山西大学学报(自然科学版)48(4):653—665,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024159

基于法条检索的生成式法律问答研究

李明达[1],邸洪波[2],孙媛媛[1*],王艳华[3],杨志豪[1],林鸿飞[1]
(1.大连理工大学 计算机科学与技术学院,辽宁 大连 116000;
2.大连市公安局网安支队,辽宁 大连 116000;
3.中国人民解放军空军通信士官学校,辽宁 大连 116000)

摘　要:基于预训练语言模型的法律问答系统无法灵活理解用户的意图,且缺乏外部知识的整合,难以达到预期的效果。为此,该文提出了基于法条库的细粒度刑法问答数据集(Fine-grained Criminal Law Question Answering,FCL-QA),并基于该数据集,提出了基于大语言模型的法条检索增强问答框架(Statutory Articles Retrieval Augmented Question Answering Framework,SaRAF)。其核心思想是通过多等级分类定位到问题所属的主题,通过主题缩小法条的范围便于进行检索,并最终利用大语言模型生成答案。实验结果表明,SaRAF框架优于无法条生成与传统检索增强生成(Retrieval-augmented Generation,RAG)的方法,在FLC-QA数据集上取得了42.27%的ROUGE-L *F*1分数、27.78%的BLEU-4分数和72.52%的BERTScore分数。

关键词:大语言模型;问答系统;检索增强
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:0253-2395(2025)04-0653-13

Research on Generative Legal Question Answering Based on Statutory Articles Retrieval

LI Mingda[1], DI Hongbo[2], SUN Yuanyuan[1*], WANG Yanhua[3], YANG Zhihao[1], LIN Hongfei[1]
(1. *School of Computer Science and Technology, Dalian University of Technology, Dalian* 116000*, China;*
2. *Cybersecurity Protection Detachment, Dalian Municipal Public Security Bureau, Dalian* 116000*, China;*
3. *Chinese People's Liberation Army Air Force Communications NCO Academy, Dalian* 116000*, China*)

Abstract: Pre-trained language model-based legal question answering systems struggle to flexibly understand users' intent and lack the integration of external knowledge, making it difficult to achieve desired results. To address this, this paper proposes a fine-grained legal question answering dataset based on the criminal law articles library (FCL-QA). Based on FCL-QA, this paper proposes a Statutory Articles Retrieval Augmented Question Answering Framework (SaRAF) based on large language model. The core idea is to locate the category of the question through multi-level classification, narrow the scope of statutory articles through the category to facilitate retrieval, and finally generate the answer using a large language model. Experimental results show that the SaRAF outperforms both without statutory articles generation method and Retrieval-augmented Generation(RAG) method, achieving ROUGE-L *F*1 score of 42.27%, BLEU-4 score of 27.78% and BERTScore of 72.52% on the FCL-QA dataset.

Key words: large language model; question answering system; retrieval augmented

收稿日期:2024-09-14;**接受日期**:2024-11-19
基金项目:国家重点研发计划项目(2022YFC3301801)
作者简介:李明达(2000—),男,河南许昌人,硕士,研究方向为自然语言处理,问答系统。E-mail:2727782303@qq.com
* **通信作者**:孙媛媛(SUN Yuanyuan),E-mail:syuan@dlut.edu.cn
引文格式:李明达,邸洪波,孙媛媛,等.基于法条检索的生成式法律问答研究[J].山西大学学报(自然科学版),2025,48(4):653-665. DOI:10.13451/j.sxu.ns.2024159.

0 引言

法律问答的目标是为用户提供高质量、高可靠的法律咨询,是自然语言处理技术在司法领域的重要应用。智能的法律问答可以帮助律师和法官等司法专业人士快速地获取法律知识和信息,也可以为普通民众提供便利的法律服务。在社会对法律咨询服务的需求日益增加的当下,智能法律问答系统能够发挥出越来越大的作用,解决法律资源稀缺的问题。

传统的问答方法以检索式为主,通过根据问题计算出的答案得分,来从答案库中筛选出能够回答问题的答案。检索式的问答依赖于庞大的答案库,没有产生新答案的能力。一方面,尽管找到的答案包含正确的信息,但答案可能并不能与问题的重点匹配。另一方面,有的答案要求额外的推理与对来自用户的知识的整合,单纯的检索无法解决这一问题。因此,检索式的系统并不具备一个直接与用户交互的问答系统所应有的基本特性[1]。

为了解决检索式中的问题,生成式的方法被广泛应用在了自动问答任务中。该类方法使用T5等生成式模型,在闭卷或参考辅助信息的情况下生成回复。相较检索式的方法,生成式的方法灵活很多,尽可能降低了答案库对回答的影响。尽管如此,生成式的模型可控性较差,且存在幻觉问题[2]。在对可靠性要求较高,需求专业知识的法律领域,模型更难以生成令人满意的答复。

目前,主流的法律问答数据集包括选择题形式的司法考试数据集JEC-QA(Judicial Examination of China Question Answering)[3]、抽取式的法律阅读理解数据集CJRC(Chinese Judicial Reading Comprehension)[4]与检索和生成相结合的司法摘要数据集。与开放领域相比,法律问答研究的发展较为滞后。一方面,现有的法律问答数据集与实际的用户需求相差较大。在现实场景下,用户的输入往往只有问题。而用户期望获得的,是自然语言形式的流畅回复。另一方面,主流的法律问答数据集缺乏法律知识的指导。在高度专业化的法律领域,法规法条是提高模型生成质量,监督模型生成准确可靠回复的有力工具。而现有的数据集往往缺乏对问题、答案与法条之间的对应关系的标注。此外,由于法律数据集的构建依赖于法律专家的人工标注,高质量的中文法律问答数据集较为稀缺。

为了解决这些问题,本文提出了基于刑法法条库的细粒度中文法律问答数据集FCL-QA(Fine-grained Criminal Law Question Answering)。该数据集包含一万条收集自中文互联网上法律咨询平台的问答数据。在数据集中,根据分类维度的不同,法条与其对应的主题被分为刑法罪名、一般规定、相关法系三大类,三大类主题下包含细致到某一条具体罪名或某一类具体行为的子类。对每条数据的标注包括两部分,分别是该数据在上述两个层面上的分类与从法条库检索出的、对回答问题有帮助的法条。数据的标注流程以法条库为核心展开,以大模型投票与人工校对相结合的半自动化形式进行,尽可能地缓解了数据标注的压力。基于该数据集,本文提出了基于大语言模型的法条检索增强问答框架SaRAF(Statutory Articles Retrieval Augmented Question Answering Framework),整体流程包括主题预测、法条检索、答案生成三部分,如图1所示。

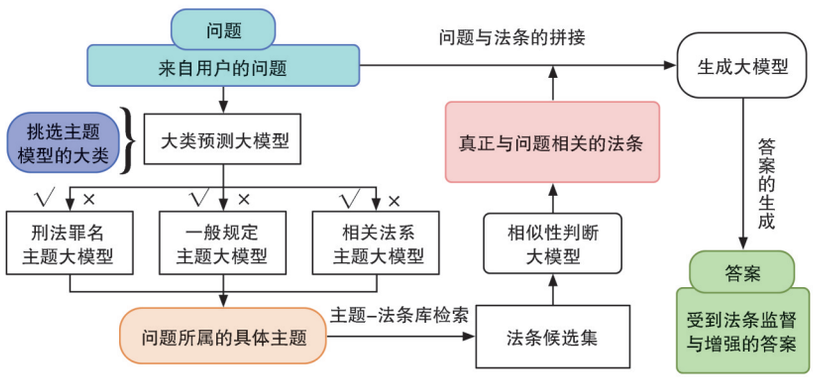
本文的主要贡献如下:

