doi:10.11887/j.issn.1001-2486.24120026 http://journal.nudt.edu.cn

数据增强与规则引导的大语言模型电网法律答辩文书生成

黄成燕[1],查小云[2],丁群晏[1],胡 伟[3*]
(1. 国网江苏省电力有限公司电力科学研究院,江苏 南京 211103;2. 国网江苏省电力有限公司,江苏 南京 210024;
3. 南京大学 计算机学院,江苏 南京 210023)

摘 要:为提升大语言模型在少样本环境下生成电网法律文书的能力,结合数据增强与规则引导技术,提出一种基于大语言模型的少样本法律文书生成方法,解决了电网企业法律文书生成中样本稀缺、专业性强、实践经验复杂等问题。实验结果表明,该方法在电网法律答辩文书生成任务中取得了优异效果,有效提升了模型生成文本的质量与专业性。

关键词:法律答辩文书生成;少样本;大语言模型;规则引导;数据增强
中图分类号:TP391.1; D9 **文献标志码**:A **文章编号**:1001-2486(2025)04-180-09

论文拓展

Large language model-based legal defense document generation for power grid enterprises with data augmentation and rule guidance

HUANG Chengyan[1], *ZHA Xiaoyun*[2], *DING Qunyan*[1], *HU Wei*[3*]
(1. State Grid Jiangsu Electric Power Co., Ltd. Research Institute, Nanjing 211103, China;
2. State Grid Jiangsu Electric Power Co., Ltd, Nanjing 210024, China; 3. School of Computer Science, Nanjing University, Nanjing 210023, China)

Abstract: To enhance the ability of large language models to generate legal documents for the power grid sector under few-shot conditions, a few-shot legal document generation method based on LLM(large language models) was proposed, integrating data augmentation and rule guidance techniques. The proposed method addressed key challenges in power grid legal document generation, such as data scarcity, high domain specificity, and the complexity of legal practice. Experimental results show that the method achieves excellent performance in generating power grid legal defense documents, significantly improving the quality and professionalism of the generated texts.

Keywords: generation of legal defense document; few-shot; large language models; rule guidance; data augmentation

近年来,大语言模型(large language model, LLM)在自然语言处理领域展现出强大能力,逐步被应用于医疗、法律等专业性较强的垂直领域[1-3]。这些任务通常要求模型具备高度的专业性和准确性,因而为特定领域设计高效的解决方案具有重要意义。本文聚焦电网法律文书生成任务,探讨如何在少样本条件下利用大语言模型为电网企业生成高质量的答辩文书。

法律文书撰写高度依赖专业知识与实践经验,尤其在法律纠纷日益复杂的背景下更具挑战性。随着人工智能的发展,特别是大语言模型在自然语言生成任务中的突破,利用大语言模型辅助法律文书生成成为值得关注的方向[4-5]。已有研究探索了大语言模型在法律摘要生成、法律问答、法条解释等任务中的应用[5-8],并关注其幻觉问题和专业性表达能力。

答辩文书作为法律文书的一种,具有逻辑严谨、结构规范和内容针对性强等特点,其生成过程需耗费律师大量的时间和精力。电网企业法律答辩文书的生成任务面临多重挑战[6-9]:一是知识高度专业,涉及电网技术与法律法规交叉,要求文书既合法合规,又体现行业特点;二是数据稀缺,因隐私限制公开数据极少,少样本条件加剧训练难度;三是文书中蕴含丰富的实践经验和规则模式,

收稿日期:2024-12-13
基金项目:国家电网有限公司科技资助项目(1400-202318637A-3-2-ZN)
第一作者:黄成燕(1992—),女,江苏仪征人,硕士研究生,E-mail:chengyanhuang@126.com
*通信作者:胡伟(1982—),男,江苏南京人,教授,博士,博士生导师,E-mail:whu@nju.edu.cn
引用格式:黄成燕,查小云,丁群晏,等. 数据增强与规则引导的大语言模型电网法律答辩文书生成[J]. 国防科技大学学报,2025,47(4):180-188.
Citation: HUANG C Y, ZHA X Y, DING Q Y, et al. Large language model-based legal defense document generation for power grid enterprises with data augmentation and rule guidance[J]. Journal of National University of Defense Technology, 2025, 47(4): 180-188.

难以被模型捕捉,影响生成质量。如何在少样本条件下整合专家经验与规则,引导大语言模型生成高质量答辩文书,是关键问题。研究该任务不仅具有学术价值,也对电网企业实践具有重要意义。

大语言模型在少样本场景下的专业任务处理受到广泛关注。数据增强是缓解少样本问题的主流方法,通过改写或生成类似文本提升训练数据多样性[10-13]。Xu 等[13]利用检索推理机制生成多样化合成数据,Zeng[11]通过提示生成跨语言数据显著提升情感分析效果。当前尚缺乏针对法律文书特点的数据增强方法。结合大语言模型的语言生成能力与电网答辩文书的层次结构特征,设计一种多策略组合的数据增强方法,提升生成效果。

为提升大语言模型在专业领域复杂推理任务中的表现,规则引导[14-15]与检索增强方法受到广泛关注[16-18]。Chen 等[14]提出规则引导的检索生成(rule guided retrieval-augmented generation, RuleRAG)方法,利用知识图谱构建规则库并实现检索增强,显著提升复杂问答性能;Wang 等[16]的逻辑推理规则生成(logic scaffolding inferential rule generation, LOIRE)框架通过本体对象提取规则并结合大语言模型进行总结。尽管已有研究尝试将大语言模型与法律知识图谱结合以增强法律问答能力[19-20],但针对电网法律文书生成的研究仍较为空白。考虑到该领域知识专有、资源稀缺,现有方法难以直接迁移。本文提出一种基于大语言模型的自动化规则抽取方法,以引导答辩观点生成,提升生成文本的合理性与专业性。

