面向法律事件检测的大模型协同主动学习框架

崔婷婷，咎红英*，籍欣萌，宋金旺，张坤丽，贾玉祥
郑州大学计算机与人工智能学院，河南省 郑州市，450001
ttcui@stu.zzu.edu.cn, iehyzan@zzu.edu.cn, jixinmeng45@gs.zzu.edu.cn
jwsong@gs.zzu.edu.cn, ieklzhang@zzu.edu.cn, ieyxjia@zzu.edu.cn

摘要

法律事件检测任务旨在识别并分类法律文本中的事件。然而，复杂的法律案件使得收集高质量标注数据面临巨大挑战。目前领域数据标注主要依赖人工，成本高昂且耗时。尽管传统的主动学习能够减少部分标注需求，但仍依赖于人工干预。大模型的发展为自动化数据标注带来了可能性，但如何确保标注的可靠性仍是亟待解决的问题。为此，本文提出了创新的协作训练范式，使用主动学习迭代选择训练数据，并利用大模型生成高质量标注，使用评估筛选机制保留高质量标注，大幅减少了人工标注的工作量。在两个事件检测基准数据集上的实验表明，该方法在低资源场景下显著降低了人工标注需求，在部分情况下可以接近监督学习的性能。

关键词： 主动学习 ； 法律事件检测 ； 大语言模型

Leveraging Large Language Model with Active Learning for Legal Event Detection

Tingting Cui, Hongying Zan*, Xinmeng Ji, Jinwang Song, Kunli Zhang, Yuxiang Jia
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou Henan, 450001
ttcui@stu.zzu.edu.cn, iehyzan@zzu.edu.cn, jixinmeng45@gs.zzu.edu.cn
jwsong@gs.zzu.edu.cn, ieklzhang@zzu.edu.cn, ieyxjia@zzu.edu.cn

Abstract

Legal Event Detection aims to identify and categorize events in legal texts. However, the complexity of legal cases poses significant challenges in collecting high-quality annotated data. Most data annotation in some domains is currently done by hand, which is expensive and time-consuming. While traditional active learning can partially reduce the need for manual annotation, their performance remains constrained by a heavy dependence on human intervention. Recent advances in Large language models have opened up new possibilities for automated data annotation, but how to ensure the reliability of the annotations they generate remains an urgent problem. To address these challenges, we propose an innovative, collaborative training paradigm, which iteratively selects informative data using active learning and employs the generative capabilities of large language models to produce and refine high-quality annotations. An evaluation and filtering mechanism is further introduced to retain only reliable annotations, significantly reducing the need for manual labeling. Extensive experiments

* 通讯作者

第二十四届中国计算语言学大会论文集，第37页-第49页，济南，中国，2025年8月11日至14日。

on two event detection benchmark datasets demonstrate that our method substantially reduces the demand for manual annotations in low-resource scenarios and, in some instances, achieves performance comparable to supervised learning.

Keywords: Active learning , Legal event detection , Large language models

1 引言

在法律领域，事实认定是做出法律判断的最基本步骤，因此检测和识别法律文件中的事件对于法律案件的分析和决策至关重要。查找和识别法律文件中的事件对于许多法律人工智能应用来说至关重要。法律事件检测（Legal Event Detection, LED）涉及识别和标注文本中的触发词（最能代表事件发生的词）。在先前的研究中，法律事实提取通常用于辅助分析下游任务 (Zhang et al., 2024; Feng et al., 2022)。随着自然语言处理技术的发展，从法律文本中自动提取关键信息已成为支持法律决策和风险评估的重要任务。然而，尽管事实提取任务得到了大量研究，但现有的标准事件提取数据集通常规模较小，例如被研究者广泛使用的ACE2005数据集 (Doddington et al., 2004)只有8个类型和33个子类型，涵盖的事件类型和标注数据有限，并且由于法律案件的独特性和多样性，专门面向法律事件的提取方法仍然相对匮乏。在实际场景下面临着大规模的事件类型却只有少量标注数据，甚至是无标注数据的问题 (Yao et al., 2022)。

目前主流的领域事件数据集构建方法大多基于人工标注，这一过程繁琐且成本高昂，使得高质量的标注数据仍然十分稀缺、昂贵。因此，如何高效的利用有限的标注数据并降低人工标注的成本是一个重要且有意义的研究方向。主动学习（Active Learning，AL）(Settles, 2009)通过迭代选择信息量最大的样本以减少标注数据的需求，已被成功应用于文本分类 (Margatina et al., 2022)、情绪分类 (Margatina et al., 2021a)等任务，与这些序列分类任务不同，将主动学习应用于事件检测任务是一个尚未充分研究的主题，带来了独特的挑战。另一方面，主动学习方法依赖人类专家作为昂贵的监督来源 (Li et al., 2024)，限制了它的广泛应用。近来大模型（Large Language Model, LLM）在多个自然语言处理任务中展现出了卓越能力为传统数据标注方法提供了一种全新的视角 (Tan et al., 2024)。通过自动化标注任务和微调以适应特定领域，大模型在一些任务上甚至可以超过人类标注者的表现 (He et al., 2024)。但是，大模型在一致性和准确性方面能否提供正确且有用的标注数据，充当主动学习中数据标注者的角色，也是一个值得深入研究的问题。