(1)提出了基于刑法法条库的法律问答数据集FCL-QA,该数据集收集自真实问答平台,与实际的法律咨询形式更加接近,能够反映法律问答实际的需求;

(2)提出了一种半自动化的数据标注流程,能够缓解数据标注的压力,低成本地扩大数据集的规模;

(3)针对法律问答任务,提出了一种基于大语言模型的法律问答框架SaRAF。该框架将法律问答所需要的能力拆分成分类、检索与生成三部分,通过将法条知识与模型经验相结合来提升模型生成回复的质量;

(4)在FCL-QA数据集上使用SaRAF框架进行了评测,结果证明,SaRAF框架可以检索出用于辅助答案生成的优质法条。与检索增强生成(Retrieval-augmented Generation, RAG)和不带任何法条相比,模型在ROUGE-L(Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence)$F1$指标上可提高



注:无色框表示参与框架的对象,深色框代表每阶段的结果。框架利用大类预测阶段的结果,挑选需要的主题预测模型来获取具体主题、利用主题缩小检索范围,找到真正需要的法条、并根据法条与问题,生成最终的答案。

图1 法条检索增强问答框架

Fig. 1 The framework of legal question answering based on statutory articles retrieval

11.17 和 4.63 个百分点,在 BLEU-4(Bilingual Evaluation Understudy-4-gram)指标上可提高 13.40 和 5.79 个百分点,在 BERTScore(Bidirectional Encoder Representations from Transformers Score)指标上可提高 5.75 和 2.00 个百分点。

1 相关工作

1.1 法律问答数据集

法律问答的发展离不开优质数据集的创建,按照任务类型,法律问答数据集可以分为:分类式、抽取式、检索式与文本生成式。其中生成式问答更符合实际问答系统的需求,因此成为研究的热点。

Zhong 等[3]提出了 JEC-QA,是目前最大的中文法律问答数据集,其中的数据来自中国国家司法考试。考虑到回答问题注重的方面,JEC-QA 中的数据被分为知识驱动与案件驱动两种类型。JEC-QA 提供了包含需要法律知识与法条的数据库,但并没有提供问题与法条之间的对应关系。Duan 等[4]通过抽取裁判文书中的事实描述内容,构建了中文法律阅读理解数据集 CJRC,填补了法律领域阅读理解研究的空白。该数据集的问题由标注者以转述的形式给出,涉及种类丰富。2021 年,中国法律智能技术评测司法摘要赛道提供了涉法舆情摘要数据集。该数据集通过融合多来源的答案数据与对多文档信息进行摘要精简,大幅度提高了答案的生成质量。Louis 等[5]提出了法语法律数据集 LleQA(Long-form Legal Question Answering),通过专家注释为每个问题提供相应的法律条款。

1.2 问答模型

在问答领域,深度学习的发展使得生成式的问答成为可能,Tan 等[6]提出了 S-Net(Synthesis Network)模型,采用抽取与生成相结合的范式,使用序列到序列模型(Sequence To Sequence,Seq2Seq)进行答案生成。预训练语言模型的出现为问答系统的发展带来了飞跃式的进步,Karpukhin 等[7]提出了 DPR(Dense Passage Retriever)模型,通过基于 BERT(Bidirectional Encoder Representations from Transformers)模型微调来完成检索任务,取得了远超 BM25(Best Matching 25)算法的效果。Garg 等[8]提出了 TANDA(Transfer and Adapt)模型,通过对预训练模型进行两次微调来提高模型的性能与鲁棒性。Roberts 等[9]以闭卷的形式微调 T5(Text-To-Text Transfer Transformer)模型,直接输入问题并获取对应的答案。Hsu 等[1]提出了 GenQA(Generative Question Answering)模型,在 T5 模型的基础上通过综合利用问题中信息与候选答案信息生成答案。

在法律问答领域,Wang 等[10]提出了法小飞(IflyLegal)系统,通过双向循环神经网(Bidirectional Recurrent Neural Networks,BiRNN)和卷积神经网络(Convolutional Neural Network,CNN)实现用户意图的判断,并通过匹配模型最终从答案库中检索出答案。Hoppe 等[11]提出了在文档检索的基础上,使用 BERT 与 BM25

构建的德语的法律问答系统。Kien 等[12]基于越南语提出了一种法律文本匹配模型,目的在于找到能够对回答问题有帮助的法律文章。Louis 等[5]通过检索增强的框架,构建了端到端的基于大语言模型的法律问答系统。

1.3 大语言模型

在生成式任务中,大模型展示出了卓越的性能,其代表性的工作是 Touvron 等[13]提出的 LlaMA(Large Language Model Meta AI)模型。LlaMA 模型使用注意力机制模型(Transformer)[14]架构,预测给定单词或元(token)作为下一个单词或 token 的概率。通过在 1.4×10^{12} 个 token 上进行训练,LlaMA 模型取得了强大的性能,在常识推理、闭卷问答、阅读理解、数学推理、代码生成与大规模语言任务理解中都取得了优秀的成果。

在 LlaMA 被提出后,许多工作都在 LlaMA 框架的基础上进行,大量开源的大语言模型被公布。Taori 等[15]提出了 Alpaca 模型,在 LlaMA 模型的基础上使用指令数据进行了进一步的微调,取得了媲美 GPT3.5(Generative Pre-trained Transformer 3.5)的水平。Chiang 等[16]提出了 Vicuna 模型,通过收集 ShareGPT 网站上的数据来进行指令微调,在低成本的情况下达到了接近 GPT 的能力。Bai 等[17]提出了千问(Qwen)模型,通过使用高达 3 万亿个 token 的数据进行预训练,为模型提供了可靠的知识源。在 LlaMA 的框架外,Du 等[18]提出了 ChatGLM(Chat Generative Language Model)模型,针对中文问答和对话进行了专门的优化,能够生成相当符合人类偏好的回答。

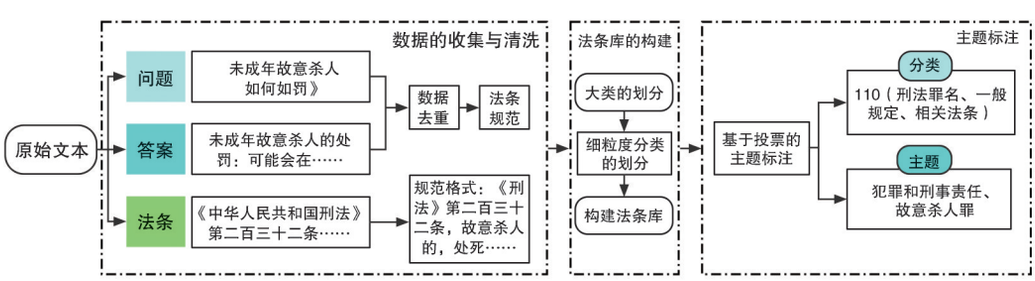
2 方法

本文构建了基于刑法法条库的法律问答数据集 FCL-QA,并基于该数据集的特点,提出了一种基于法条检索的法律问答框架 SaRAF。数据集的构建过程包括数据的收集与清洗、法条库的构建、主题标注三个阶段,如图 2 所示。根据数据集的特点,问答流程同样被划分为主题预测、法条检索与答案生成三部分。