综上,针对电网法律文书生成中样本稀缺与推理复杂的问题,提出一种基于大语言模型生成方法,主要包括少样本数据增强与规则引导微调两部分。设计了自动数据扩充与改写框架,生成结构与语义相近的文本,提升训练语料多样性;提出基于大语言模型的自动规则抽取方法,通过少量标注规则扩展有效规则,并用于引导模型生成,增强生成结果的合理性与针对性。实验结果表明,本文方法在多个大语言模型上均表现优异,数据增强与规则引导策略在提升生成质量方面发挥了关键作用,为类似任务提供了可行的参考路径。

1　电网法律答辩文书生成方法设计

1.1　任务描述与方法总体思想

本文研究在少样本条件下,利用预训练大语言模型生成电网企业法律答辩文书。任务目标是基于少量电网法律文书与答辩文书样本,训练生成模型以自动生成合理有效的答辩观点。该任务需同时理解电网专业术语和技术背景,并遵循法律文书的规范与逻辑,具有较高难度。

总体方法设计思路主要围绕如何有效解决“少样本”问题,同时提高生成结果的专业性和准确性。针对当前电网法律答辩文书生成中的主要挑战,设计了一种基于数据增强与规则引导的大语言模型电网法律答辩文书生成方法(large language model-based legal defense document generation with data augmentation and rule guidance, LADDER),如图 1 所示,本文方法设计可以分为两个关键技术部分:一是为了应对训练

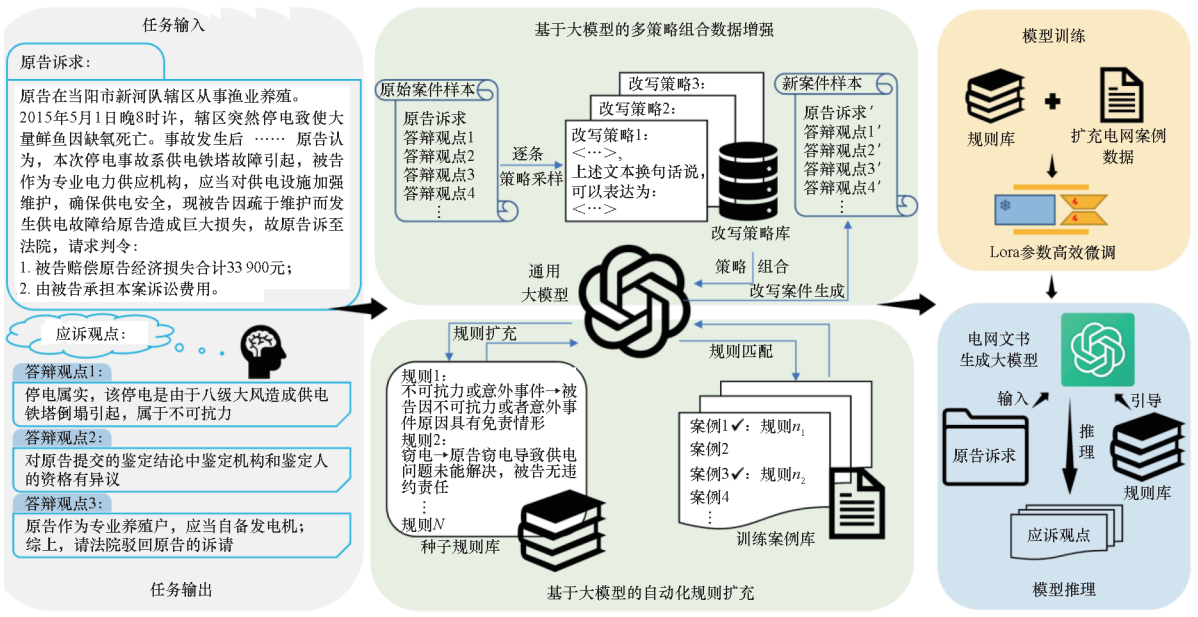

图 1　基于数据增强与规则引导的大语言模型电网法律答辩文书生成方法框架

Fig. 1　Framework of the proposed LLM-based model for generating legal defense documents in power grid

数据稀缺的问题提出的基于大语言模型的多策略组合数据增强技术。该技术的核心目标是通过增强训练数据的多样性,来提升大语言模型的泛化能力。二是基于大语言模型的自动化规则扩充技术。为提高生成答辩观点的精度和准确性,还提出了一种基于大语言模型的自动化规则扩充技术。其设计目标是通过自动提取和扩充规则,帮助大语言模型在生成过程中遵循特定的法律逻辑和观点框架。综上,本文所提基于数据增强与规则引导的生成方法,旨在提升电网法律答辩文书的生成质量,并解决少样本训练中的挑战,为实际应用提供切实可行的技术支持。

1.2 基于大语言模型的多策略组合数据增强

为应对电网法律答辩文书生成任务中训练样本不足的问题,提出并实现了一种基于通用大语言模型的数据扩充与改写框架。该方法利用大语言模型的自然语言生成能力自动扩充数据,缓解了少样本带来的训练难题,并兼顾了扩充样本的语义一致性与多样性。