在这项工作中，我们提出了一种新颖的协作训练范式ALLED（**A**ctive Learning with **L**arge Language Model for **L**egal **E**vent **D**etection），它通过利用主动学习的思想，迭代地查询出下一轮模型所需要的训练数据，训练LLM在特定领域上数据标注任务的能力，利用其作为传统的主动学习中的数据标注者，对新的训练数据标注完成后进入下一轮的迭代中，新标注数据的质量会在很大程度上影响模型的质量，为了更好的利用LLM标注器的标注数据，我们采用了波束搜索增强技术，引导LLM生成多个标注结果，并设计了一种伪数据筛选规则对多个标注数据进行筛选，只有符合标准的标注数据才可以进入下一轮迭代中使用。

总的来说，本文的主要贡献可以概括如下：提出了基于主动学习方法的框架ALLED，提高了低资源法律事件检测场景下的学习性能，甚至在某些情况下接近有监督学习性能；探索了使用LLM作为主动学习中的数据标注者的性能，并设计不同的人工标注数据占比来分析LLM的表现对事件检测任务的影响；在两个广泛使用的事件检测基准上进行了实验，评估了三类基线方法以及ALLED框架下的多种查询策略。

2 相关工作

由于标注数据的匮乏，模型的训练和优化往往面临瓶颈。为应对这一挑战，主动学习方法应运而生，其旨在从数据池中选择信息丰富的示例，以在所需的数据预算下最大化性能，或最小化数据预算以实现所需的性能。在自然语言处理领域，主动学习已成功应用于资源匮乏环境下的语言模型的优化 (Dor et al., 2020)，多数研究集中于情感分类 (Margatina et al., 2021a)、文本分类 (Schröder et al., 2023)和命名实体识别任务 (Vacareanu et al., 2024; Radmard et al., 2021)。在主动学习的每次迭代中，模型会使用查询策略在数据中选择最具信息量的数据来进行

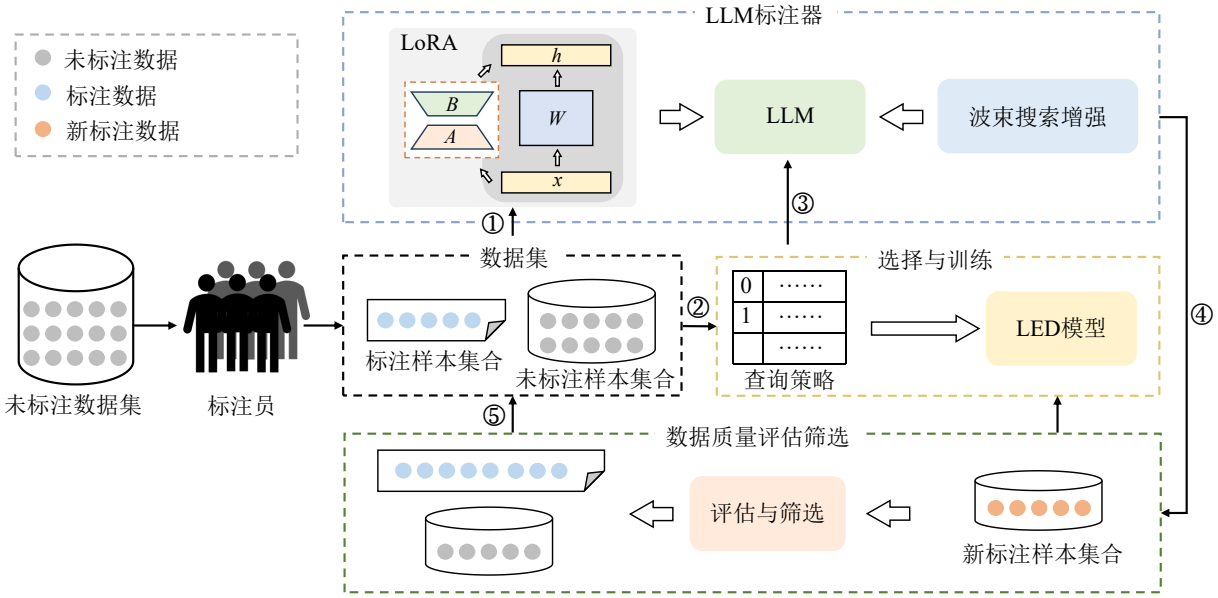


Figure 1: ALLED整体架构：利用LLM进行法律事件检测的主动学习

标注。主动学习的传统查询策略通常分为基于不确定性的方法 (Prabhu et al., 2019; Margatina et al., 2021b)和基于多样性的方法 (Ash et al., 2020)。本文参考Vacareanu (2024)的研究，将主动学习应用于低资源的法律事件检测任务，并探索各种查询策略对模型性能的影响，以增强法律事件检测模型的有效性。