2.1 数据集的构建

在数据收集与清洗阶段,本文从中文互联网上爬取了 49 060 条附有参考法条的问答对数据。本文过滤掉了重复的问答对,依照 2017 年修正版本的《中华人民共和国刑法》与 2018 年修正版本的《中华人民共和国刑事诉讼法》,依靠人工对爬取到的法条进行了校对。最终,本文将数据限定在了 38 251 条。并随机抽取了其中一万条数据作为第一批次的数据进行进一步的处理。

在法条库的构建阶段,本文首先将法条在刑法罪名、一般规定与相关法系三个大层进行了划分。刑法罪名分类中的法条涉及刑法中的某一特定罪名。一般规定分类包括适用于所有罪名的相关规定。相关法系分类中的法条并不属于刑事的范畴,但因为与刑事领域相关常被提及。在每一大类下,本文对主题的粒度进行了扩展。比如对于《中华人民共和国刑法》中某一章节名,本文将其下面的每一条法条作为单独的类别对待,从中拆分出抢劫罪、盗窃罪等具体的主题。最终,每一法条都被归入到了大类下的某一具体主题中。



注:数据的收集与清洗阶段筛选掉重复与不合规的数据,并按照格式对法条进行了整理。法条库的构建阶段在大类与主题两个层面上,将法条分入所属的类别。主题标注阶段依据法条的法条库分类与问题的性质,为问题分配分类与主题的标签。

图 2 数据集 FCL-QA 构建的流程

Fig. 2 The process of constructing the FCL-QA dataset

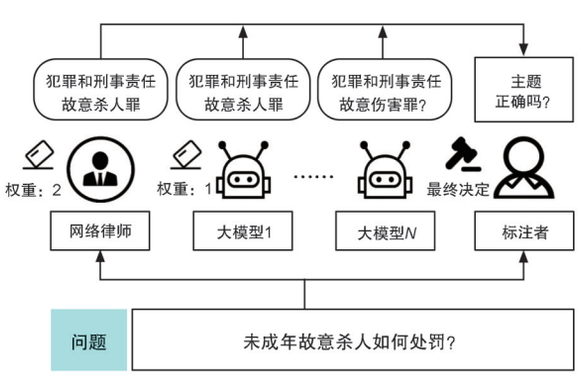
在主题标注阶段,通过将爬取到的法条与法条库进行对照,本文能够得到问题所属的主题。然而,网络上爬取到的法条存在与问题不匹配情况。此外,对于涉及多种主题的问题,单靠法条所属的主题不足以涵盖该问题涉及的所有情况。为了解决这一问题,本文通过人工的标注者对数据的主题标签进行修正。本文使用基于大语言模型的投票方法,在降低标注成本的前提下尽可能地获取高质量的数据,如图3所示。

为了训练标注模型,本文首先依靠人工对数据集中三千条数据涉及的主题进行了标注。在标注过程中,标注者被要求给出问题所涉及的所有主题,即使回答问题并不需要该主题下的法条。基于这三千条数据,本文使用低秩自适应(Low-Rank Adaptation,LoRA)[19]微调的方法,在ChatGLM-6B模型上使用不同的LoRA秩、LoRA缩放因子、学习率等超参进行了指令微调,训练出五个不同的主题预测模型。为了确保生成的主题在规定的主题集内,本文使用文本到向量(text2vec)模型,通过余弦相似度计算找到模型结果在主题库内对应的主题,具体如式(1)所示:

$$a_{\text{similarity}}(t_1,t_2)=\frac{\boldsymbol{e}_1\cdot\boldsymbol{e}_2}{\|\boldsymbol{e}_1\|\|\boldsymbol{e}_2\|},\tag{1}$$

其中t_1表示不符合标准的主题,t_2为来自主题库的主题。$\boldsymbol{e}_1$、$\boldsymbol{e}_2$分别表示t_1和t_2输入预训练BERT模型获得的表示向量。

对单个主题来说,当它的票数大于等于3时,如果该主题来自互联网律师,则表明该主题获得了大模型的再次确认,能够排除法条与问题无关的情况。如果该主题并非来自法条,则表明该主题获取了超过半数以上的大模型的认可,虽然没有相应的法条,但该主题应同问题强相关。因此,本文将票数大于等于三的主题视为可信。完成主题的获取后,本文通过人工校对来确保问题、答案与主题之间的关联。



注:投票的参与者包括互联网律师与五个经过微调后的大模型,权重分别为2和1。当主题获取到三票以上时,则认为该主题是可信的。

图3 主题标注的投票流程

Fig. 3 The process of voting for topic annotation

2.2 主题预测

对于主题预测部分,本文希望能够由浅入深,通过两阶段的分类任务来预测出问题所属的正确分类。本文通过对大模型进行指令微调来训练分类模型,框架如图4所示。

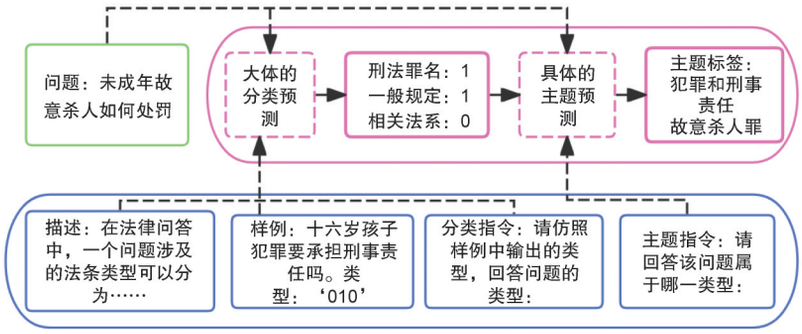
在大体分类预测阶段,本文根据法条所属的大类,按照刑法罪名、一般规定与相关法系的顺序,为问题分配标签信息,形式如"100"所示。在输入数据的描述部分中,本文对此分类任务进行了介绍并给出了详细的分类定义。为了保证模型生成的结果与预期一致,本文还在输入中给出了生成的样例,在加深模型对任务理解的同时,限制输出为期望的格式。在输入的最后,本文加入了指令,对模型的输出进行引导。

在具体主题预测阶段,本文根据问题所属大类下的主题标签构建了大类对应的主题预测训练集,通过指令微调训练出每个大类下的主题预测模型。在预测时,当大类预测阶段的结果在某一位置上为"1",本文便选择出相对应的模型用于主题预测。通过这一流程,原本的多主题多标签分类任务被转换为单标签多分类任务,在降低了主题预测的难度的同时,使得预测的结果更加准确可靠。在该阶段中,本文同样使用text2vec模型对不在主题集内的主题进行纠正。

2.3 法条检索

对于法条检索部分,本文将检索任务转换为相关性判别任务,通过指令微调引导大模型输出"0"或"1"来对问题与法条之间的关系进行判断,如图5所示。

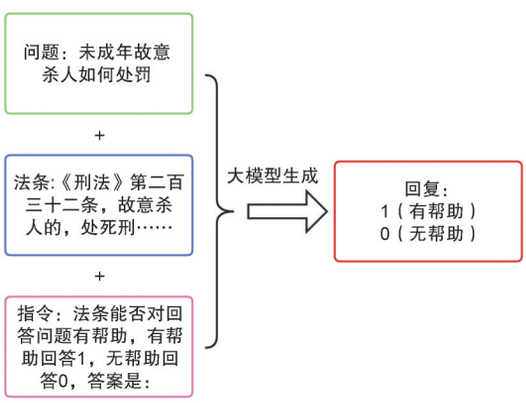
为了训练检索模型,本文利用数据集中标注好的主题标签与相关法条,构建了相关性判断训练集。对于每条与问题相关的法条,本文将其与问题拼接在一起,构成一个正样本对。