确保改写后的文本与原始文本核心语义一致是数据增强过程中的首要考虑因素。为此,本文精心设计了特定的提示(Prompt),通过引导大语言模型理解并遵循这些提示,使得它能够在生成新文本时保留原有文本的核心信息和意图。为了实现这一目标,本文为提示设定了以下约束,首先是明确指令,即在向大语言模型发出指令时,明确要求其在改写过程中保持原有语义。接着是示例引导,利用大语言模型的上下文学习能力,提供一些示例,帮助大语言模型更好地理解要求。

保证语义一致性的基础上,本文设计了多种数据增强策略来提升样本的多样性:

1)表达风格变换:通过调整文本的表达方式来增加多样性。如将正式的表达转换为口语化的表达,或将复杂句转换为简单句。用“换句话说”“意味着”等自然语言引导大语言模型更换表达方式。

2)非关键实体替换:随机替换文本中不影响核心语义的实体,如人名、地名等。

3)摘要扩写式改写:文本摘要进行扩写,从而生成语义相似但表达方式不同的文本。

为进一步提高增强样本的多样性,本文采用一种多策略组合的方式。具体来说,对于每个待增强样本,本文方法将其划分为多个部分(如原告诉称以及被告辩称的不同观点),并为每个部分随机选择一种增强策略进行处理。这种随机组合的方式能够有效地增加生成样本的多样性。需要注意的是,数据增强不止需要提升数据表达的多样性,还需要保证增强前后数据核心语义的一致性。因此本文方法在每次改写后,增加了过滤步骤,即使用通用大语言模型对单个改写部分前后一致性进行判断,若语义不一致则放弃本次改写,重新采样策略进行改写。上述方法中使用的通用大语言模型为 ChatGPT,但不限于 ChatGPT。

本文提出的数据增强方法充分利用了大语言模型的强大能力,通过设计灵活的 Prompt 和组合多种增强策略,实现了对电网法律文书的高质量数据增强。该方法不仅能够有效缓解训练数据不足的问题,而且能够提升模型的泛化能力,为电网法律答辩文书的自动生成提供了支持。

1.3 基于大语言模型的自动化规则扩充

为提高大语言模型生成电网法律文书中答辩观点的精度和准确性,提出了一种基于大语言模型的自动化规则扩充方法。该方法通过从训练数据中自动抽取答辩规则,并利用这些规则引导大语言模型生成符合法律逻辑和实践经验的答辩观点。本任务中每一条规则均源于诉讼文书与对应答辩观点的组合。规则通过分析文书和观点之间的逻辑关系与法律依据提取而来,为大语言模型提供生成过程中法律框架和观点指导。如算法 1 所示,自动规则抽取方法包括规则匹配(见算法 2)和规则生成(见算法 3),均基于大语言模型的上下文学习能力。其中算法 2 中的规则匹配算法通过为大语言模型提供规则与(诉讼书、答辩观点)匹配的样例,使其判断每个答辩观点是否符合规则库中的某条规则,该步骤具体输出为观点编号及其对应的规则编号。此时模型会保存与每条规则匹配的诉讼书和答辩观点数量,即规则的匹配度。当答辩观点未命中规则库中的规则时,模型将调用规则生成算法根据规则库中的规则形式生成新的规则。通过从多个样例中抽取信息,模型能够总结出新的规则,进一步扩充规则库,具体操作见算法 3。每轮迭代中生成的新规则会通过规则匹配算法进行验证,并将匹配度高的 k 条规则保留在规则库中。需要注意的是,一个案例(具有一个或多个观点)可能匹配多条规则,覆盖率 t 指所有案例中与至少一条规则匹配的案例数量占总案例数量的比例。规则匹配覆盖率在规则匹配算法遍历整个案例数据集后计算(算法 1 的步骤 5)。当覆盖率大于等于预定阈值 t 时算法结束循环(本文该阈值设置为 0.7)。上述算法中使用的通用大语言模型为 ChatGPT,但不限于 ChatGPT。

算法 1　规则库扩充算法

Alg. 1　Rule base expansion algorithm

输入: 规则集合 *rule_set*;按类别分类的案例集合 *typed_examples_set*;覆盖率阈值 t;每轮新规则添加数量 k

输出: 规则集合 *rule_set*

rule_set←*RMA*(*rule_set*, *typed_examples_set*)
rule_cover_rate←计算当前规则库 *rule_set* 对所有案例 *typed_examples_set* 的覆盖率
while *rule_cover_rate* < t **do**
rule_set←*NREA*(*rule_set*, *typed_examples_set*, k)
rule_cover_rate←计算当前规则库 *rule_set* 对所有案例 *typed_examples_set* 的覆盖率
return 规则集合 *rule_set*

算法 2　规则匹配算法

Alg. 2　Rule matching algorithm

输入: 规则集合 *rule_set*;案例集合 *examples*

输出: 规则集合 *rule_set*

for each *example* **in** *examples* **do**
　for each *opinion* **in** *example* **do**
　　rules←使用大语言模型找到 *rule_set* 中与 *opinion* 匹配的规则
　end for each
　　rule_set←将 *example* 的编号记录在 *rules* 中
end for each
return 规则集合 *rule_set*

算法 3　新规则生成算法

Alg. 3　New rule generation algorithm

输入: 规则集合 *rule_set*;按类别分类的案例集合 *typed_examples_set*;新规则添加数量 k

输出: 规则集合 *rule_set*

for each *typed_examples* **in** *typed_examples_set* **do**
　examples←在 *typed_examples* 中随机采样若干条案例
　new_rules←使用大语言模型从 *examples* 中提取新规则
　new_rule_set←*new_rule_set* ∪ *new_rules*
　new_rule_set←*RMA*(*new_rule_set*, *typed_examples_set*)
　new_rule_set←筛选 *new_rule_set* 中匹配数量最多的 k 个规则
end for each
　rule_set←*rule_set* ∪ *new_rule_set*
return 规则集合 *rule_set*