最近，大模型在多项自然语言处理任务中展现出卓越的零样本和少样本能力 (Zhao et al., 2023)。一些工作研究了在缺乏特定于任务的数据时LLM的零样本学习能力 (Savelka and Ashley, 2023)。另一些研究关注上下文学习（in-context learning）的应用 (Dong et al., 2024)，通过少量输入实例作为演示，已经显示出有希望的少样本性能 (Zhang et al., 2022)。此外，针对特定域的微调 (Devlin et al., 2019)也可以显著减少传统标注方法面临的挑战。目前的事件检测数据集大多基于广泛使用的ACE（Automatic Content Extraction）事件模式进行人工标注。一些工作通过自建数据集 (Gao et al., 2024)进行实验，或者对原本的公开数据集进行扩充 (Parekh et al., 2023)，以及在低资源语言 (Touileb et al., 2024)、低资源领域 (Ma et al., 2023)上构建数据集。但是这些数据集的构建严重依赖于人类专家作为昂贵的标注来源。因此高质量的公开标注数据仍然十分稀缺、昂贵。为了降低费用，一些工作以牺牲数据质量为代价，使用成本较低的ChatGPT作为数据标注源 (Gilardi et al., 2023)。通过引入数据质量筛选机制，可以进一步提升标注数据的可靠性，目前主流的质量筛选机制可分为基于规则的方法 (Zheng et al., 2023; Kim et al., 2023)、基于外部源的方法 (Dou et al., 2024)和LLM驱动的方法 (Wang et al., 2023; Lu et al., 2023)。本研究尝试利用LLM的丰富知识作为低成本监督的来源，在无需人力的情况下保证标注数据的准确性和一致性，提高模型的泛化性能。

3 主要方法

本节详细介绍了ALLED框架，该方法探索了大模型与主动学习策略相结合在法律事件检测任务中的应用。整体流程如图1 所示。与传统主动学习方法需要在每轮训练迭代中依赖人工标注不同，ALLED 充分利用了LLM 作为自动标注器的强大能力，从而显著降低了人工成本。在该框架中按照以下步骤依次执行：

1. 首先，人工标注一小部分高质量事件数据，作为初始训练集，并基于该数据训练一个事件检测任务的大模型标注器，模拟人工标注行为。

2. 在主动学习框架下，采用多种查询策略从未标注数据中选取当前迭代的候选样本。

3. 使用已构建的大模型标注器对所选候选样本进行自动标注，获取对应的事件触发词与类型。

4. 引入质量控制机制，对大模型自动生成的标注结果进行质量筛选，过滤掉预测不确定或与标注规范偏离过大的样本。

5. 将筛选后的新标注样本加入已有训练集中，更新用于训练事件检测模型的数据集。

6. 重复执行第2至第5步，通过多轮主动学习迭代不断优化训练数据与模型性能，直至无更多高质量样本可用或达到预设的迭代次数上限。

3.1 查询策略

本研究采用了三种广泛使用的基于不确定性的查询策略，以及一个随机查询的基线方法。在样本选择过程中，针对每种查询策略，将结果列表按得分升序排序，并选取预设数量的数据用于标注。各查询策略的具体说明如下。

Random (RD)：无论模型的预测如何，它都会以随机方式获取要标注的数据标注。随机查询策略不需要复杂的计算。

Breaking Ties (RT)：根据模型预测结果，选择前两个预测概率差距最小的实例进行标注。具体来说，通过以下公式来选择数据x_i：

$$\arg\min_{x_i} [P(y_i = l_1 \mid x_i) - P(y_i = l_2 \mid x_i)] \tag{1}$$

其中，l_1和l_2分别是最可能的标签和第二可能的标签。

Least Confidence (LC)：根据模型预测结果，挑选对预测标签置信度最低的实例进行标注。具体通过以下公式来选择数据x_i：

$$\arg\max_{x_i} [1 - P(y_i = l_1 \mid x_i)] \tag{2}$$

其中l_1是最可能的标签。

Prediction Entropy (PE)：根据模型的预测，选择标签分布熵值最高的样本进行标注，以降低整体预测熵。具体通过以下公式来选择数据x_i：

$$\arg\max_{x_i} \left[-\sum_{j=1}^{c} P(y_i = j \mid x_i) \log P(y_i = j \mid x_i) \right] \tag{3}$$

其中c是x_i所有可能的预测标签的数量。

3.2 大模型标注器

大模型的兴起激发了人们对其在生成高质量、具备上下文感知能力的标注数据方面的广泛关注。微调是将LLM应用于特定下游任务的常见做法，但由于模型参数规模庞大，全面微调的代价十分高昂。研究人员提出了参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）方法，其核心思想是在保持主干模型参数冻结的前提下，仅调整一小部分外部参数，从而获得相当或更优的性能。当前主流的PEFT技术包括基于适配器的方法、前缀调优和低秩自适应微调（Low-Rank Adaptation，LoRA） (Hu et al., 2022)等。其中，LoRA由于其优越的性能和良好的兼容性，成为应用最广泛的PEFT方法。本文主要采用LoRA作为参数高效微调的实现方式。

LoRA假设模型权重矩阵的更新可以通过低秩分解进行近似，即将权重更新表示为两个低秩矩阵的乘积，从而有效降低可训练参数的数量。

$$\Delta W = \alpha BA \tag{4}$$

其中，ΔW是模型权重矩阵的更新，$B \in \mathbb{R}^{d\times r}$和$A \in \mathbb{R}^{r\times k}$是秩为$r$的矩阵，$\alpha$是常数缩放因子。在训练过程中，仅对$\Delta W$进行优化，而模型权重矩阵$W$保持不变。应当注意，$A$是随机初始化的，而$B$则初始化为零。在训练开始时，模型状态满足$W + \Delta W = W$，与原始模型参数一致。