注:主题预测包括大体分类预测与具体主题预测两项子任务,大类预测按照描述、问题、样例、分类指令的顺序构造输入,输出中0与1分别代表是否属于该大类。主题预测使用问题、主题指令的顺序构造输入,输出为具体主题。

图4 主题预测的流程

Fig. 4 The process of topic prediction



注:法条检索的输入为问题、法条与指令三部分的拼接,输出为0或1,分别代表无帮助和有帮助。

图5 法条检索的流程

Fig. 5 The process of statutory articles retrieval

根据主题标签,本文从整体的法条库中检索出了问题对应的候选法条集合,并使用集合中与问题无关的法条构成负样本对。Cai等[20]的研究表明,处理数据集中的不平衡问题对分类模型的正确性和准确性十分重要。因此,本文在构建检索训练集时对负样本的数目加以限制,保证正负样本的数目能够保持一致。

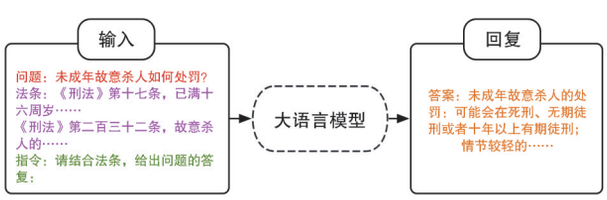
在实际的问答场景下,本文根据主题预测阶段的结果为问题筛选出候选法条集,并按照指令数据的格式拼接问题与法条,输入给大模型进行判断。为了防止问题找不到对应的法条,本文利用全量的正负样本集合,额外训练了SBERT(Sentence Bidirectional Encoder Representations from Transformers)[21]模型。SBERT模型的训练目标是,给定问题q、相关法条p和无关法条n,使得q和p之间的距离尽可能小,q和n之间的距离尽可能大,具体如式(2)所示。

$$L=\max(\|s_q-s_p\|-\|s_q-s_n\|+\epsilon,0), \quad (2)$$

其中s_x表示q、p和n对应的句向量,边缘参数ϵ表明s_q与s_p的距离应至少比s_q与s_n近ϵ。当问题找不到对应的法条时,本文通过SBERT模型来进行相似度判断,寻找到候选集合中与问题最匹配的法条。

2.4 答案生成

对于答案生成部分,本文将问题与法条进行了拼接,输入到大模型中进行答案生成,如图6所示。



注:答案生成部分的输入为问题、法条、指令的拼接,经过大模型后,输出为最终的答案回复。

图6 答案生成流程

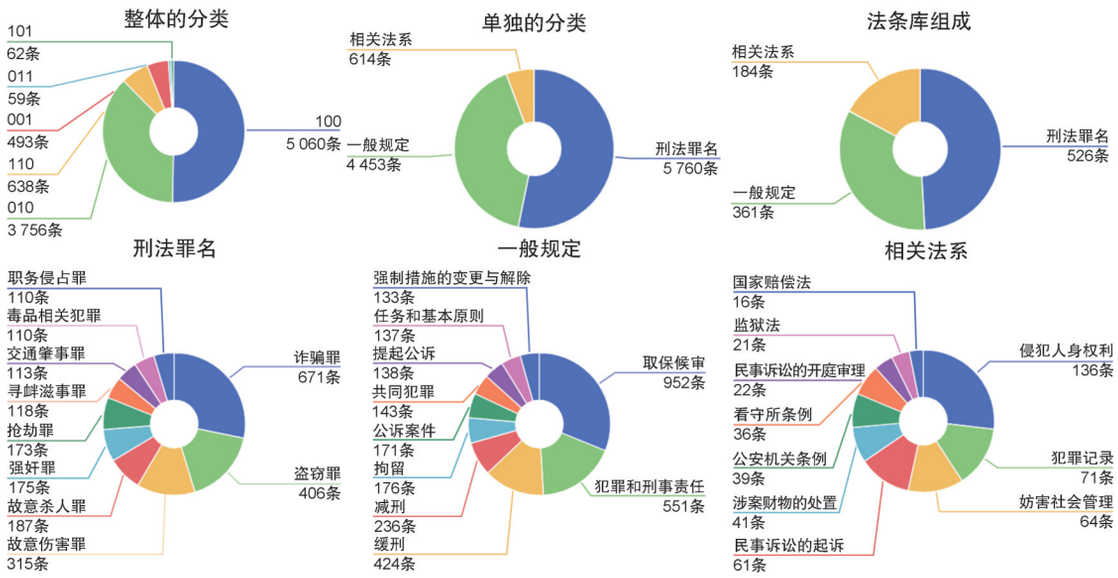
Fig. 6 The process of generating answers

在本部分中,本文利用了大模型优秀的自然语言生成能力,实现了法条知识与模型经验信息的融合。在训练阶段,使用的法条数据为经过标注后的准确法条。在测试阶段,使用的法条数据为从法条检索阶段获得的法条。

3 实验与分析

3.1 数据集

本文主要在本文构建的FLC-QA数据集上进行实验。图7上半部分展示了数据集与法条库的具体分布,下半部分展示了问答数据中每



注:上半部分中,中间为归类在三种大类下的数据分布,左边为归类在三大类的六种组合下的数据分布,右边为三大类下的法条数目的分布。下半部分中,左边、中间、右边分别为刑事罪名、一般规定、相关法系下的标注数目最多的十类主题。

图7 FCL-QA数据集的分布

Fig. 7 The distribution of the FCL-QA dataset

个大类下数目最多的十个主题的统计结果。在数据集中,每条数据包含一对问答对、若干条问题的相关法条、问题所属的大类与问题所属的主题。

本文按9∶1的比例,对训练集与验证集进行划分。在主题预测任务中,训练集通过根据大类标签对训练集进行归类获得。主题预测、法条检索任务的验证集分别根据流程中上一步的结果获得。表1中为ChatGLM3-6B(Chat Generative Language Model 3-6 Billion Parameters)模型上的实验结果,不同模型间存在一定的误差。

表1 训练集与测试集的分布

Table 1 The distribution of the training set and test set

任务类型	训练集	验证集	总计
大类预测、答案生成	9 061	1 007	10 068
主题预测(刑法罪名)	5 324	602	5 926
主题预测(一般规定)	4 067	432	4 499
主题预测(相关法系)	610	21	631
法条检索	21 689	8 091	29 780

3.2 评价指标

对于大类预测任务,本文使用准确率(Acc)作为评价指标。对于主题预测任务,本文使用Micro-$F1$分数(Top_$F1$)作为评价指标。本文使用N_{correct}、N_{predict}与N_{true}分别代表正确预测出的主题数、模型预测出来的主题数与实际预期的主题数。通过计算对应的精确率P与召回率R,进而计算Micro-$F1$分数,如式(3)—式(5)所示。

$$P=\frac{N_{\text{correct}}}{N_{\text{predict}}}, \tag{3}$$

$$R=\frac{N_{\text{correct}}}{N_{\text{true}}}, \tag{4}$$

$$F1=\frac{2\times P\times R}{P+R}。 \tag{5}$$

对于法条检索任务,本文实际将其转换成进行相关性判断的二分类任务,采用Micro-$F1$分数(Re_$F1$)进行评价。本文将模型检索到的相关法条数目、模型检索到的所有法条数目与问题相关的法条数目分别记为N_{correct}、N_{predict}与N_{true},代入式(3)—式(4)中计算精确率P与召回率R,并最终根据式(5)计算$F1$分数。