基于大语言模型的自动化规则扩充方法通过规则匹配与生成的结合,能够有效扩充规则库,并提高大语言模型生成电网法律文书的准确性和专业性。此方法为少样本场景下的法律文书生成提供了有效的技术支持。

1.4　模型训练与推理

为了提高训练效率并减少参数量,本文采用了低秩微调[21](low-rank adaptation, LoRA)方法。LoRA 通过引入低秩矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 来对预训练模型的权重矩阵进行微调,其微调公式为:

$$\boldsymbol{W} = \boldsymbol{W}_0 + \boldsymbol{A} \cdot \boldsymbol{B} \tag{1}$$

其中,$\boldsymbol{W}_0$ 是原始权重,$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是低秩矩阵。LoRA 方法能够在不增加大量参数的情况下,提升模型在特定任务上的表现。

在训练过程中,本文使用 LoRA 方法基于扩充的数据和答辩规则微调大语言模型。规则在训练过程中以 Prompt 的形式被加入输入中,帮助模型生成合理的答辩观点。损失函数如下:

$$L = -\sum_{t=1}^{T} \ln P(y_t \mid y_{<t}, X) \tag{2}$$

其中:y_t 为第 t 个时间步生成的词;$y_{<t}$ 为之前的生成结果;X 为输入诉讼文书及规则信息;$P(y_t \mid y_{<t}, X)$ 在给定上下文和输入信息的条件下,模型生成下一个词的概率。通过最大化该条件概率,模型能生成符合上下文和规则约束的法律文书。

为了使模型能够在推理过程中灵活地应用自动抽取的规则,本文训练了一个基于 Transformer 的双向编码器表示(bidirectional encoder representations from Transformer, BERT)的规则分类器。该分类器的任务是根据新发案件的原告诉称,判别其中包含哪些规则。本文将该任务建模为一个多标签分类问题,其中包含 13 个类别:12 个为从 1.3 节中提取出的答辩规则类别以及 1 个无规则类别。每个案件可能对应多个规则,或不包含任何规则,因此该分类任务的类别数为 13。因此其损失函数如下:

$$L_{\mathrm{CE}} = -\sum_{i=1}^{C} \left[y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i) \right] \tag{3}$$

其中,C 是类别数,y_i 是实际标签,$\hat{y}_i$ 是模型预测标签。训练后多标签模型的实际标签准确率为 0.93,被用于推理阶段的规则预测。在推理阶段,将文书输入规则分类器,从中提取出对应的规则。规则与诉讼文书一起作为输入传递给大语言模型。大语言模型接收到信息后,在规则引导下生成合理且符合法律规范的答辩文书。生成合理的

答辩文书,确保生成内容的法律合规性和专业性。

2 实验

2.1 实验设计

2.1.1 数据集

实验所用数据集是基于电网企业的真实法律诉讼及答辩案例文书数据集,内容涵盖供用电合同纠纷、触电人身损害责任纠纷和财产损害赔偿纠纷等多个方面,经过清洗整理后得到1 249个样本,从中随机挑选105个案例数据作为测试集,剩余部分按照9∶1的比例划分为训练集和验证集。

2.1.2 比较方法

本文选择的对比方法分为:①现有的大语言模型,包括通用大语言模型LLaMA-2-7B-Instruct[22]、QWen2-7B-Instruct[23]和GPT-4o-mini,以及面向中国法律应用的开源大语言模型LawGPT[24];②规则引导的方法RuleRAG-ICL,该方法在输入中添加相关的规则引导模型的生成过程,无须额外训练即可提升大语言模型针对特定任务的表现;③基于LawGPT进行指令微调的方法LawGPT(微调)。在本文实验部分,RuleRAG-ICL选择GPT-4o-mini作为文本生成模型,①和②中的方法都提供3个演示进行上下文学习。本文提出的方法LADDER分别选用通用大语言模型QWen2-7B-Instruct和具有法律领域知识的LawGPT作为基础大语言模型进行实验,用以验证法律专业领域知识对方法性能的影响。在没有特殊说明的情况下,LADDER默认使用LawGPT。

2.1.3 评估方法

答辩文书生成属于自然语言文本生成任务,文本生成任务常见的评价指标包括双语评估替代方法(bilingual evaluation understudy, BLEU)、面向召回的自动摘要评估方法(recall-oriented understudy for gisting evaluation, ROUGE)和Perplexity。其中:BLEU计算生成文本和参考文本之间的n元组重合度,常用于评估机器翻译的质量;ROUGE主要用于摘要生成任务,评估生成文本和参考文本之间的重叠情况,侧重召回率;Perplexity通过语言模型计算生成文本的困惑度,反映生成文本的流畅性和合理性。

答辩文书生成的目标是为电网企业生成合理的答辩文书,与传统的文本生成任务侧重于生成文本的流畅性、与参考文本的相似性等不同的是,更加关注生成文本中是否包含参考答辩文书中对应的答辩观点,现有的评价指标并不适用于该任务。BLEU主要关注n元组的精确匹配,忽略了语义信息,而答辩文书中的观点表达可能多样化,仅靠n元组匹配无法捕捉到语义上的相似性。ROUGE侧重评估文本的重合程度,但答辩文书中的观点需要更加细致的语义理解,该指标无法很好地反映观点之间的相似性。Perplexity用于评估语言的流畅性,不直接衡量文本的语义,而流畅的答辩文书可能在语义上与参考答辩文书不一致。