3.3 数据质量评估筛选

在ALLED框架的每次训练迭代中，都将使用由LLM标注器生成的数据，而新标注数据的质量对LED模型的训练效果有着重要影响。因此，有效评估并筛选LLM生成的标注对于充分发挥其潜力至关重要。为解决该问题，本文设计了一种基于波束搜索增强技术的伪数据筛选规则。传统的问题生成方法通常只能生成一个概率最高的问题，而波束搜索增强技术则能够引导LLM生成多个候选标注结果供进一步使用。具体来说，在波束搜索的每一步，它都会跟踪当前概率最高的N_{beam}序列，N_{beam}是波束大小。对于句子s_i，它可以生成一组候选标注$A_i = \{(s_i, e_{i1}, r_{i1}), (s_i, e_{i2}, r_{i2}), \ldots, (s_i, e_{in}, r_{in})\}$，$e_{ij}$表示为$s_i$生成的第$j$个候选标注，$r_{ij}$是候选标注$e_{ij}$的得分。根据$r_{ij}$的得分，$A_i$中的标注结果按降序排序。候选标注的得分$r_{ij}$计算如下：

$$P(e_{ij} \mid s) = \sum_{t=1}^{T} \log P(x_t \mid x_{<t}, s_i) \tag{5}$$

$$r_{ij} = \exp(P(e_{ij} \mid s_i)) \tag{6}$$

其中x_t是生成的第t个数据。$P(x_t \mid x_{<t}, s)$是生成模型输出的条件概率。

过滤标准依赖于标注结果的可信度和独特性。具体来说，当一个候选标注的最高置信度得分r_{i1}超过预定义的阈值τ_1，且最高得分r_{i1}与第二高得分r_{i2}之间的差值超过另一个阈值τ_2时，该标注被认为是有效的。具体筛选规则如下：

$$D = \{s_i \in S \mid r_{i1} \geq \tau_1 \text{ and } r_{i1} - r_{i2} \geq \tau_2\} \tag{7}$$

其中，S表示根据查询策略选择的句子集。经过筛选后，D是最终从S中选出的句子集。当前训练迭代所需的标注句子数量（用k表示）等于D中的句子数量。

筛选标准旨在确保保留下来的标注既具有可靠性，又能与其他候选标注明显区分。若当前选中的标注数据未达到质量要求，则继续从剩余数据中选择样本进行标注，直到获取满足下一轮迭代所需的训练数据为止。

4 实验

4.1 数据集

我们在广泛使用的事件检测数据集ACE05-C (Doddington et al., 2004)和LEVEN (Yao et al., 2022)上开展实验。ACE05-C包含了33种事件类型和599份从不同来源的文档。尽管该数据集属于通用领域，但其中包含的13 种法律相关事件类型对法律事件检测任务具有重要意义。LEVEN是目前中国规模最大的法律事件检测数据集，包含8116个法律文档和108种事件类型。表1展示了两个数据集的详细统计信息。

数据集	ACE05-C	LEVEN
句子数量	7,955	63,616
事件类型	33	108
事件提及	4090	150,977

Table 1: 事件检测数据集的统计信息

4.2 基线与评估

基线方法划分为以下三组：

- 上下文学习：使用LLMs分别进行少样本的上下文学习，示例从训练集中随机选取。由于示例中并未涵盖所有可能出现的事件类型，实验中还比较了将LoRA应用于这些LLMs的结果。在模型选择上，选取了以下几种主流的大模型进行测试：（1）闭源模型：由OpenAI发布的ChatGPT (Brown et al., 2020)系列模型，具备强大的语言理解与生成能力，广泛应用于各类自然语言处理任务。（2）开源模型：阿里云推出的大规模开源语言

模型Qwen2.5-32B和Qwen2.5-7B (Yang et al., 2024)；书生·浦语团队开发的专门针对国内企业级使用场景设计并优化的第二代开源大语言模型InternLM2-7B (Cai et al., 2024)；百川智能发布的第二代开源大语言模型Baichuan2-7B (Yang et al., 2023)；DeepSeek团队发布的基于Qwen架构蒸馏优化的DeepSeek-R1-Distill-Qwen-7B。

- 监督微调：使用LEVEN (Yao et al., 2022)中提供的模型用于微调。所用的训练数据比例与ALLED中使用的数据比例相对应，即30%的训练数据。另外还提供了使用完整数据进行监督微调的结果作为参考值。
  - 分类：使用深度神经网络对输入句子进行编码，包括了预训练语言模型BERT和RoBERTa。编码后提取每个候选触发词的隐藏表示，并将其输入至分类层，以预测其对应的事件类型。该过程本质上是一个Token级别的多类分类任务，旨在识别文本中所有潜在的事件触发词及其类型。
  - 序列标注：引入了序列标注方法来更有效地建模句子中不同事件触发词之间的上下文依赖与相互关系。通过将条件随机场层叠加于编码器之上，模型能够捕捉标签序列间的结构信息，从而提升整体标注性能。
- 采用随机策略的ALLED：随机策略是主动学习中最简单的查询策略。

在评估中使用了与Li (2013)的工作中相同的标准：如果触发词的偏移量与参考触发词一致（Trig-I），则触发词被正确识别。如果触发词的事件类型与参考触发词的事件类型一致（Trig-C），则触发词被正确分类。我们报告Trig-I和Trig-C的F1分数。所有实验均使用不同的随机种子进行重复，结果展示为最终的平均值和标准差。