对于答案生成任务,本文采用ROUGE-L和BLEU-4,根据重叠词语的相似度度量评估生成质量,采用BERTScore,根据语义相似度评估生成质量。ROUGE-L指标在计算时使用了机器输出C与参考答案S的最长公共子序列LCS,如式(6)—式(8)所示。

$$R_{\text{LCS}}=\frac{l(C,S)}{l(S)}, \tag{6}$$

$$P_{\mathrm{LCS}}=\frac{l(C,S)}{l(C)}, \tag{7}$$

$$m_{\mathrm{ROUGE\text{-}L}}=\frac{(1+\beta^2)R_{\mathrm{LCS}}P_{\mathrm{LCS}}}{R_{\mathrm{LCS}}+\beta^2P_{\mathrm{LCS}}}, \tag{8}$$

BLEU通过计算单词级别的准确性来衡量句子的流畅性,使用长度惩罚因子BP使得模型的评估更加准确,如式(9)—式(11)所示。

$$p_n=\frac{\sum_{C\in\{A_{\mathrm{Candidate}}\}}\sum_{n_{\mathrm{gram}}\in C}N_{clip}(n_{\mathrm{gram}})}{\sum_{C'\in\{A_{\mathrm{Candidate}}\}}\sum_{n_{\mathrm{gram}}'\in C'}N(n_{\mathrm{gram}})}, \tag{9}$$

$$BP=\begin{cases}1 & \text{if } c>r\\ \mathrm{e}^{1-r/c} & \text{if } c\leqslant r\end{cases}, \tag{10}$$

$$m_{\mathrm{BLEU}}=BP\cdot\exp\left(\sum_{n=1}^{N}w_n\log p_n\right)。 \tag{11}$$

其中$A_{\mathrm{Candidate}}$指候选结果项,n_{gram}指文本中连续出现的n个词语,$N(n_{\mathrm{gram}})$指某个n_{gram}在候选结果中出现的次数,$N_{clip}(n_{\mathrm{gram}})$指$n_{\mathrm{gram}}$在候选结果中出现的次数,但最大不超过它在参考文本中出现的次数。

BERTScore将问题和回答编码为BERT向量,通过计算生成文本与参考文本的余弦相似度,捕获两个句子的语义相似性。生成文本在参考文本中的相似度最大值为该词语的精确率P,参考文本在生成文本中的相似度最大值为该词语的召回率R,综合P与R,可以计算$F1$分数,如式(12)—式(14)所示。本文使用chinese-roberta-large模型来获取BERT向量表示。

$$P_{\mathrm{BERT}}=\frac{1}{|\boldsymbol{x}|}\sum_{x_j\in\boldsymbol{x}}\max_{x_i\in x}\left(x_i^{\mathrm{T}}\boldsymbol{x}_j\right), \tag{12}$$

$$R_{\mathrm{BERT}}=\frac{1}{|x|}\sum_{x_i\in x}\max_{x_j\in\boldsymbol{x}}\left(x_i^{\mathrm{T}}\boldsymbol{x}_j\right), \tag{13}$$

$$F_{\mathrm{BERT}}=2\times\frac{P_{\mathrm{BERT}}\times R_{\mathrm{BERT}}}{P_{\mathrm{BERT}}+R_{\mathrm{BERT}}}。 \tag{14}$$

3.3 基线模型

实验中采用了10个模型作为数据集上的基线模型,分别为:

(1) ChatGLM3-6B:智谱AI和清华大学KEG实验室联合发布的对话预训练模型。

(2) LLaMA3-8B(Large Language Model Meta AI 3-8 Billion Parameters):Meta发布的大语言模型,是目前最先进的开源大模型。

(3) Qwen-7B(千问-7B):阿里云开源的中文大语言模型,在中英文评测任务中表现出色。

(4) LegalEagle:在InternLM2-chat-7B的基础上使用法律问答进行了指令微调,具备出色的文本处理、法律推理和知识检索能力。

(5) LaWGPT:以Chinese-LlaMA-7B为大模型底座,进行了大规模的中文法律语料预训练,并基于法律对话数据集进行指令精调。

(6) 太令:以Qwen-7B为基座模型,结合裁判文书、合同等专业司法数据进行深度训练的司法大模型。

(7) GPT3.5:openai发布的自然语言处理模型,目前提供可调用的接口。

(8) T5 PEGASUS(Text-to-Text Transfer Transformer with PEGASUS Pre-training):生成式预训练模型,使用伪摘要式的预训练提高模型的文本生成表现。

(9) T5 Copy:在T5 PEGASUS的基础上加入了Pointer结构,同样进行摘要式的预训练任务。

(10) SBERT:使用孪生网络衡量两个输入文本的相似程度,获取包含语义信息的句向量。

3.4 实验设置

本文的所有实验均基于深度学习框架Pytorch,在Linux平台上使用一张RTX4090显卡训练。实验环境如表2所示。

在实验中,本文在框架上使用不同的大模型,基于Firefly项目进行量化低秩适配(Quantized Low-Rank Adaptation,QLoRA)指令微调,完成框架下的各项任务流程。除了大模型外,本文还使用相同的输入,在法条检索阶段训练了SBERT模型、在答案生成阶段训练了T5模型作为对照。本文在数据集上以2×10^{-4}的学习率微调2个迭代轮次,训练批大小设置为4。对于lora_rank、lora_alpha等参数,本文均设置为16。本文使用训练中最后保存的LoRA权重,用于验证集上的测试。

表2 实验环境配置

Table 2 The settings of experimental environment

项目	参数配置
操作系统	Ubuntu 22.04
处理器	Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10 GHz
显卡	NVIDIA GeForce RTX 4090
CUDA版本	12.1

3.5 实验结果

表3展示了不同的模型在SaRAF框架中的大类预测与主题预测阶段上的实验结果。本文设计了单阶段的主题预测实验,把各类型的主题混杂在一起作为生成目标来微调大模型,并将其实验结果与在SaRAF框架中进行两阶段预测时的结果进行了对比。实验结果表明,在使用两阶段预测后,所有模型在主题预测阶段中的表现都取得了提升。其中ChatGLM3-6B模型表现优异,在*F*1指标上提高了14.33个百分点。在单阶段预测时,LegalEagle模型取得了最好的成绩,在*F*1指标上分别领先ChatGLM3-6B模型、太令模型和LawGPT模型0.92、4.81和12.87个百分点。经过两阶段预测后,ChatGLM3-6B模型实现了性能上的反超,后二者与LegalEagle模型间的差距被减小到了3.38和6.99个百分点。这一结果表明,两阶段的主题预测可以在一定程度上弥补模型之间性能的差距,为较弱模型提供追平甚至超越更强的模型的可能性。

在SaRAF框架中,检索阶段使用的检索器为使用正负样本数目平衡的训练集训练,基座模型与前一阶段中相同的大模型。本文将同样经过平衡训练后的SBERT模型作为检索器,与SaRAF框架的实验结果进行对比,结果如表4所示。实验结果表明,大部分模型在使用SBERT模型检索时的精确率优于使用SaRAF框架,二者间可相差1.11~7.84个百分点。在召回率上,除了LawGPT模型外,使用SaRAF框架的结果优于SBERT模型,存在1.21~3.35个百分点之间的差距。在从知识库中检索知识的场景下,与衡量检索出的法条中有多少正样本的精确率相比,正样本的覆盖率,即衡量检索到的知识能否满足回答问题的需求的召回率更加重要,因此SaRAF框架更符合检索增强问答任务的需求。在3.6节中,本文将对这一点进行详细分析。