基于BERTScore[25],本文提出了一种基于语义相似度的答辩文书评估方法。具体而言,给定参考文本T_r和生成文本T_g,首先将其分别划分成若干答辩观点$S_r=\{s_r^1,s_r^2,\cdots,s_r^n\}$和$S_g=\{s_g^1,s_g^2,\cdots,s_g^m\}$。之后,利用Sentence-BERT[26]对观点文本进行编码,得到各个观点语义的向量化表示:

$$\boldsymbol{e}=Sentence\text{-}BERT(s) \tag{4}$$

其中,$\boldsymbol{e}$是观点s经过Sentence-BERT编码得到的向量表示。最后,通过余弦相似度判断观点是否相似:

$$Match(s^i,s_j)=\begin{cases}1, & \cos(\boldsymbol{e}_i,\boldsymbol{e}_j)\geqslant\theta\\ 0, & \cos(\boldsymbol{e}_i,\boldsymbol{e}_j)<\theta\end{cases} \tag{5}$$

其中,$\cos(\boldsymbol{e}_i,\boldsymbol{e}_j)=\dfrac{\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{e}_j}{|\boldsymbol{e}_i|\cdot|\boldsymbol{e}_j|}$是余弦相似度,$\theta$是指定的阈值(本文实验中该阈值设置为0.8)。对于来自生成文本的观点s_g^i,只要其与参考文本的观点集合S_r中任意一个观点匹配即可,反之,对于来自参考文本的观点s_r^j,只要其与生成文本的观点集合S_g中任意一个观点匹配即可。在判断观点是否相似之后,本文基于查准率、查全率和F1值计算整个测试集上的指标,其中查准率为生成文本中匹配成功的答辩观点数量除以生成文本中的答辩观点总数;查全率为参考文本中匹配成功的观点数量除以参考文本的观点总数。

综上,为更好地衡量生成文本是否包含参考答辩文书中对应的答辩观点,满足答辩文书生成的目标,采用上述3种指标作为模型的最终测试指标。

2.2 主实验

对比不同方法在电网企业法律文书生成任务上的表现。选择3个随机数种子进行独立重复实验,汇报了不同模型在3个指标下的平均值以及标准差(结果见表1)。可以看到,LADDER在各项指标上均取得了显著的优势。此外,对实验结果进行显著性检验,采用配对t检验对比本文模型与各对比模型的指标进行统计显著性检验,F1值对应的每对p值最高为0.019 7(小于0.05),表明本文模型性能的优势具有统计学显著性。

表 1　主实验

Tab. 1　Main experiment

方法	查准率	查全率	F1 值
LLaMA-2-7B-Instruct	0.314 ±0.003	0.487 ±0.002	0.382 ±0.002
QWen2-7B-Instruct	0.531 ±0.001	0.592 ±0.003	0.560 ±0.002
GPT-4o-mini	0.493 ±0.000	0.607 ±0.003	0.544 ±0.001
LawGPT	0.546 ±0.009	0.546 ±0.008	0.546 ±0.008
RuleRAG-ICL	0.567 ±0.002	0.685 ±0.007	0.621 ±0.003
LawGPT (微调)	0.611 ±0.004	0.626 ±0.007	0.619 ±0.003
LADDER 基于 QWen2-7B-Instruct	0.643 ±0.004	0.751 ±0.006	0.693 ±0.001
LADDER	**0.660 ±0.005**	**0.781 ±0.006**	**0.715 ±0.005**

通过对比不同方法的表现,有以下结论:由于缺乏法律相关的背景知识,通用的大语言模型表现都欠佳,QWen2-7B-Instruct 在处理中文文本上有一定优势,因此表现明显优于 LLaMA-2-7B-Instruct,且整体效果略优于 GPT-4o-mini。

由于提供了法律相关的规则,RuleRAG-ICL 可以更好地指导文本生成的过程,与基于现有大语言模型的方法相比,表现有明显的提升。得益于规则的指导,RuleRAG-ICL 在查全率上提升明显。

微调后的 LawGPT 表现更佳,说明针对特定任务的微调能有效提升模型的专业性和准确性。

基于 LawGPT 的 LADDER 效果优于基于 QWen2-7B-Instruct 的效果,这是因为比起 QWen2-7B-Instruct,LawGPT 还学习了法律领域的相关知识,因此更加有助于法律文书的生成。

与微调后的 LawGPT 相比,LADDER 在所有指标上都有明显提升,证明了所提数据增强和规则指导两个模块的效果与对结果提升的重要性。

综上所述,LADDER 成功提升了大语言模型在电网企业法律文书生成任务中的表现。

2.3　消融实验

对 LADDER 的关键模块进行消融实验,结果见表 2。移除一致性限制表示在数据增强时取消对样本合理性与一致性的约束。实验表明,这 3 个模块均对模型性能有显著贡献。完整 LADDER 模型 F1 分数最高。去除数据增强后,F1 下降至 0.654,说明数据增强可丰富样本多样性和提升泛化能力,对模型帮助明显。去除规则引导后,F1 降至 0.687,表明规则引导通过提供法律规则和结构化信息,有效提升了生成内容的专业性和准确性。移除一致性限制导致性能下降,主要因为增强数据质量下降,影响模型对法律要点的准确表达。综上,数据增强和规则引导模块相辅相成,共同提升了 LADDER 在法律答辩文书生成任务中的表现。