4.3 实验设置

本实验参考了Vacareanu (2024)工作中的实验设置，并根据计算资源的限制对模型规模和训练策略进行了适当调整。具体而言，LED模型采用了参数量为107M的BERT-base-chinese，并使用AdamW优化器进行训练。在主动学习设置中，每个数据集共进行30轮迭代，每轮从未标注数据集中选取1%的样本加入到训练集中。迭代结束时，累计约有30%的样本用于训练LED模型。在LLM标注器部分，选用Qwen2.5-7B-Instruct作为基础模型，并通过LoRA微调方法在测试集上进行轻量化适配。对于LLM的上下文学习、LoRA以及标注器的训练，均统一采用表2中的数据格式。LoRA微调时，r设置为64，缩放因子设置为16。每条样本由系统提示（System）、用户输入（User）以及模型返回（Assistant）三部分组成。其中，系统提示用于限定模型的角色和任务，用户输入为待标注的原始文本，模型返回则为结构化的事件标注结果，包含被识别的事件触发词及其对应的事件类型。

System:
你是一个语言学家，你需要标注出一段文本中所有的事件触发词及其事件类型。

User:
2016年5月19日，被告人孙金才主动到检察机关投案，并如实供述了自己的罪行

Assistant:

```
{
  "sentence":  "2016年5月19日，被告人孙金才主动到检察机关##投案@@，并
  如实##供述@@了自己的罪行",
  "type":  ["投案", "供述"]
}
```

Table 2: LLM训练数据示例

在数据质量评估与过滤模块中，设置$N_{beam} = 3$，$\tau_1 = 0.95$，$\tau_2 = 0.05$，用于筛选高质量的自动标注样本。为提升结果的可靠性，每组实验均在随机种子为2、4、6、8的设置下重复四次，最终报告平均结果及标准差。其余训练超参数包括：学习率设置为5×10^{-5}，$epoch = 10$，并启用提前停止机制，若连续3个epoch验证集性能无提升，则提前终止训练。

5 实验结果与分析

5.1 主要实验结果

表3展示了三组不同的基线方法与ALLED在ACE05-C与LEVEN两个数据集上的实验结果，并提供了使用完整训练集进行监督微调的结果作为参考。除该参考结果外，其余实验均仅使用了30%的训练数据。表中数值表示多个随机种子实验的均值及其标准差，标准差以下标括号形式给出，最佳结果以粗体标出。

Models		ACE05-C		LEVEN	
		Trig-I	Trig-C	Trig-I	Trig-C
GPT-3.5$_{\text{5 shot}}$		$10.83_{(0.30)}$	$7.75_{(0.49)}$	$14.83_{(0.11)}$	$12.84_{(0.11)}$
GPT-4o$_{\text{5 shot}}$		$35.12_{(0.24)}$	$25.71_{(0.33)}$	$32.80_{(0.13)}$	$28.24_{(0.17)}$
Qwen2.5-32B$_{\text{5 shot}}$		$21.10_{(0.34)}$	$16.09_{(0.55)}$	$28.14_{(0.06)}$	$22.71_{(0.07)}$
Qwen2.5-7B$_{\text{5 shot}}$		$3.23_{(0.64)}$	$2.73_{(0.56)}$	$12.49_{(1.04)}$	$8.77_{(1.02)}$
Qwen2.5-7B$_{\text{LoRA}}$		$31.04_{(0.24)}$	$27.40_{(0.11)}$	$85.76_{(0.03)}$	$81.68_{(0.03)}$
InternLM2-7B$_{\text{LoRA}}$		$25.11_{(0.18)}$	$22.18_{(0.65)}$	$84.66_{(0.05)}$	$79.68_{(0.04)}$
Baichuan2-7B$_{\text{LoRA}}$		$40.32_{(1.16)}$	$35.65_{(0.67)}$	$83.69_{(0.04)}$	$79.15_{(0.08)}$
DeepSeek-R1-Qwen-7B$_{\text{LoRA}}$		$22.72_{(0.71)}$	$16.37_{(0.21)}$	$83.05_{(0.03)}$	$78.38_{(0.06)}$
BERT		$68.50_{(1.64)}$	$57.48_{(1.32)}$	$86.61_{(0.07)}$	$82.29_{(0.08)}$
RoBERTa		$69.09_{(1.69)}$	$58.43_{(1.64)}$	$86.66_{(0.16)}$	$82.23_{(0.14)}$
BERT+CRF		$68.90_{(0.40)}$	$57.78_{(1.43)}$	$86.48_{(0.14)}$	$82.22_{(0.20)}$
ALLED(Ours)	RD	$64.32_{(3.95)}$	$60.76_{(4.04)}$	$87.06_{(0.21)}$	$81.71_{(0.37)}$
	LC	$67.86_{(1.85)}$	$63.79_{(2.46)}$	$\mathbf{87.69_{(0.14)}}$	$82.37_{(0.21)}$
	PE	$68.54_{(2.79)}$	$65.00_{(2.68)}$	$87.22_{(0.33)}$	$82.21_{(0.12)}$
	BT	$\mathbf{69.14_{(0.91)}}$	$\mathbf{66.02_{(1.08)}}$	$87.55_{(0.05)}$	$\mathbf{82.49_{(0.20)}}$
BERT(full)		$75.03_{(0.62)}$	$64.32_{(1.30)}$	$87.55_{(0.21)}$	$83.80_{(0.26)}$