表5中展示了不同模型在答案生成阶段的各项指标上的实验结果。在使用RAG的情况下,本文将输入部分中的法条部分替换为通过向量相似度检索获得的法条库片段。对于预训练模型T5与GPT3.5,本文仅对比其在准确法条下的生成结果。实验结果显示,使用SaRAF进行检索增强的效果全面优于使用RAG。在ROUGE-L *F*1和BLEU指标上,二者之间分别存在2.62~11.17和2.63~13.40个百分点之间的差距。在BERTScore指标上,二者之间存在1.54~5.75个百分点的差距。这是由于RAG检索出的片段无规则且质量较低,对回答问题帮助有限,可能导致模型无法正确地遵循指令,从而输出无意义的结果。同样经过指令微调,即使使用了准确法条,预训练T5模型的表现仍然不如SaRAF框架下的大模型,在三项指标上分别存在0.25~7.78、1.59~6.57和1.21~5.80个百分点之间的差距。在大模型内部,GPT3.5在各项指标上的表现同样不如经过指令微调的基座大模型,在三项指标上存在10.85~13.87、10.67~14.01和8.44~10.41个百分点之间的差距。GPT3.5生成了大量高质量的正确回复片段,但未经过微调的GPT3.5同时也生成了大量

表3 大类预测和主题预测实验结果(%)

Table 3 Experimental results of category prediction and topic prediction (%)

模型	单阶段预测			SaRAF			
	P	*R*	*F*1	*P*	*R*	*F*1	Acc
GLM3-6B	80.63	74.10	77.23	**93.47**	**89.72**	**91.56**	88.08
LLaMA3-8B	76.26	68.17	71.98	78.33	71.59	74.81	87.68
Qwen-7B	**82.02**	73.09	76.86	81.70	74.26	77.81	88.77
LegalEagle	81.75	**74.85**	**78.15**	81.53	75.61	78.46	**89.37**
LaWGPT	71.25	60.23	65.28	75.82	67.58	71.47	86.29
太令	77.36	69.93	73.34	78.44	72.01	75.08	88.68

注:*P*(Precision)代表精确率,衡量了主题预测结果中的正确主题数目。*R*(Recall)代表召回率,衡量了所有正确的主题中有多少被预测出来。*F*1是综合*P*与*R*计算出的分数。Acc代表大类预测阶段的准确率,衡量了大类预测阶段的结果的正确性。

表4 法条检索实验结果(%)

Table 4 Experimental results of statutory articles retrieval (%)

模型	SBERT			SaRAF		
	P	*R*	*F*1	*P*	*R*	*F*1
GLM3-6B	**61.09**	**78.02**	**68.52**	**55.98**	**81.37**	**66.33**
LLaMA3-8B	53.55	64.16	58.38	55.89	65.41	60.28
Qwen-7B	56.54	66.83	61.25	50.89	68.76	58.49
LegalEagle	56.01	68.42	61.60	48.17	71.35	57.51
LaWGPT	51.41	61.06	55.82	55.97	59.57	57.71
太令	54.59	65.49	59.55	53.48	66.67	59.35

的无意义片段,导致精确率偏低,拉低了$F1$指标。

表5 答案生成实验结果(%)

Table 5 Experimental results of answer generation (%)

模型	ROUGEL-$F1$		BLEU		BertScore	
	RAG	SaRAF	RAG	SaRAF	RAG	SaRAF
T5 PEGASUS	—	34.49	—	21.22	—	66.72
T5 COPY	—	39.00	—	24.61	—	69.34
GPT3.5	—	28.40	—	13.78	—	62.11
GLM3-6B	29.72	40.89	13.09	26.49	65.86	71.61
LLaMA3-8B	32.19	40.51	16.16	26.67	66.89	71.39
Qwen-7B	**39.33**	41.95	**25.16**	**27.79**	**70.89**	72.43
LegalEagle	35.88	**42.27**	20.15	27.50	68.85	**72.52**
LaWGPT	31.31	39.25	14.07	24.45	66.06	70.55
太令	34.77	40.69	19.27	26.20	68.41	71.37

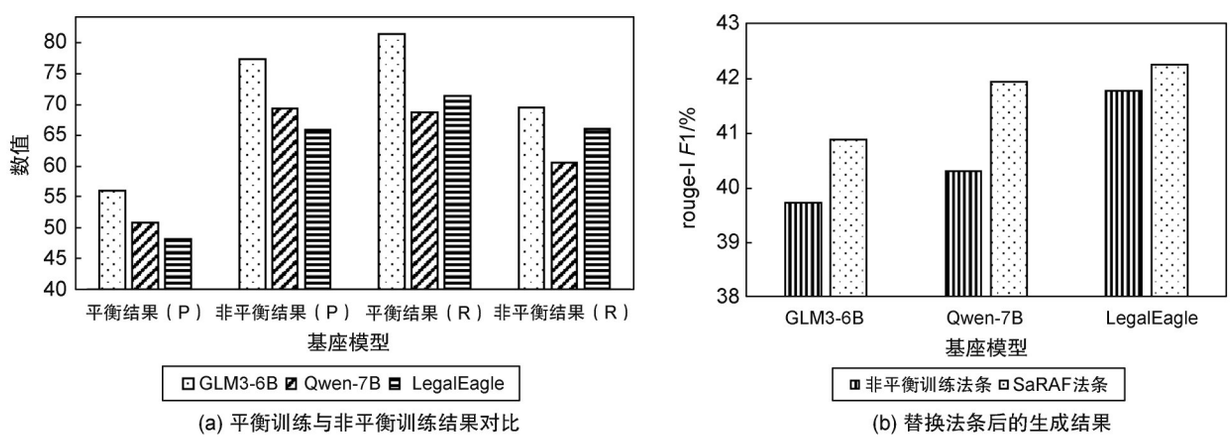
注:ROUGE-L $F1$、BLEU、BERTScore为衡量生成质量的常用指标。前两者从词法、语法的相似度上对结果进行衡量,后者从语义相似度上对结果进行衡量。

3.6 检索指标分析

在检索阶段,本文构造了正负样本平衡的训练集,来防止过多负样本干扰模型对正样本的识别能力。为了验证这一策略是否有效,本文构造了包含76 890条训练数据的非平衡全量检索训练集。由于大规模数据导致训练成本增加,本文选取了在先前实验中表现出色的Chat-GLM3-6B、Qwen-7B和Legal-Eagle模型进行了实验,结果如图8(a)所示。实验结果表明,在使用非平衡训练集进行训练后,模型检索出的正例更少,正例中真正例的数目也随之减少,导致了召回率的下降。负样本上的训练提高了模型对负样本的识别能力,降低了结果中假正例的占比,导致精确率的提高。其中Chat-GLM3-6B模型变化显著,在精确率上提高了21.32个百分点,在召回率上降低了11.95个百分点。本文将答案生成阶段中输入的法条替换为经过非平衡训练后检索到的法条,生成的答案在ROUGE-L $F1$指标上的结果如图8中(b)所示。可以看到,与SaRAF框架的生成结果相比,使用非平衡训练获取的法条时,$F1$指标下降了0.48~1.64个百分点。该结果说明,在检索增强的场景下,召回率与答案的质量之间存在正向的关联,这可能是由于用户生成的大模型依靠自身能力对法条中的有用信息进行了提取。大模型不擅长无中生有,但具备优秀的自然语言理解能力。在这种情况下,更高的召回率更有助于模型获取信息。因此相比于SBERT模型,SaRAF框架是更合适的检索器。

