料上进行了继续预训练(continual pretraining),并借助ChatGPT收集了一批对中国国家统一法律职业资格考试客观题的分析和对法律咨询的回答,利用收集到的数据对模型进行指令微调,让模型习得将法律知识应用到具体场景中的能力。LexiLaw基于ChatGLM-6B架构,通过在法律领域的数据集上进行微调,使其在提供法律咨询和支持方面具备更高的性能和专业性。ChatLaw法律大模型开源版本的底座基于姜子牙-13B、Anima-33B,使用了大量法律新闻、法律论坛、法条、司法解释、法律咨询、法考题、判决文书等原始文本来构造对话数据。

## 2 中文智慧法律大模型LawLLM

### 2.1 模型框架及训练方法

中文智慧法律大模型LawLLM的模型整体框架如图1所示,该模型具有3个核心能力。

#### (1) 基础的法律文本处理能力

指法律领域的文本理解和文本生成的不同基础处理能力,包括信息抽取、文本摘要等,该部分能力主要来源于基于NLP司法任务的公开数据集和真实世界的法律

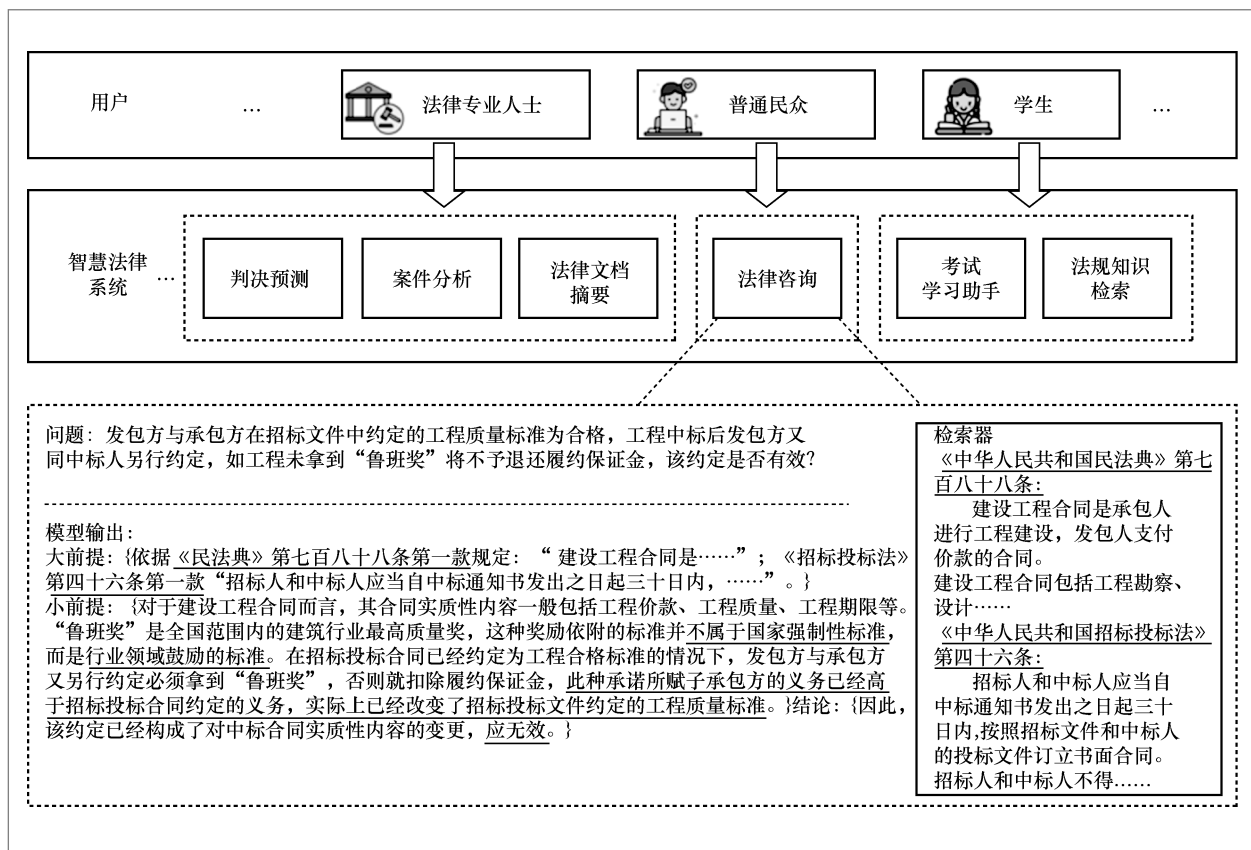


图1 中文智慧法律大模型 LawLLM 的模型整体框架

相关文本构建的指令微调数据集。

### (2) 法律推理思维能力

本文使用法律三段论的基本推理过程重构指令数据集，提高模型的法律推理能力。

### (3) 司法领域知识检索能力

为加强模型对与输入问题相关的背景法条或案例的知识检索能力，增加检索增强模块。

为了构建具有推理能力和检索能力的智慧法律大模型 LawLLM，模型训练被分为监督微调 (supervised fine-tuning, SFT) 和检索增强两个阶段。

监督微调阶段的训练目标和自回归模型一样，即根据已知文本预测下一个 Token。经过监督微调，模型能够具备基础的法律文本处理能力，包括法律语言的理解和

生成能力，并具备法律推理思维。但在很多法律场景下，例如法律咨询和判决预测场景，模型的输出如果能有法律法规作为支持依据则更有说服力。同时在监督微调之后，模型可能会因为幻觉或过时的知识而产生不准确的输出。为了解决这个问题，检索增强阶段利用开源的检索框架 Langchain 来增强模型的能力。首先建立包括中国宪法、刑法、行政诉讼法、著作权法、专利法等 50 多类法律的知识库，将这些文档编码为向量并保存在本地知识库。检索过程中，对于每一个用户输入，检索器计算输入与知识库中文本块 (chunk) 的相似性，从知识库中召回最相关的 Top-K 个文本块，候选文档和原始用户输入经过设计的提示 (prompt) 组合后，再输入模型中得到输出。图2为该过程的示意图。