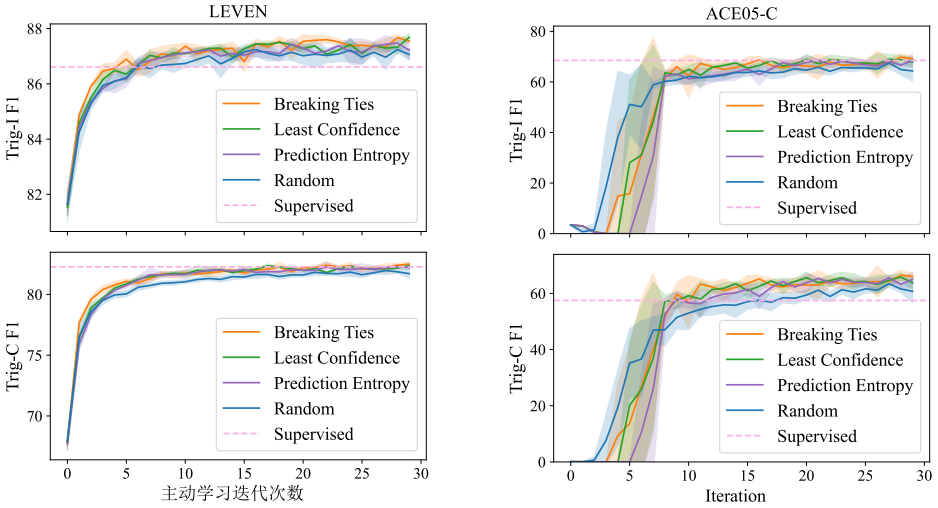
Table 3: 不同基线以及ALLED的实验结果

在这些基线中，大模型在基于上下文学习的方法的表现整体较弱。这表明尽管大模型具有强大的语言理解能力，用有限的示例定义有效的提示来促使LLM进行LED仍然具有挑战性，尤其在结构化输出要求较高的任务中略显不足。

在对大模型进行LoRA 微调后，性能显著提升，特别是在LEVEN 数据集上，接近部分预训练语言模型的表现。这表明，在有限的训练样本下，通过高效的参数微调策略，可以充分激发大模型在事件识别任务中的潜力。然而，LoRA 调优的大语言模型在ACE05-C 数据集上的表现低于监督微调的基线。此差异可能与标签语言的不同有关。尽管ACE05-C 的文本为中文，但标签却采用英文形式，这种语言不一致性可能影响了模型的理解和预测能力。LoRA 微调通常依赖于语言特定的模式，在处理带有英文标签的中文文本时，模型可能未能有效捕捉到标签与文本之间的语义映射。而监督微调方法直接将模型输出映射到预定义的标签，从而避免了语言不一致性所带来的影响。BERT和RoBERTa作为主流预训练语言模型，整体表现稳定且较强。另外，尽管尝试利用条件随机场捕捉标签序列间的结构信息，但结果并不符合预期。BERT+CRF在两个数据集上的表现均略低于BERT模型。

本文所提的ALLED方法在多个设置下都取得了优异成绩，尤其是在LEVEN数据集上，BT策略下的ALLED超越了多数基线模型。在ACE05-C数据集上，BT策略同样在这些对比结果中取得最高的效果，进一步验证了主动学习策略对模型性能的积极影响。ALLED方法将大模型引入主动学习循环中作为自动标注器，通过数据质量控制机制，有效降低了低质量标注样本对模型训练的负面影响，缓解了高质量人工标注数据稀缺的问题并确保了主动学习过程具有良好的稳定性与鲁棒性。另外，通过引入基于不确定性的查询策略，优先选择信息增益更高的样本，使模型在有限迭代中获得更强泛化能力。但是，当使用全部训练数据时，ALLED与

其相较仍有一定的差距，说明主动学习在高质量数据受限的场景下具有明显优势，而在数据充足的情形下，传统监督学习仍然具备强竞争力。



(a) LEVEN (b) ACE05-C

Figure 2: ALLED在四种查询策略下的学习曲线

接下来，我们进一步探讨了模型的性能如何随着主动学习的迭代次数的增加而变化。图2展示了针对不同数据集、查询策略以及评估指标下的学习曲线。整体来看，与其他三种基于不确定性的查询策略相比，随机查询的基线的性能提升过程明显更为平缓，最终达到的F1分数也相对较低。这一结果验证了在主动学习框架中，优先选择信息量更高的样本用于训练，确实能够更高效地提升模型性能。另外，在ACE05-C数据集上，所有策略的学习曲线都表现出较大的波动性。推测这一现象主要与该数据集本身的特点有关：一方面，其整体规模相对较小，训练样本有限；另一方面，事件在数据样本中的密度较低，图3展示了两个数据集中触发词在句子中的分布情况。ACE05-C数据集中超过70%的样本没有触发词，这表明该数据集中的事件密度较低。这一特点使得模型在每次迭代中获取的信息差异较大，从而导致预测性能的波动性。这一发现进一步强调了在小规模、高稀疏度的数据场景中，合理设计主动学习策略与样本筛选机制的重要性。