3.7 消融实验

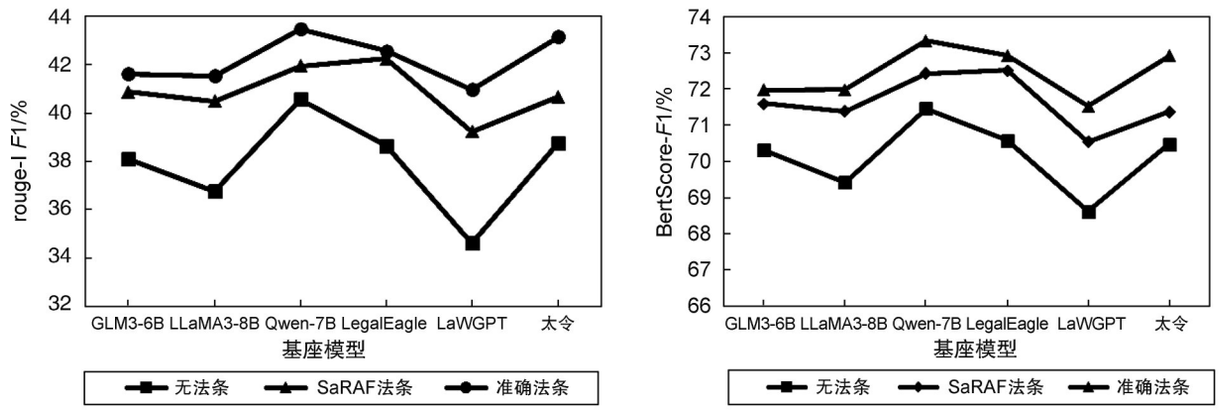
为了验证法条在答案生成中发挥的作用,本文消融掉法条检索的过程,设计了无法条与使用正确法条的两种变体。图9展示了两种变体与SaRAF框架分别在ROUGE-L $F1$和BERTScore指标上的对比结果。结果显示,在没有法条辅助生成时,模型的生成结果全面劣



注:(a)横坐标中平衡结果(P)和(R)分别代表使用正负样本平衡训练集训练后,检索阶段在精确率和召回率上的结果。非平衡结果(P)和(R)代表使用非平衡训练集训练后的相应结果。纵坐标代表精确率和召回率的数值。不同的基座模型以不同的柱形图案加以区分。(b)横坐标代表使用的基座模型类型,纵坐标使用ROUGE-L $F1$的数值,来衡量对应基座模型在生成阶段的表现,两种来源不同的法条的结果以不同的柱形图案加以区分。

图8 检索指标分析

Fig. 8 The analysis of statutory articles retrieval metrics



(a) ROUGE-L *F*1指标上的实验结果　　(b) BERTScore指标上的实验结果

注:(a)和(b)分别采用ROUGE-L *F*1和BertScore指标,展示了不同基座模型的生成质量与使用法条类型的关系。使用法条质量越高,模型的生成效果越好。

图9　法条作用消融实验

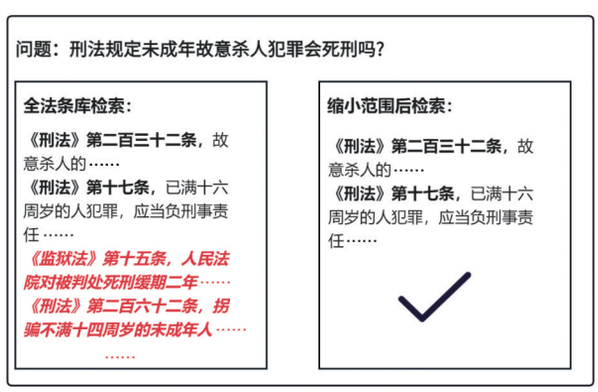
Fig. 9　Ablation study on the impact of statutory articles retrieval

于SaRAF框架下的生成结果。而在使用正确的法条后,模型生成的质量都获得了提升。这说明SaRAF框架中通过检索获取的法条是有效的,能够对答案生成产生积极的影响。通过进一步优化问答的流程,SaRAF框架仍然有着提升的空间。

3.8　案例分析

图10中展示了不使用多等级主题分类检索限定范围,直接利用向量相似度在整个法条库中进行检索的结果。可以看到,在获取到法条的过程中,主题起到了过滤掉无关法条的作用。在问题长度较短,包含信息有限的情况下,向量相似度无法准确地表示出问题的意图,导致检索出大量与回答问题无关的法条。在答案生成阶段,无关的法条会成为噪声拉低生成质量,误导模型输出错误的答案。

图11中展示了使用SaRAF框架为FCL-QA中问题生成答案的一个错误样例。可以看到,法条的存在与否对答案质量有着直接的影响。模型1未能找到与诈骗罪的量刑金额相关的规定,其回复直接忽略了问题中“500”的影响,错误回复了刑法中有关诈骗的处罚。在缺乏来自治安管理处罚法的法条的情况下,模型2的回复则仅停留在知道诈骗罪存在立案标准的程度上,对于500元是否达到立案标准,达到标准后如何处罚,模型仍然无法给出准确回复。



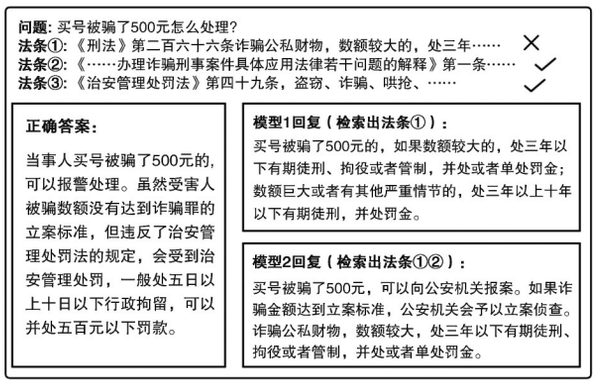
注:图中左边文本框中,以斜体表示的部分为错误地被检索出的法条,右边文本框为真正正确的法条。

图10　法条检索案例分析

Fig. 10　The case study of one FCL-QA dataset example on articles retrieval

3.9　幻觉分析

对于法律问答这种需要高度准确性的场景来说,小部分关键信息的不同就可能完全改变答案的正误,基于词重叠率与向量相似度的评价指标往往无法正确反映出生成的质量。此外,大模型往往缺乏法律领域的知识,容易产生幻觉,生成看似流畅,但错误地解释问题内容的答案。为了更全面地对生成的答案进行评估,本文从验证集中随机选取了100条数据,在使用ChatGLM3-6B作为基座模型的情况下,从RAG检索增强、SaRAF、无法条和正确法条等框架下获取到数据中问题的答案,并邀请人工标注者,对答案中是否存在幻觉进行了判断。本文定义了三种类型的幻觉,包括是否看似合