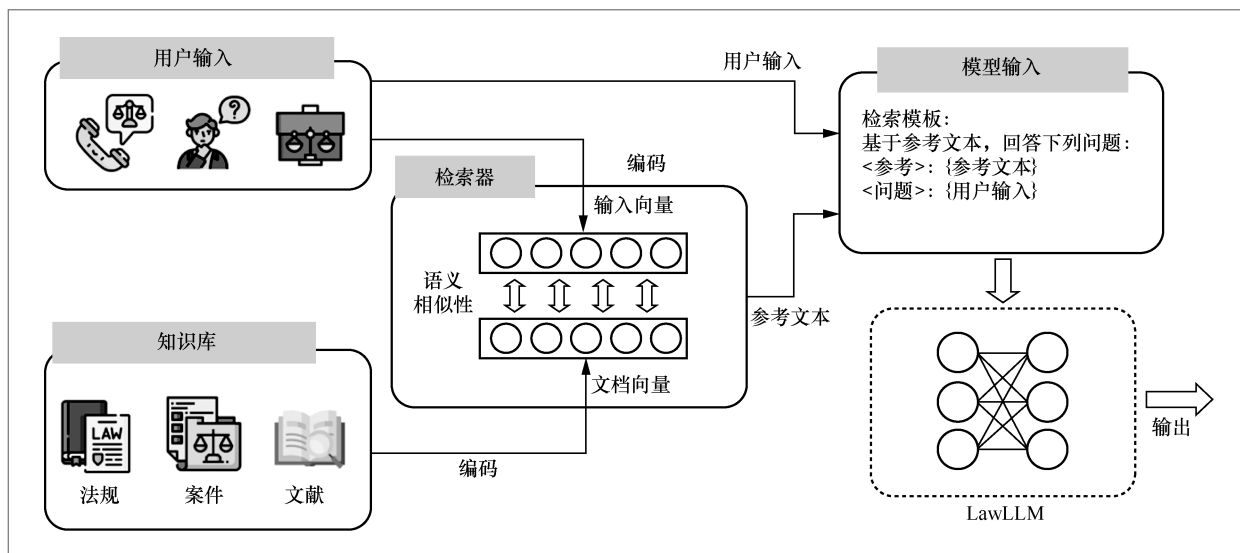


图2 检索增强过程示意图

整个训练过程基于Baichuan-13B-Base模型做全参数微调, Baichuan-13B-Base模型是在约1.4万亿Tokens的高质量预料上训练的130亿参数基模型。训练过程的超参数设置如下: 批大小(batch size)为64、学习率为 $5 \times 10^{-5}$ 、2个epoch训练阶段、上下文长度为4 096个Token、训练过程在8\*A800GPU上进行。

## 2.2 数据集构建

法律智慧大模型需要具备不同法律场景下的文本理解、文本生成能力, 因此在构建数据集时考虑了法律信息抽取、法律咨询问答、法律文档摘要等不同场景或任务。构建数据集的数据主要来自三大部分: 一是公开的NLP法律任务数据集; 二是来自真实世界的原始法律文本; 三是开源的指令微调数据集。

数据构建采用下面3种方法对数据进行重构。

### (1) 行为塑造

在法律判决三段论中, 法官的基本推理过程是: 大前提(major premise)是适

用的法律, 小前提(minor premise)是相关事实, 结论(conclusion)是最终判决。即每个案例都可以用法律三段论得出一个明确的结论, 形式如下。

- 大前提: 适用法律法条。
- 小前提: 案件事实。
- 结论: 判决结果。

受法律三段论提示(legal syllogism prompting)<sup>[21]</sup>和自构造(self-construct)<sup>[22]</sup>的启发, 利用GPT-3.5-Turbo模型来对数据按照上面的格式进行改写, 完成行为塑造的重构, 确保“<输入-输出>”指令对数据的输出是通过法条和案件事实推论得到的。

### (2) 知识扩展

这类重构方法针对选择题数据集, 由于许多与法律相关的考试数据集的输出仅提供选项答案, 用GPT-3.5-Turbo模型来扩展输出中的法律知识, 以提供更多的推理细节。

### (3) 思维发展

已有多项研究证明思维链(chain of thought, CoT)能够有效地提高模型的推理能力<sup>[23]</sup>。为进一步赋予模型法律推

理的能力,参考思维链方法设计法律思维链(law-specific chains of thought, LCoT)来要求模型按照法律三段论推理答案。LCoT将按下面的形式转换输入 $X$ (LCoT prompt)。

- 在法律三段论中,大前提是适用的法律法规,小前提是案件事实,结论是对案件的法律判断。

- 案件:  $X$ 。

- 用法律三段论来思考和输出判断。

通过以上3种方法对数据进行重构后,数据集的总体规模约为403 000条,其中指令对数据约为256 000条,指令三元组数据约为39 000条,通用指令数据约为108 000条,表1是LawLLM指令微调数据集的详细统计。

## 2.3 评测方案

LawBench是一个基于中国法律体系的开源大语言模型综合评估基准(Benchmark)<sup>[24]</sup>,LawBench数据集包括了20个不同的任务,涵盖了记忆、理解、应用3个认知水平。