表 2　消融实验

Tab. 2　Ablation experiment

方法	查准率	查全率	F1 值
LADDER	**0.660 ±0.005**	**0.781 ±0.006**	**0.715 ±0.005**
移除数据增强	0.628 ±0.004	0.683 ±0.007	0.654 ±0.004
移除规则引导	0.641 ±0.004	0.741 ±0.002	0.687 ±0.003
移除一致性限制	0.638 ±0.002	0.743 ±0.003	0.687 ±0.002

2.4　超参分析实验

本文对答辩观点余弦相似度匹配阈值进行了超参分析实验,比较了本文提出的方法和 3 个效果较好的对比方法,结果如图 2 所示。从图中可以看出,随着阈值升高,对能够匹配上观点的限制越强,验证阶段模型认为成功匹配的观点越少,因此验证时计算出的 F1 值越低。然而不论相似度阈值选择多少,本文方法比对比方法均取得更好的效果。

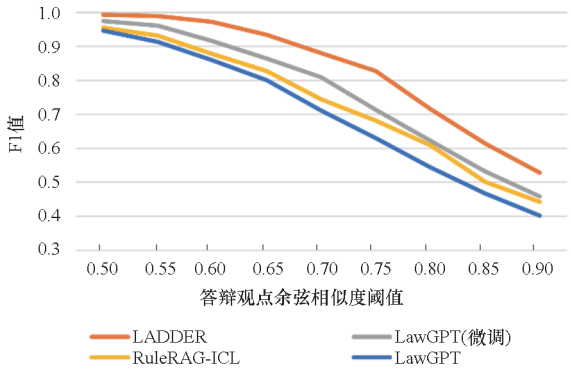

图 2　模型 F1 值与答辩观点余弦相似度阈值之间的关系

Fig. 2　Relationship between F1 score and cosine similarity threshold for evaluating opinions

对规则覆盖率阈值进行超参分析实验,结果如图 3 所示。从图中可以看出,由于种子规则本

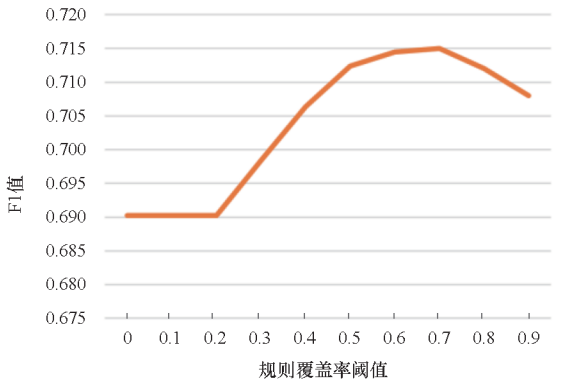

图 3　模型性能与规则覆盖率阈值之间的关系

Fig. 3　Relationship between model performance and rule coverage threshold

身可以覆盖27%的样例,因此当覆盖率阈值设置为0到0.2时模型效果不变;模型整体效果随覆盖率的提升而上升;但当覆盖率超过0.7时模型效果有所下滑,本文认为这是由于随着覆盖率的提升,大语言模型生成的规则泛化性下滑,过于个性化的规则对一些案例的引导产生了负面作用。

2.5 人工评估实验

为避免主实验中相似度可能存在的局限性,进行电网企业法律工作者的人工评估实验。首先针对每个案件的生成结果分别从3个维度进行评分:"观点有效性(生成的答辩观点是否符合法律逻辑、事实依据和行业规范,能否有效支持被告方立场)""内容完整性(生成内容是否覆盖案件关键争议点,无重要信息遗漏)""表达流畅性"。3个维度分别进行1~5分评分(1分最差,5分最佳),最终汇报测试方法在各个维度上的平均值,在Fleiss'κ系数中评估者间一致性为0.61。将本文方法与3个次优对比方法在测试集上进行比较,由3位电网企业法律工作者进行独立评分,实验结果如表3所示。从实验结果中可以看出,LADDER在观点有效性和内容完整性上均优于各对比方法。在使用QWen2-7B-instruct作为基础模型后(第3行结果),表达流畅性有一定的提升,本文认为这主要是由于相较于LawGPT,QWen2-7B-instruct通用的中文语言表达力更强。

表3 生成的法律答辩文书的人工评估实验

Tab.3 Manual evaluation experiment on generated legal defense documents

方法	观点有效性	内容完整性	表达流畅性
RuleRAG-ICL	2.45	2.94	4.29
LawGPT(微调)	2.41	2.65	4.12
LADDER基于QWen2-7B-instruct	2.63	3.10	4.31
LADDER	**2.81**	**3.33**	**4.14**

增加对规则质量的人工评估来验证本文方法的稳定性。更多的生成规则覆盖了更多的样例,在覆盖率分别达到0.3、0.5、0.7和0.9时统计所生成规则的质量。从规则合理性(如规则是否符合法律逻辑、事实依据和行业规范,能否有助于支持被告方立场)和表达流畅性两个维度进行1~5分人工评分。如表4所示,覆盖率达到0.3、0.5、0.7和0.9时,生成规则的数量分别为2、5、8和14。表中汇报了在两个指标上的平均值,Fleiss'κ系数显示在3位电网企业法律工作者间的一致性为0.58。可以看出,本文方法生成规则的质量相对稳定。覆盖率超过0.7时模型指标的下降可能是由于更加多样规则的引导使模型生成了一些超出真值参考范围的答辩观点,这些观点虽具有一定的参考价值,但不在提供的真值范围内。