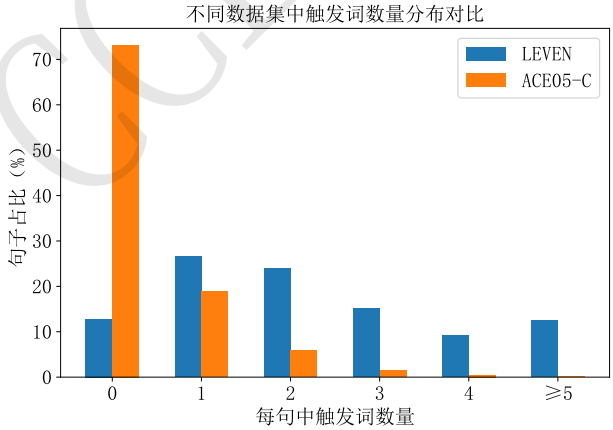


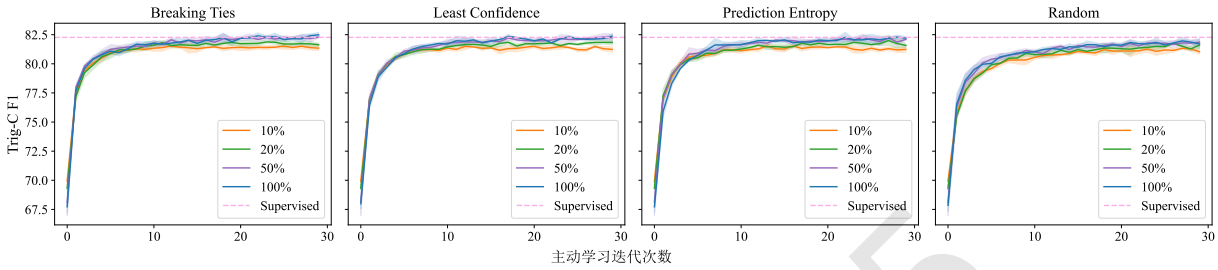
Figure 3: 触发词数量分布对比

5.2 LLM标注器性能

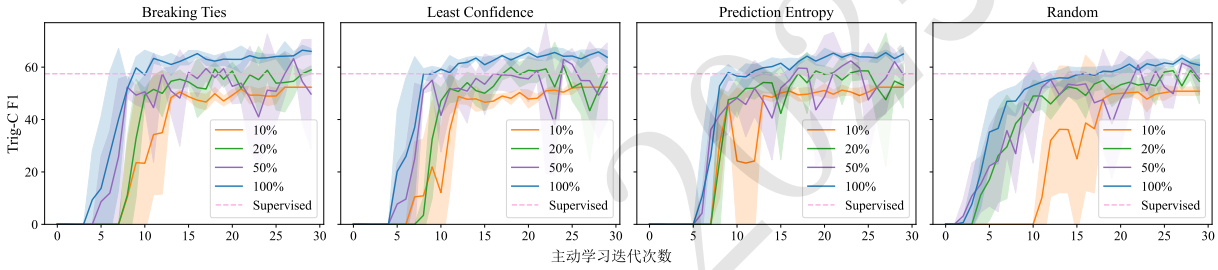
在本节中，比较了ALLED在LEVEN和ACE05-C两个数据集上使用不同占比的人工标注数据对性能的影响。图4分别显示了四种查询策略下，人工标注数据在不同占比时Trig-C指标的学习曲线。随着人工标注数据的减少，意味着LLM标注器的工作量也相应增加，LLM标注器在LEVEN数据集上展示了最显着的改进，减少近50%的人工标注就可以获得与全量标注相当

甚至更佳的性能，如果模型降低后的性能仍在可接受的范围内，人工标注的工作甚至可以减少到20%或10%。

另外，对于四个不同的选择策略，虽然策略的具体机制存在差异，但这些不同策略下的模型性能随着人工标注数据的减少呈现出相似的性能变化，所以说不同的选择策略下并不会显著影响标注器的整体性能。这表明，标注器的性能主要受标注数据质量和数量的影响，而非具体的样本选择策略。相较于LEVEN上的表现，LLM标注器在ACE05-C中的表现则相对较差些，这可能是因为该数据集的原始规模较小，LLM标注器可利用的训练数据也相对较少，影响到了后续的训练数据的标注质量。受限于初始标注的准确性，模型在前几轮主动学习迭代中难以有效学习，直到平均第5轮迭代之后，模型性能才开始逐步提升。这一现象表明，在数据资源相对稀缺的场景下，LLM标注器的性能较为依赖初始数据的质量和数量，突显了主动学习策略在提升标注效率和模型性能中的重要性。