- 记忆:大语言模型能否记住法律概念、法条、术语等。

- 理解:大语言模型能否理解法律文本中的意义和内涵。

- 应用:大语言模型能否正确利用其知识和推理能力并完成下游的现实法律任务。

LawBench的20个不同任务每项都有500个示例,不同任务的评价指标有一定的差异。表2是Lawbench任务列表。

## 2.4 实验结果

本次实验除了笔者提出的智慧法律大模型LawLLM,还对比了其他通用大模型

表1 LawLLM 指令微调数据集详细统计

数据集	任务	规模/条	场景
指令对数据	司法要素抽取	32 000	专业工具
	司法事件检测	27 000	
	司法案件分类	20 000	
	判决预测	11 000	
	类案匹配	8 000	
	司法文本摘要	9 000	
指令三元组数据	涉法舆情摘要	6 000	
	司法问答	93 000	法律咨询
	司法阅读理解	38 000	法考助手
	司法考试	12 000	
通用数据	Alpaca-GPT4	48 000	通用场景
	Firefly	60 000	
总量		403 000	

以及法律领域大模型,各个模型如下。

- 通用大模型:GPT-3.5-Turbo、Llama2-13 B-Chat、Baichuan-13 B-Chat、Chinese-Alpaca2-13 B。

- 法律领域大模型:LaWGPT、Lawyer-LLaMa、ChatLaw、LexiLaw。

图3是在Lawbench上的Zero-shot、Few-shot评测实验结果,从图3中可以看到,本文的模型LawLLM在Lawbench上的Zero-shot的平均表现超过了所有对比的大模型,均值比具有 $175 \times 10^9$ 个参数的GPT-3.5-Turbo高0.19%。LawLLM在Lawbench上的Few-shot的平均表现仅次于GPT-3.5-Turbo,相比其低0.02%。

## 2.5 模型Demo服务情况

模型Demo发布后,截至2024年3月底,已有超过两万人次使用中文智慧法律

表2 Lawbench 任务列表

认知水平	任务ID	任务	指标	类型
法律知识记忆	1-1	法条背诵	Rouge-L	生成
	1-2	知识问答	Accuracy	单选
法律知识理解	2-1	文件校对	F0.5	生成
	2-2	纠纷焦点识别	F1	多选
	2-3	婚姻纠纷鉴定	F1	多选
	2-4	问题主题识别	Accuracy	单选
	2-5	阅读理解	RC-F1	抽取
	2-6	命名实体识别	Soft-F1	抽取
	2-7	舆情摘要	Rouge-L	生成
	2-8	论点挖掘	Accuracy	单选
	2-9	事件检测	F1	多选
	2-10	触发词提取	Soft-F1	抽取
法律知识应用	3-1	法条预测(基于事实)	F1	多选
	3-2	法条预测(基于场景)	Rouge-L	生成
	3-3	罪名预测	F1	多选
	3-4	刑期预测(无法条内容)	Normalizedlog-distance	回归
	3-5	刑期预测(给定法条内容)	Normalizedlog-distance	回归
	3-6	案例分析	Accuracy	单选
	3-7	犯罪金额计算	Accuracy	回归
	3-8	咨询	Rouge-L	生成

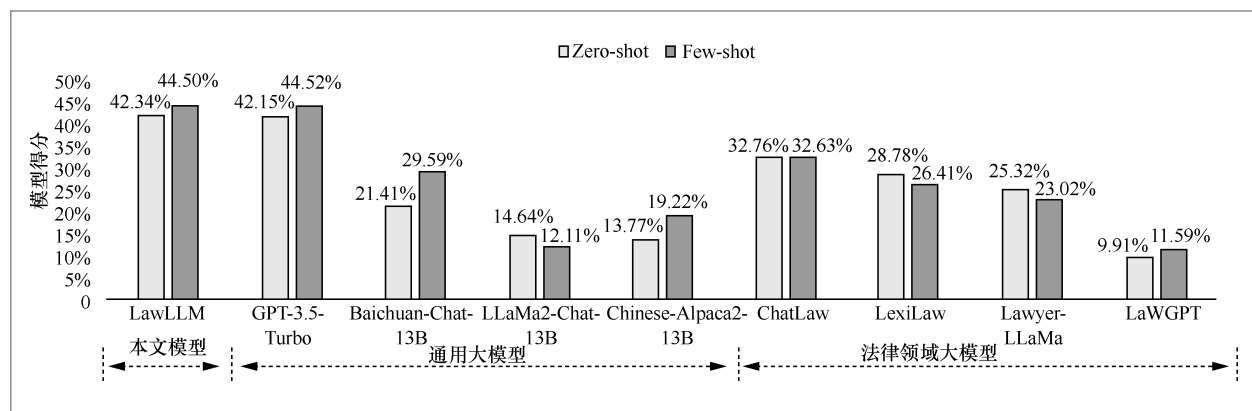


图3 在Lawbench上的Zero-shot、Few-shot评测实验结果

大模型LawLLM, 对话交互超过4万次, 反响良好。排名位于前三的用户咨询问题分别为民间借贷纠纷类、劳动人事争议类、离婚财产分割类问题, 表3是排名位于前三的用户咨询问题对应的部分问题。表4是模型在实际使用过程中的部分对话实

例, 从表4中可以看出, 模型能出色地完成众多法律领域的任务, 不限于常规问答, 对实际应用更有帮助。但模型也有不足之处, 主要体现在模型的多轮对话能力不强, 这是由于训练过程中使用的是单轮对话数据, 图4所示是模型多轮对话实例。

表3 排名位于前三的用户咨询问题对应的部分问题

问题类型	对应具体问题
民间借贷纠纷类问题	① 请问,公司A欠我司款20万元,我们应该用什么法条将20万元追回? ② 我是债权人,债务人去世了,我该如何拿回我的出借款? ③ 起草一个借款起诉状。
劳动人事争议类问题	① 公司经营困难,提前一个月口头通知裁员,员工能获得什么补偿? ② 公司强迫加班如何维权? ③ 劳动者拒签劳动合同的,用人单位是否需要承担未签订劳动合同二倍工资责任?
离婚财产分割类问题	① 写一篇离婚财产分割协议。 ② 离婚诉讼中双方对财产价值有争议时,是否必须委托中介机构评估? ③ 婚后分居,离婚的时候,一方的个人收入属于共同财产吗?