表4 规则的人工评估实验

Tab.4 Manual evaluation experiment on rules

指标	覆盖率			
	0.3	0.5	0.7	0.9
规则数量	2	5	8	14
规则合理性	4.00	3.60	3.92	3.88
表达流畅性	3.67	3.67	3.63	3.74

2.6 案例分析

对模型生成的答辩观点以及自动抽取出的规则进行分析。对于生成的答辩观点,本文选择两个具体的样例进行分析,对比LawGPT(微调)和LADDER的输出文本与标准的参考文本之间的差异。案例信息如表5所示。

案例一中,LawGPT(微调)的输出包含幻觉"被告已经向当地政府部门申请了火灾事故认定书",原告的诉讼中并未提及该信息,也无事实依据,因此可能导致误导性答辩。LADDER的输出与标准答案非常接近,准确识别了火灾原因不明和原告起诉金额过高的问题,没有出现不实信息。这表明LADDER由于使用数据增强和规则引导,在信息提取和生成方面更为准确,有效避免了幻觉。

案例二中,LawGPT(微调)的输出完全偏离事实,声称李某某的死亡与被告之间存在因果关系,这与案件事实不符,显示出对案件细节的误解。LADDER的输出虽然指出"我方不是适格被告",但未提供具体原因,且后续观点存在幻觉,例如对起火和死亡原因的错误推断。本文认为这主要是因为模型在训练过程中受相似案件答辩观点的影响,进行了一定程度的推测。虽然与参考答案不符,但的确是可供采纳的答辩观点。这表明LADDER在处理复杂案件时仍有改进空间,仍需提高对细节的准确性和因果关系的理解能力,避免模型在使用类似案例训练而形成的细粒度幻觉。后续工作可以考虑通过结合因果推理或图模型引入因果推理模块,使模型能够更好地理解和

表 5　案例分析

Tab. 5　Case study

案例信息	案例一	案例二
原告诉讼	742:原告陈某某向本院提出诉讼请求:判令被告赔偿我 4.5 亩尖椒损失 31 500 元;诉讼费由被告承担。事实与理由:2017 年 6 月 19 日,被告高压线路存在安全隐患,小鸟撞到线路上起火,引起火灾,将我家辣椒全部烧毁、绝收	602:原告诉称:2016 年 4 月 21 日凌晨 1 时 30 分许,本案受害人李某某(已逝)租住房屋发生火灾(起火原因不明)。被告在事发房屋上方架设空裸导线,且一直存在安全隐患,违反了《电力设施保护条例》的规定,更没有尽到安全监督、警示教育的法定义务
参考答辩文本	本案火灾原因不明,是否是小鸟撞到被告线路上引起,没有火灾调查报告予以证明; 原告起诉的数额过高,没有法律依据	我方不是本案的适格被告,根据鉴定报告具体起火原因不明,并非由于电击引起的; 死者李某某的死亡原因与我方无关,我方不应承担责任,请求法院依法驳回原告对我方的诉讼请求
LawGPT(微调)	本案火灾原因不明,是否是小鸟撞到被告线路上引起,被告不知道; 火灾后,被告已经向当地政府部门申请了火灾事故认定书,但当地政府一直没有做出认定; 原告的损失数额没有依据,被告不同意承担责任	李某某的死亡与被告之间存在因果关系,因被告的违法行为导致李某某死亡,因此应当承担相应责任
规则分类器结果	适用规则:原告请求赔偿且无证据⟶原告起诉的金额过高,需要提交证据且给出法律依据	无适用规则,分类器结果为 0
LADDER	本案火灾原因不明,是否是小鸟撞到被告线路上引起,没有火灾调查报告予以证实; 原告起诉的数额过高,没有法律依据	我方不是本案的适格被告; 起火原因为电器线路老化所致,与我方并无因果关系; 死者李某某的死亡原因不明,我方不承担责任

处理复杂的因果关系,也可通过多任务学习,增强模型在不同法律情境下的泛化能力,使其在面对复杂案件时能更准确地生成合适的答辩文书。

规则一共有 12 条,包括人工构建的种子规则 4 条,模型自动抽取的规则 8 条。本文选择 2 条种子规则和 2 条自动抽取的规则进行分析。规则包括规则头(前提)和规则体(结论)两个部分,一条规则为:规则头→规则体。种子规则为:

1) 计划停电→被告按计划停电符合法律规定,停电程序正当,没有过错。原告收到计划停电通知后没有采取相应措施导致的损失由原告自行承担。

2)供电公司进行电表校验,且原告拒绝配合供电公司核查→原告未提供证据证明其未窃电,用电检查人员发现窃电行为可当场制止,并中止供电,此时供电公司断电程序合法,原告损失无须供电公司赔偿。

下面 2 条自动抽取的规则为:

1)不可抗力或意外事件免责→被告因不可抗力或者意外事件原因具有免责情形,被告对损害事实的发生无过错。

2)原告窃电→原告窃电导致供电问题未能及时解决,被告无违约责任,不应承担任何责任,窃电造成的损失应该由原告承担。

种子规则的规则头为原告诉求中事实的总结抽象,规则体则为对应的答辩策略。在自动规则扩充的过程中,大语言模型根据种子规则和已有案例,充分理解了规则头和规则体的形式,并根据答辩观点总结出对应的答辩策略。

3　结论

本文提出了一种基于大语言模型的少样本电网法律文书生成方法,结合数据增强与规则引导技术,旨在为电网企业生成高质量的答辩文书。针对电网法律领域的专业性和数据稀缺性,本研究从以下几个方面做出了创新性贡献:

1)数据增强与改写框架的设计:通过自动生成结构和语义相近的文本,缓解了少样本问题,提

升了数据多样性和模型训练效果。

2)基于大语言模型的自动规则抽取方法:本文提出了一种高效的自动规则抽取方法,利用少量人工标注,通过大语言模型扩充规则库并引导模型微调,提升了文书生成的准确性和合规性。

本文通过实验验证了所提方法在电网法律文书生成任务中的有效性,尤其是在少样本条件下,数据增强与规则引导方法对生成质量的提升起到了至关重要的作用。

参考文献(References)

[1] CHEN Z, MA J, ZHANG X L, et al. A survey on large language models for critical societal domains: finance, healthcare, and law[EB/OL]. (2024-11-21)[2024-12-01]. https://arxiv.org/abs/2405.01769.