(a) LEVEN



(b) ACE05-C

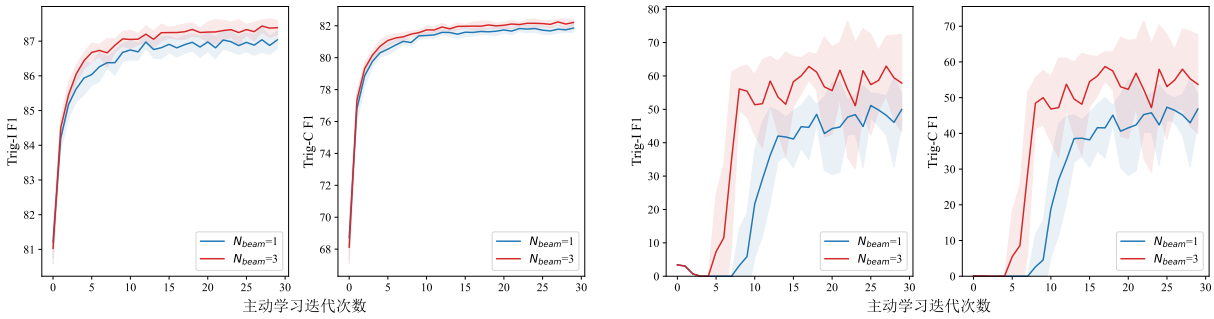
Figure 4: 不同标注比例对各个查询策略的影响

5.3 数据质量控制

本节旨在深入分析ALLED框架中数据质量评估与过滤机制对模型性能的具体影响。为了确保实验的公平性，统一采用了50%的人工标注比例进行训练，并选取三种基于不确定性的查询策略下的平均F1分数作为评估指标，以尽可能消除不同主动学习策略带来的干扰。如图5所示，当N_{beam}设置为1，即跳过质量筛选步骤，直接采用LLM的原始标注结果作为训练数据时，F1分数出现了显著下降。这一结果表明，未经筛选的自动标注数据可能存在较多噪声，从而影响模型的学习效果。相比之下，引入数据质量控制机制后，训练数据的准确性得以提升，有助于模型更有效地学习，从而显著改善ALLED框架的整体表现。这进一步验证了在自动标注驱动的主动学习框架中，质量评估与筛选机制的重要性。

此外，为了评估数据选择机制中两个关键阈值参数对模型性能的具体影响，我们在不同组合设置下进行了系统实验，相关结果如表4所示。

其中，τ_1反映了标注器对标注结果的自信程度，其取值越高，表示模型越确信当前预测具有较高可信度；τ_2则衡量标注结果的确定性程度，取值越大表明预测结果在多个候选中更具唯一性和明确性。实验结果表明，τ_1和τ_2的取值并非越高越好。具体而言，较高的τ_1虽然有助于提升标注样本的整体质量，但也可能导致可选样本数量骤减，从而影响模型泛化；而τ_2若设定过高，则可能过度过滤边界样本，降低数据多样性，使模型难以学习，反而对模型训练的性能产生负面影响。因此，为二者设置合理的阈值，综合考虑置信度与确定性，以实现高质量数据选择与高效模型学习之间的最佳平衡。



(a) LEVEN (b) ACE05-C

Figure 5: 不同N_{beam}取值下的F1值变化趋势图

数据集		LEVEN		ACE05-C	
τ_1	τ_2	Trig-I	Trig-C	Trig-I	Trig-C
0.92	0.05	$87.35_{(0.29)}$	$82.16_{(0.32)}$	$55.55_{(15.48)}$	$51.86_{(13.60)}$
0.95	0.05	$87.39_{(0.21)}$	$82.21_{(0.19)}$	$\mathbf{57.87_{(14.65)}}$	$\mathbf{53.72_{(13.86)}}$
0.98	0.05	$87.40_{(0.42)}$	$82.27_{(0.27)}$	$50.85_{(17.45)}$	$46.85_{(17.08)}$
0.95	0.02	$\mathbf{87.66_{(0.14)}}$	$\mathbf{82.40_{(0.14)}}$	$57.00_{(9.42)}$	$53.15_{(8.94)}$
0.95	0.08	$87.26_{(0.27)}$	$82.18_{(0.34)}$	$56.06_{(12.87)}$	$52.63_{(11.20)}$

Table 4: 不同τ_1和τ_2取值对F1值的影响

6 结论

在这项工作中，提出了一种基于LLM的新型协同训练范式ALLED，旨在解决低资源法律事件检测任务中的数据标注问题。该框架利用LLM进行数据标注，并引入质量检测机制筛选出有价值的样本，通过协作训练的方式实现事件检测模型的迭代优化。实验结果表明，ALLED在两个广泛使用的事件检测数据集上取得了显著的效果。此外，我们还进一步探讨了人工标注数据占比对ALLED框架性能的影响，结果表明，ALLED能大幅减少对高质量人工标注数据的依赖，并且在有限的人工监督下仍能获得显著的性能提升。

致谢

感谢审稿人帮助改善论文的建议。本文受国家自然科学基金（U23A20316）资助。

参考文献

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: an empirical study. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7949–7962.

Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. 2024. Re-rest: Reflection-reinforced self-training for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15394–15411.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664.

Qiang Gao, Zixiang Meng, Bobo Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024. Harvesting events from multiple sources: Towards a cross-document event extraction paradigm. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1913–1927.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, et al. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2024. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. 2023. Self: Language-driven self-evolution for large language model. *arXiv preprint arXiv:2310.00533*.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023. Dice: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021a. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021b. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663. Association for Computational Linguistics, November.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686.

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4058–4068. Association for Computational Linguistics, November.

Puria Radmard, Yassir Fathullah, and Aldo Lipani. 2021. Subsequence based deep active learning for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4310–4321.

Jaromir Savelka and Kevin D Ashley. 2023. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6:1279794.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. Small-text: Active learning for text classification in python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95.

Burr Settles. 2009. Active learning literature survey.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Samia Touileb, Jeanett Murstad, Petter Mæhlum, Lubos Steskal, Lilja Charlotte Storset, Huiling You, and Lilja Øvrelid. 2024. Eden: A dataset for event detection in norwegian news. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5495–5506.

Robert Vacareanu, Enrique Noriega-Atala, Gus Hahn-Powell, Marco A Valenzuela-Escárcega, and Mihai Surdeanu. 2024. Active learning design choices for ner with transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 321–334.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. Leven: A large-scale chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148.

Yuxin Zhang, Songlin Zhai, Yuan Meng, Sheng Bi, Yongrui Chen, and Guilin Qi. 2024. Event is more valuable than you think: Improving the similar legal case retrieval via event knowledge. *Information Processing & Management*, 61(4):103729.

第二十四届中国计算语言学大会论文集，第37页-第49页，济南，中国，2025年8月11日至14日。

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.