表4 模型在实际使用过程中的部分对话实例

示例编号	具体对话内容
示例1	输入:哥哥不照顾父母,遗产可不可以少分一点? 模型回复:可以。根据《民法典》第一千一百三十条规定:同一顺序继承人继承遗产的份额,一般应当均等。对生活有特殊困难又缺乏劳动能力的继承人,分配遗产时,应当予以照顾。对被继承人尽了主要扶养义务或者与被继承人共同生活的继承人,分配遗产时,可以多分。有扶养能力和有扶养条件的继承人,不尽扶养义务的,分配遗产时,应当不分或者少分。继承人协商同意的,也可以不均等。
示例2	输入:写一份离婚财产分割协议。 模型回复:男方:____,身份证号码:____住址:____女方:____,身份证号码:____住址:____男女双方于__年__月__日登记结婚,于__年__月__日生儿子。现因夫妻感情破裂,已无和好可能,在平等、自愿的基础上,经双方共同协商,达成协议如下:一、双方自愿离婚。二、孩子由男方监护抚养,女方支付抚养费____元/月,一年一付,每年__月__日前支付,上述款项包括学费、生活费等一切费用。三、夫妻共同财产按以下分割:1.男方一次性支付女方____万元(该款包括但不限于女方拆迁补偿、安置、过渡、奖励等费用),在双方办理离婚手续后当日付清;2.车辆归男方所有;3.现有家具、家电等家庭用品全部归__所有。四、双方无共同债权、债务。婚姻关系存续期间个人债务由个人承担,与对方无涉。五、女方对儿子享有探视权,男方应积极配合,原则上每月__次,具体时间、方式双方另行协商。六、本协议一式3份,男女双方各执一份,婚姻登记机关存档一份。男方:____女方:____年__月__日
示例3	输入:王女士在某公司工作已满4年,在怀孕5个月时告知公司人事并申请公司配合办理生育保险申请等事项,孕期工作表现均符合公司公示制度要求。两周后,公司借口以王女士无法胜任工作为由要与其解除劳动合同关系,王女士如何维护自己的合法权益? 模型回复:根据《劳动法》的规定,王女士在怀孕5个月时告知公司并申请了公司配合办理生育保险申请等事项,孕期工作表现均符合公司公示制度要求,属于合法的。根据《女职工劳动保护特别规定》的规定,用人单位不得因女职工怀孕、生育、哺乳降低其工资、予以辞退、与其解除劳动合同或者聘用合同,因此公司不得以王女士无法胜任工作为由与其解除劳动合同关系。因此,如果公司以此为理由要与王女士解除劳动合同关系,王女士有权维护自己的合法权益,包括要求公司按照法定程序办理生育保险申请等事项,并可向劳动仲裁机构提起劳动仲裁,要求公司支付相应的经济补偿或其他合法赔偿。

### 3 法律场景下的长文本信息抽取

#### 3.1 任务介绍

裁判文书原文中通常包括文书基本信息、人物信息、案件信息、判决信息、依据法条等。文书基本信息包含审理法院、文书类

型等;人物信息包含公诉机关名称(刑事)、刑事被告信息(刑事)等;案件信息包含案件审理阶段、案由等;判决信息包含裁判日期、本院认为、争议焦点等;依据法条包含法律名称及法律条款。图5所示是裁判文书信息抽取任务示例,左侧为裁判文书原文,右侧为抽取结果。在现实场景下,裁判文书的平均长度为1万字以上,这远远超出大部分中文基座模型的处理长度(4 096字)。