注:左边文本框为正确的答案,右边上面文本框为检索出错误法条后生成的答案,右边下面文本框为检索错误且检索不全时生成的答案。

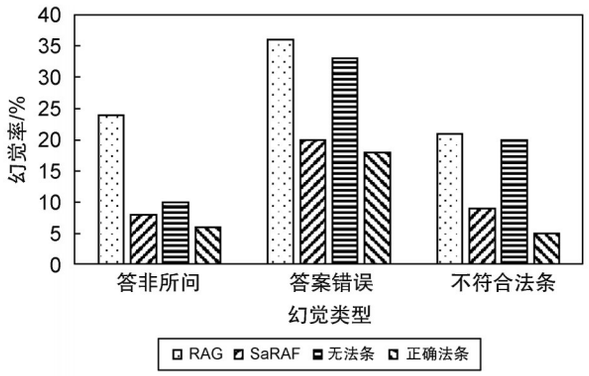
图11 答案生成案例分析

Fig. 11 The case study of one FCL-QA dataset example on answer generation

理但答非所问、是否回答出错误答案、是否不受相关法条的约束,结果如图12所示。实验结果表明,模型生成的答案中存在幻觉的问题,其中答案错误类型的幻觉率最高,在SaRAF框架上可高达20.00个百分点。答非所问、答案与法条矛盾的现象也普遍存在,在SaRAF框架的结果里可分别占8%和9%。在所有种类的框架中,RAG框架的幻觉率在三方面都是最高的,其次是不带法条的生成,再之后是SaRAF。当提供给模型正确的法条时,幻觉率也随之降低,这说明法条在答案生成的过程中起到了监督与引导的作用。在法律问答场景下,因其天然的规范性,法条可以成为应对大模型幻觉问题的有力武器。通过利用法条对答案进行修正与参考,答案的真实性与质量都能得到较大的提高。

4 结论

针对生成式的法律问答任务,本文提出了一个细粒度的刑法问答数据集FCL-QA,该数据集收集了自然语言形式的刑法领域问答对,并对该类型问答中频繁使用的法条进行了整理,为问答对与法条之间建立了明确的检索路径。基于FCL-QA,本文提出了基于大模型的法条检索增强框架SaRAF。该框架通过多级分类获取问题所属的细粒度主题,通过主题辅助法条检索,并最终将大模型的经验与法条的知识相结合,生成高质量的回复。在FCL-QA数据集上进行的实验证明了SaRAF框架的可靠性。



注:横坐标为定义的幻觉类型,纵坐标为幻觉率的百分比数值,使用法条类型的不同以不同的柱形图案区分。

图12 幻觉率指标分析

Fig. 12 The analysis of Hallucination Rate

在未来工作中,本文将进一步对数据集进行扩充,并基于数据集的特点,探索如何提高大模型问答系统的性能。

参考文献:

[1] HSU C C, LIND E, SOLDAINI L, *et al*. Answer Generation for Retrieval-based Question Answering Systems [EB/OL]. (2021-06-02) [2024-07-21]. https://arxiv. org/abs/2106.00955

[2] JI Z W, LEE N, FRIESKE R, *et al*. Survey of Hallucination in Natural Language Generation[J]. *ACM Comput Surv*, 2023, **55**(12): 1-38. DOI: 10.1145/3571730.

[3] ZHONG H X, XIAO C J, TU C C, *et al*. JEC-QA: A Legal-domain Question Answering Dataset[J]. *Proc AAAI Conf Artif Intell*, 2020, **34**(5): 9701-9708. DOI: 10.1609/aaai.v34i05.6519.

[4] DUAN X Y, WANG B X, WANG Z Y, *et al*. CJRC: A Reliable Human-annotated Benchmark DataSet for Chinese Judicial Reading Comprehension[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 439-451. DOI: 10.1007/978-3-030-32381-3_36.

[5] LOUIS A, VAN DIJCK G, SPANAKIS G. Interpretable Long-form Legal Question Answering with Retrieval-augmented Large Language Models[J]. *Proc AAAI Conf Artif Intell*, 2024, **38**(20): 22266-22275. DOI: 10.1609/aaai.v38i20.30232.

[6] TAN C Q, WEI F R, YANG N, *et al*. S-net: From Answer Extraction to Answer Synthesis for Machine Reading Comprehension[J]. *Proc AAAI Conf Artif Intell*, 2018, **32**(1). DOI: 10.1609/aaai.v32i1.12035.

[7] KARPUKHIN V, OĞUZ B, MIN S, *et al*. Dense Passage

Retrieval for Open-domain Question Answering[EB/OL]. (2020-04-10) [2024-07-21]. https://arxiv. org/abs/2004.04906.

[8] GARG S, VU T, MOSCHITTI A. TANDA: Transfer and Adapt Pre-trained Transformer Models for Answer Sentence Selection[J]. *Proc AAAI Conf Artif Intell*, 2020, **34** (5): 7780−7788. DOI: 10.1609/aaai.v34i05.6282.

[9] ROBERTS A, RAFFEL C, SHAZEER N. How Much Knowledge Can You Pack into the Parameters of a Language Model?[EB/OL]. (2020-02-10)[2024-07-21]. https://arxiv.org/abs/2002.08910.

[10] WANG Z Y, WANG B X, DUAN X Y, *et al*. IFlyLegal: a Chinese Legal System for Consultation, Law Searching, and Document Analysis[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP−IJCNLP): System Demonstrations. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 97−102. DOI: 10.18653/v1/d19-3017.

[11] HOPPE C, PELKMANN D, MIGENDA N, *et al*. Towards Intelligent Legal Advisors for Document Retrieval and Question-answering in German Legal Documents[C]//2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). New York: IEEE, 2021: 29−32. DOI: 10.1109/AIKE52691.2021.00011.

[12] 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, PA, USA: International Committee on Computational Linguistics, 2020: 988−998. DOI: 10.18653/v1/2020.coling-main.86.

[13] TOUVRON H, LAVRIL T, IZACARD G, *et al*. LLaMA: Open and Efficient Foundation Language Models[EB/OL]. (2023-02-27) [2024-07-21]. https://arxiv.org/abs/2302.13971.

[14] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention Is All You Need[EB/OL]. (2017-06-12) [2024-07-21]. https://arxiv.org/abs/1706.03762.

[15] TAORI R, GULRAJANI I, ZHANG T, *et al*. Alpaca: A Strong, Replicable Instruction-following Model[EB/OL]. (2023−03−13)[2024−07−21]. https://crfm.stanford.edu/2023/03/13/alpaca.html.

[16] CHIANG W L, LI Z, LIN Z, *et al*. Vicuna: an Open-source Chatbot Impressing Gpt-4 with 90%* Chatgpt Quality[EB/OL]. (2023−03−30) [2024−07−21]. https://lmsys.org/blog/2023−03−30−vicuna/.

[17] BAI J Z, BAI S, CHU Y F, *et al*. Qwen Technical Report[EB/OL]. (2023-09-28) [2024-07-21]. https://arxiv.org/abs/2309.16609.

[18] DU Z X, QIAN Y J, LIU X, *et al*. GLM: General Language Model Pretraining with Autoregressive Blank Infilling[EB/OL]. (2021-03-18)[2024-07-21]. https://arxiv.org/abs/2103.10360.

[19] HU E J, SHEN Y L, WALLIS P, *et al*. LoRA: Low-rank Adaptation of Large Language Models[EB/OL]. (2021-06-17)[2024-07-21]. https://arxiv.org/abs/2106.09685.

[20] CAI X X, XIAO M, NING Z Y, *et al*. Resolving the Imbalance Issue in Hierarchical Disciplinary Topic Inference via LLM-based Data Augmentation[C]//2023 IEEE International Conference on Data Mining Workshops (ICDMW). New York: IEEE, 2023: 1424−1429. DOI: 10.1109/ICDMW60847.2023.00181.

[21] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks[EB/OL]. (2019-08-27) [2024-07-21]. https://arxiv. org/abs/1908.10084.