[2] YANG C H, XU C J, QI Y Y. Financial knowledge large language model [EB/OL]. (2024-6-29)[2024-12-01]. https://doi.org/10.48550/arXiv.2407.00365.

[3] XU Z, WU F, ZHANG Y Y, et al. Retrieval-reasoning large language model-based synthetic clinical trial generation[EB/OL]. (2024-10-16)[2024-12-01]. https://arxiv.org/pdf/2410.12476.

[4] ARBEL I, REFAEL Y, LINDENBAUM O. TransformLLM: adapting large language models via LLM-transformed reading comprehension text[EB/OL]. (2024-10-28)[2024-12-01]. https://arxiv.org/abs/2410.21479v1.

[5] COLOMBO P, PIRES T, BOUDIAF M, et al. SaulLM-54B & SaulLM-141B: scaling up domain adaptation for the legal domain[EB/OL]. (2024-07-28)[2024-12-01]. https://doi.org/10.48550/arXiv.2407.19584.

[6] HOU A B, JURAYJ W, HOLZENBERGER N, et al. Gaps or hallucinations? Gazing into machine-generated legal analysis for fine-grained text evaluations[EB/OL]. (2024-09-23)[2024-12-01]. https://arxiv.org/abs/2409.09947v2.

[7] BARRON R C, EREN M E, SERAFIMOVA O M, et al. Bridging legal knowledge and AI: retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization [EB/OL]. (2025-02-27)[2025-05-28]. https://arxiv.org/abs/2502.20364v2.

[8] FARZINDAR A, LAPALME G. Legal text summarization by exploration of the thematic structure and argumentative roles[C]//Proceedings of the Text Summarization Branches Out, 2004.

[9] GALGANI F, COMPTON P, HOFFMANN A. Citation based summarisation of legal texts[C]// Proceedings of PRICAI 2012: Trends in Artificial Intelligence, 2012.

[10] SHNARCH E, LEVY R, RAYKAR V, et al. GRASP: rich patterns for argumentation mining[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1345-1350.

[11] ZENG L D. Leveraging large language models for code-mixed data augmentation in sentiment analysis[C]//Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024), 2024.

[12] SHIM H S, PARK H, LEE K, et al. A persuasion-based prompt learning approach to improve smishing detection through data augmentation[EB/OL]. (2024-09-06)[2024-12-01]. https://arxiv.org/abs/2411.02403.

[13] GOCERI E. Medical image data augmentation: techniques, comparisons and interpretations[J]. Artificial Intelligence Review, 2023, 56: 12561-12605.

[14] CHEN Z W, XU C J, WANG D M, et al. RuleRAG: rule-guided retrieval-augmented generation with language models for question answering[EB/OL]. (2024-10-15)[2024-12-01]. https://arxiv.org/pdf/2410.22353.

[15] WANG S Y, WEI Z Y, CHOI Y, et al. Can LLMs reason with rules? Logic scaffolding for stress-testing and improving LLMs[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 7523-7543.

[16] EDGE D, TRINH H, CHENG N, et al. From local to global: a graph RAG approach to query-focused summarization[EB/OL]. (2024-04-24)[2024-12-12]. https://arxiv.org/abs/2404.16130v2.

[17] JIMENEZ GUTIERREZ B, SHU Y H, GU Y, et al. HippoRAG: neurobiologically inspired long-term memory for large language models [EB/OL]. (2024-05-23)[2024-12-12]. https://arxiv.org/abs/2405.14831.

[18] RAM O, LEVINE Y, DALMEDIGOS I, et al. In-context retrieval-augmented language models[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1316-1331.

[19] LI J, QIAN L, LIU P F, et al. Construction of legal knowledge graph based on knowledge-enhanced large language models[J]. Information, 2024, 15(11): 666.

[20] SHU D, ZHAO H R, LIU X K, et al. LawLLM: law large language model for the US legal system[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024: 4882-4889.

[21] HU J E, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB/OL]. (2024-10-16)[2024-12-01]. https://arxiv.org/abs/2106.09685.

[22] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models [EB/OL]. (2023-07-19)[2024-12-01]. https://doi.org/10.48550/arXiv.2307.09288.

[23] YANG A, YANG B S, HUI B Y, et al. Qwen2 technical report [EB/OL]. (2024-07-15)[2024-12-12]. https://doi.org/10.48550/arXiv.2407.10671.

[24] ZHOU Z, SHI J X, SONG P X, et al. LawGPT: a Chinese legal knowledge-enhanced large language model [EB/OL]. (2024-06-07)[2024-12-01]. https://arxiv.org/abs/2406.04614.

[25] ZHANG T Y, KISHORE V, WU F, et al. BERTScore: evaluating text generation with BERT [EB/OL]. (2020-02-24)[2024-12-01]. https://arxiv.org/abs/1904.09675.

[26] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.