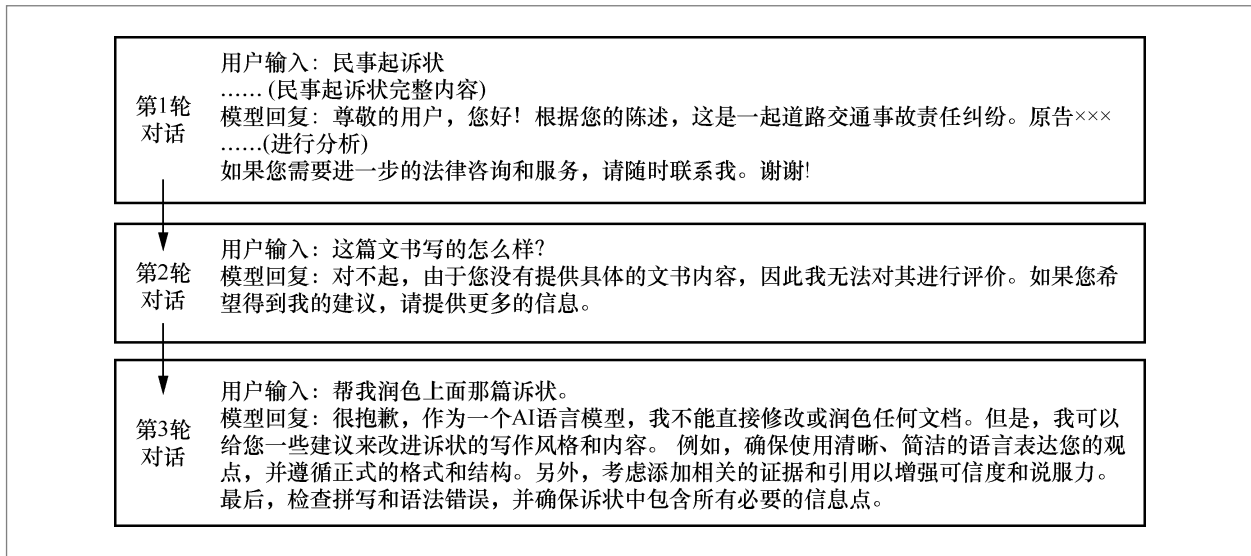


图 4 模型多轮对话实例

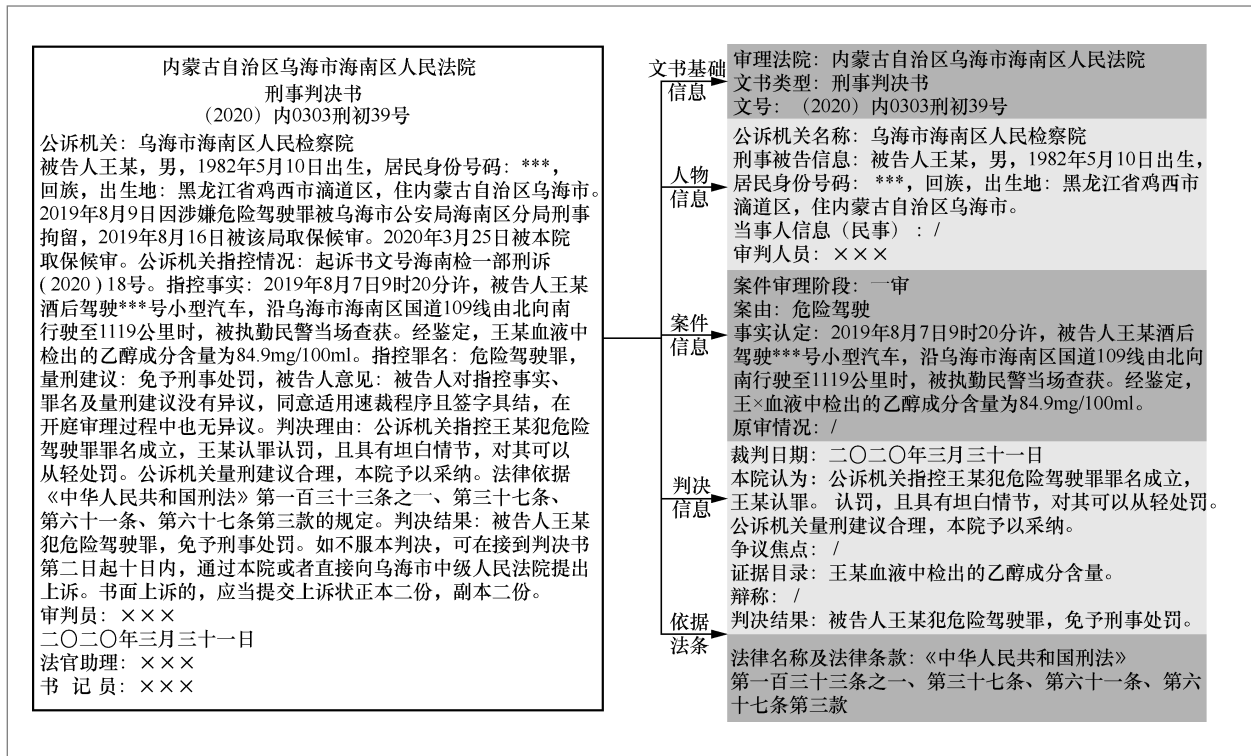


图 5 裁判文书信息抽取任务示例

### 3.2 实验方案

法律场景下的长文本信息抽取应用主要针对裁判文书，评测集共有320篇裁判

文书，文书总长度在0~4 000字、4 000~8 000字、8 000~16 000字、16 000字以上的各有80篇，每篇裁判文书均抽取具体项，共5 760个测试问题，裁判文书信息抽取项的类别、问题ID及对应具体项见表5。

裁判文书均来自中国裁判文书网,该网站提供了不同年份、不同法院、不同审理阶段的民事和刑事裁判文书,并不断更新。

实验对比模型有LawLLM、Chat-GLM2-6B-32K、ChatGLM3-6B-32K、GPT-3.5-Turbo-0125。LawLLM是本文提出的中文智慧法律大模型,该模型未在裁判文书信息抽取任务上进行微调,可以处理的上下文长度为4 000字。ChatGLM2-6B经过了1.4 TB中英标识符的预训练与人类偏好对齐训练。ChatGLM3-6B的基础模型采用了更多样的训练数据、更充分的训练步数和更合理的训练策略。ChatGLM2-6B-32K、ChatGLM3-6B-32K分别在Chat-GLM2-6B、ChatGLM3-6B的基础上进一步强化了对长文本的理解能力,能够更好地处理多达32 000字长度的上下文。Long-Bench<sup>[25]</sup>是一个多任务、中英双语、针对大语言模型长文本理解能力的评测基准,Long-Bench的测评结果表明,在等量级的开源模型中,ChatGLM2-6B-32K和ChatGLM3-6B-32K在中文长文本处理上有较明显的竞争优势。GPT-3.5-Turbo-0125具备16 000字上下文长度的处理能力。

对预测文本和参考文本分词处理后,基于Rouge-L评价指标计算它们之间的相似度分数,关注预测文本和参考文本之间的重叠程度。

在测试过程中,对于超出模型处理长度能力的文本,从文本中间进行截断,保留前后部分的信息。该设置参考了Liu等<sup>[26]</sup>的研究,其实验结果表明,这种截断方式对模型性能影响最小。

### 3.3 实验结果

(1)不同长度文本上的实验结果及分析

表6是裁判文书信息抽取任务在不同长度文本上的总体实验结果,加粗为最

表5 裁判文书信息抽取项的类别、问题ID及对应具体项

类别	问题ID	具体项
文书基础信息	1-1	审理法院
	1-2	文书类型
	1-3	文号
人物信息	2-1	公诉机关名称(刑事)
	2-2	刑事被告信息(刑事)
	2-3	当事人信息(民事)
	2-4	审判人员
案件信息	3-1	案件审理阶段
	3-2	案由
	3-3	事实认定
	3-4	原审情况
判决信息	4-1	裁判日期
	4-2	本院认为
	4-3	争议焦点
	4-4	证据目录
	4-5	辩称
	4-6	判决结果
依据法条	5-1	法律名称及法律条款

佳结果,加下划线为次佳结果。从表6中可以看到, LawLLM的平均结果仅次于 $175 \times 10^9$ 个参数、16 000字上下文长度的GPT-3.5-Turbo-0125模型。在不同上下文长度上的效果几乎都超过了32 000字上下文长度的ChatGLM系列,仅在4 000~8 000字长度上比ChatGLM3-6B-32k低0.87%。LawLLM未在裁判文书信息抽取任务上进行微调,且可以处理的上下文长度仅为4 000字,由此可见,其他法律类任务的微调使模型具备了完成新的法律任务的能力,若其长度突破4 000字,可能在裁判文书这类长文本上具备更强的处理能力。

表6 裁判文书信息抽取任务在不同长度文本上的实验结果

模型	参数量/Billion	上下文长度/字	0~4 000字	4 000~8 000字	8 000~16 000字	16 000字以上	平均值
ChatGLM2-6B-32K	6	32 000	28.16	24.74	23.11	20.96	24.33
ChatGLM3-6B-32K	6	32 000	33.06	<u>29.70</u>	26.14	24.41	28.43
GPT-3.5-Turbo-0125	175	16 000	<b>47.14</b>	<b>42.78</b>	<b>40.62</b>	<b>39.26</b>	<b>42.53</b>
LawLLM	13	4 000	<u>36.67</u>	28.83	<u>26.54</u>	<u>25.21</u>	<u>29.40</u>

(2)不同抽取问题上的实验结果及分析

表7是裁判文书信息抽取任务在不同抽取问题上的实验结果,加粗为最佳结果,加下划线为次佳结果。从表7中可以看到, LawLLM在18类问题中,有2项问题取得了最佳的结果,6项问题取得了次佳的结

表7 裁判文书信息抽取任务在不同抽取问题上的实验结果

问题ID	ChatGLM2-6B-32K	ChatGLM3-6B-32K	GPT-3.5-Turbo-0125	LawLLM
平均	24.33	28.43	<b>42.53</b>	<u>29.4</u>
1-1	60.99	36	<b>67.39</b>	<u>62.49</u>
1-2	31.02	34.28	<u>62.06</u>	<b>63.57</b>
1-3	38.47	<u>46.88</u>	<b>78.48</b>	38.93
2-1	<b>33.96</b>	14.1	<u>32.89</u>	30.02
2-2	48.11	43.66	<b>69.44</b>	<u>69.37</u>
2-3	14.2	<u>28.27</u>	<b>44.19</b>	12.7
2-4	7.66	<u>12.22</u>	<b>26.04</b>	6.86
3-1	<u>48.58</u>	41.35	<b>61.2</b>	37.28
3-2	28.2	23.85	<u>57.97</u>	<b>59.28</b>
3-3	15.64	<b>26.88</b>	<u>24.13</u>	13.32
3-4	7.81	<b>20.34</b>	<u>10.88</u>	8.73
4-1	22.46	23.41	<b>74.03</b>	<u>28.27</u>
4-2	7.11	<b>62.24</b>	<u>32.03</u>	21.01
4-3	<b>17.39</b>	10.91	12.02	<u>15.2</u>
4-4	8.07	<b>18.54</b>	<u>15.35</u>	6.94
4-5	18.82	<b>30.12</b>	<u>29.21</u>	15.74
4-6	6.47	<u>23.15</u>	<b>25.47</b>	12.4
5-1	23.05	15.49	<b>42.79</b>	<u>27.11</u>

果,这6项均次于GPT-3.5-Turbo-0125模型。LawLLM表现较差的问题是证据目录,在问题4-4上取得最佳结果的模型是ChatGLM3-6B-32K,可能是因为这项问题不仅需要模型有抽取式阅读理解的能力,还需要具备对抽取到的内容总结、归纳的能力,本文构建的指令微调数据集中忽略了这类能力的训练,导致了模型对通用能力的遗忘。

### 3.4 未来研究思路

受Transformer结构的限制,大模型在训练阶段直接训练长文本需要消耗大量计算资源,效果往往也欠佳,主要原因如下。

- 模型训练: GPU显存的占用与序列长度的平方成正比,使训练量急剧上升。
- 模型结构: 序列越长,模型的Attention越分散,模型越容易忘记前序内容。
- 推理速度: 模型序列增长,将大幅度降低模型推理速度。

Huang等<sup>[27]</sup>研究了基于Transformer结构的大模型处理长文本的方法,包括高效注意力机制、长期记忆、外推位置编码、上下文处理等方法,具体如下。

#### (1) 高效注意力

这类方法致力于实现具有降低计算需求能力的高效注意力机制,甚至实现线性复杂度<sup>[28-30]</sup>。注意力机制是Transformer

计算瓶颈的核心模块,优化注意力机制的方法在预训练阶段直接增加超参数,从而扩展大模型在推理过程中有效上下文长度的边界。

#### (2) 长期记忆

大模型缺乏显式的记忆机制,仅依靠KV缓存(key-value cache)来存储序列中所有先前Token的表示。这种设计意味着一旦在一个调用中完成了查询(query),Transformer在后续调用中将不会保留或记起任何先前的状态或序列,除非整个历史记录被逐个Token地重新加载到KV缓存中。为了突破上下文工作记忆的局限性,一些方法旨在设计显式的记忆机制,针对KV缓存进行优化<sup>[31-32]</sup>,弥补大模型中缺乏高效和有效的长期记忆的不足。

#### (3) 改进位置编码

通过改进现有位置编码方案的外推性能,提高大模型的长度扩展能力。最初的Transformer架构提出了正弦位置编码(sinusoidal PE, SinPE),通过优化位置编码来扩展上下长度,常见的有外推和内插<sup>[33]</sup>两种方式。外推将原来的编码方法直接应用到更大的位置上,内插将目标位置等比例放缩到模型支持的位置处。

#### (4) 上下文处理

除了针对Transformer模块的修改方法,有一些方法将现成的大模型与额外的上下文预处理/后处理相结合<sup>[34-35]</sup>,将预训练的大模型视为黑盒或灰盒模型,通过对模型进行多次调用来处理超出模型长度限制的上下文。这些方法满足了每次调用时实际输入始终满足最大长度的要求,并通过引入多次调用开销,打破上下文的窗口限制。

考虑到高效注意力机制和长期记忆这两种方法涉及模型的预训练,而预训练过程通常会比微调消耗更多的计算资源和训练时间,未来LawLLM处理法律场景下的

长文本信息抽取可以从改进位置编码和上下文处理这两种方法着手做深入研究。

## 4 结束语

本文的工作对智慧司法有实际的应用价值,除法律咨询场景以外,还考虑到了很多不同的法律任务,应用到实际中可以惠及不同的使用群体,能为实现法律资源的平衡带来更多的可能性,为法律行业带来全新的发展机遇和技术解决方案。本文的训练数据和训练方法可以被适配到任何底座大模型上,研究法律领域的指令微调工作也能被迁移到其他垂直领域,如金融、教育等。目前的工作还有不少可以优化的地方,可以从以下3个方面进行改进。

- 本文训练的中文法律智慧大模型使用的是单轮对话训练,模型没有利用历史对话信息的能力。在实际的应用过程中,尤其是在法律咨询场景下,多轮对话能力可以提升性能,更贴近实际需求。

- 本文的应用部分着重研究裁判文书这类长文本的处理,而案例也是智慧法律应用过程中非常重要的文本。针对这类文本任务,如类案匹配,也有两大难点需要解决,一是案例大多涉及长文本的处理,二是类案匹配任务需要建立一个完整的案例库。

- 在长文本信息处理中,未来将深入研究一些不需要训练就能实现低资源消耗的方法,如改进位置编码、上下文处理等。

## 参考文献:

- [1] BOMMARITO M J, KATZ D M, DETTERMAN E M. LexNLP: natural language processing and information extraction for legal and regulatory texts[EB]. arXiv preprint, 2018, arXiv:

- 1806.03688.
- [2] YUAN J, WEI Z Y, GAO Y X, et al. Overview of SMP-CAIL2020-argmine: the interactive argument-pair extraction in judgement document challenge[J]. *Data Intelligence*, 2021, 3(2): 287-307.
- [3] YUAN J, CHENG L Y, HE R D, et al. Overview of argumentative text understanding for AI debater challenge[M]. Cham: Springer, 2021: 548-568.
- [4] MA Y X, SHAO Y Q, WU Y Y, et al. LeCaRD: a legal case retrieval dataset for chinese law system[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 2342-2348.
- [5] KIEN P M, NGUYEN H T, BACH N X, et al. Answering legal questions by learning neural attentive text representation[C]//Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg, USA: International Committee on Computational Linguistics, 2020: 988-998.
- [6] 张虎, 张振, 范越, 等. 基于因果图分析的可解释司法判决预测方法研究[J]. *大数据*, 2024, 10(2): 109-121.
- ZHANG H, ZHANG Z, FAN Y, et al. Research on interpretable legal judgment prediction method based on causal graph analysis[J]. *Big Data Research*, 2024, 10(2): 109-121.
- [7] 孙倩, 秦永彬, 黄瑞章, 等. 结合案件要素序列的罪名预测方法[J]. *大数据*, 2021, 7(6): 30-40.
- SUN Q, QIN Y B, HUANG R Z, et al. Charge prediction method combined with case elements sequence[J]. *Big Data Research*, 2021, 7(6): 30-40.
- [8] ZHONG H X, XIAO C J, TU C C, et al. How does NLP benefit legal system: a summary of legal artificial intelligence[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020.
- [9] ZHANG K P. ChatGPT: Optimizing Language Models for Dialogue[M]. USA: University of Maryland, 2022.
- [10] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[EB]. arXiv preprint, 2023, arXiv: 2302.13971.
- [11] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models[J]. arXiv preprint arXiv: 2307.09288.
- [12] YANG A Y, XIAO B, WANG B N, et al. Baichuan 2: open large-scale language models[EB]. arXiv preprint, 2023, arXiv: 2309.10305.
- [13] DU Z X, QIAN Y J, LIU X, et al. GLM: general language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, USA: Association for Computational Linguistics, 2022.
- [14] ZENG A H, LIU X, DU Z X, et al. GLM-130B: an open bilingual pre-trained model[EB]. arXiv preprint, 2022, arXiv: 2210.02414.
- [15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. *Advances in Neural Information Processing Systems*, 2017, 30.
- [16] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[EB]. arXiv preprint, 2021, arXiv: 2109.01652.
- [17] OUYANG L, WU J, XU J, et al. Training language models to follow instructions with human feedback[EB]. arXiv preprint, 2022, arXiv: 2203.0215.
- [18] HUANG Q Z, TAO M X, ZHANG C, et al. Lawyer LLaMA technical report[EB]. arXiv preprint, 2023, arXiv: 2305.15062.
- [19] CUI J X, NING M N, LI Z J, et al. Chatlaw: a multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language

- model[EB]. arXiv preprint, 2023, arXiv: 2306.16092.
- [20] CUI Y M, YANG Z Q, YAO X, et al. Efficient and effective text encoding for Chinese LLaMA and alpaca[EB]. arXiv preprint, 2023, arXiv: 2304.08177.
- [21] JIANG C, YANG X L. Legal syllogism prompting: teaching large language models for legal judgment prediction[C]// Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. New York: ACM, 2023: 417-421.
- [22] WANG Y Z, KORDI Y, MISHRA S, et al. Self-instruct: aligning language models with self-generated instructions[EB]. arXiv preprint, 2022, arxiv: 2212.10560.
- [23] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB]. arXiv preprint, 2022, arXiv: 2201.11903.
- [24] FEI Z W, SHEN X Y, ZHU D W, et al. LawBench: benchmarking legal knowledge of large language models[EB]. arXiv preprint, 2023, arxiv: 2309.16289.
- [25] BAI Y S, LYU X, ZHANG J J, et al. LongBench: a bilingual, multitask benchmark for long context understanding[EB]. arXiv preprint, 2023, arXiv: 2308.14508.
- [26] LIU N F, LIN K, HEWITT J, et al. Lost in the middle: how language models use long contexts[EB]. arXiv preprint, 2023, arxiv: 2307.03172.
- [27] HUANG Y P, XU J W, LAI J Y, et al. Advancing transformer architecture in long-context large language models: a comprehensive survey[EB]. arXiv preprint, 2023, arXiv.: 2311.12351.
- [28] CHEN Y K, QIAN S J, TANG H T, et al. LongLoRA: efficient fine-tuning of long-context large language models[EB]. arXiv preprint, 2023, arXiv: 2309.12307.
- [29] BELTAGY I, PETERS M E, COHAN A. Longformer: the long-document transformer[EB]. arXiv preprint, 2020, arXiv: 2004.05150.
- [30] XIAO G X, TIAN Y D, CHEN B D, et al. Efficient streaming language models with attention sinks[EB]. arXiv preprint, 2023, arXiv: 2309.17453.
- [31] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context[EB]. arXiv preprint, 2019, arXiv: 1901.02860.
- [32] DING S Y, SHANG J Y, WANG S H, et al. ERNIE-doc: a retrospective long-document modeling transformer[EB]. arXiv preprint, 2020, arXiv: 2012.15688.
- [33] CHEN S Y, WONG S, CHEN L J, et al. Extending context window of large language models via positional interpolation[EB]. arXiv preprint, 2023, arXiv: 2306.15595.
- [34] RATNER N, LEVINE Y, BELINKOV Y, et al. Parallel context windows for large language models[EB]. arXiv preprint, 2022, arXiv: 2212.10947.
- [35] HAO Y R, SUN Y T, DONG L, et al. Structured prompting: scaling In-context learning to 1, 000 examples[EB]. arXiv preprint, 2022, arXiv: 2212.06713.

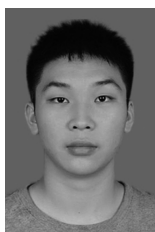
## 作者简介



沈晨晨(1995-),女,复旦大学大数据学院硕士生,主要研究方向为论辩挖掘、领域大模型微调。



岳圣斌(1998- ),男,复旦大学上海数学中心博士生,主要研究方向为多媒体内容分析、自然语言处理、大语言模型等。



刘书隽(2002- ),男,复旦大学大数据学院本科生,主要研究领域为大模型微调与对齐。



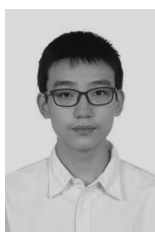
周宇轩(2001- ),男,复旦大学大数据学院硕士生,主要研究方向为自然语言处理、大语言模型。



王思远(1997- ),女,博士,南加州大学INK实验室研究助理,主要研究方向为自然语言处理、大模型。



陈伟(1994- ),男,博士,华中科技大学软件学院助理教授,主要研究方向为大语言模型与多模态智能。



萧尧(2002- ),男,上海纽约大学文理学院本科生,主要研究方向为机器学习与系统。



**李秉轩** (1999- ), 男, 硕士毕业于复旦大学大数据学院, 主要研究方向为自然语言处理、文本生成, 硕士期间参与本文工作, 现就职于百度(中国)有限公司。



**宋莹** (1986- ), 女, 博士, 西北政法大学法治学院讲师, 主要研究方向为中国古代法制史、社会治理。



**沈晓宇** (1994- ), 男, 博士, 宁波东方理工大学(暂名)数字孪生研究院助理教授, 主要研究方向为隐变量贝叶斯模型、文本生成、问答、大语言模型的微调、多模态对齐。



**黄莹菁** (1972- ), 女, 博士, 复旦大学计算机科学技术学院教授, 主要研究方向为自然语言处理、信息检索、数据密集型计算。



**魏忠钰** (1987- ), 男, 博士, 复旦大学大数据学院副教授, 主要研究方向为大语言模型、社会计算、论辩挖掘。

收稿日期: 2024-06-14

通信作者: 魏忠钰, zywei@fudan.edu.cn