专题 大模型训练数据

面向类案检索的法律大模型评测数据集

王怡文[1,2]，戚肖克[3*]，赵小兵[1,2]，马传连[2,4]，陈波[1,2]

1. 中央民族大学信息工程学院，北京 100081
2. 国家语言资源监测与研究少数民族语言中心，北京 100081
3. 中国政法大学法治信息管理学院，北京 102249
4. 中央民族大学中国少数民族语言文学学院，北京 100081

摘要：近年来，大语言模型在法律领域的应用评测成为一大热点。类案检索是法律领域重要的任务之一，然而目前的法律大模型评测方法缺乏对类案检索任务的深度评估，没有对类案检索子任务进行细分。为了充分探索大语言模型在法律类案检索相关任务上的性能，本研究根据最高人民法院指出的类案定义，将类案检索评测分解为事件检测、裁判文书事实生成、触发词提取、关键词提取、法律长文本摘要、争议焦点识别、裁判文书说理生成 7 种子任务。针对每个子任务收集了开放的源数据集，经过预处理、重新构建、整合等步骤，构建了面向类案检索任务的大语言模型评测数据集，其数据结构包含每个法律子任务的名称、法律文本（针对任务抽取刑法、民法判决文书相关部分）、标签（任务目标输出内容）等信息。本数据集整合了 7 种类案检索子任务，共计 2424 条数据。为保证数据质量和多样化，采用 5 种大语言模型进行评估，实验表明，效果较好的评估指标达到 70%以上。本数据集的潜在利用价值在于评估模型法律多任务的处理能力，为选取类案检索处理任务效果更好的模型提供评估数据。

关键词：大语言模型；法律应用评测；法律多任务数据集；类案检索

数据库（集）基本信息简介

数据库（集）名称	面向类案检索的法律大模型评测数据集
数据通信作者	戚肖克（qixiaoke@cupl.edu.cn）
数据作者	王怡文，戚肖克，赵小兵
数据时间范围	2022–2024 年
数据量	11.3 MB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.17529
基金项目	国家社科基金重大项目（22&ZD035）
数据库（集）组成	数据集共包括 2 个数据文件，其中：Case_Retrieval_Datasets.json 文件是收集并预处理的多种任务数据集，用于大语言模型的评估测试集，数据量 3.73 MB。Case_Retrieval_Datasets_llm.json 文件是增加了大语言模型预测后的综合数据集，数据量 7.66 MB。

www.sciengine.com/CSD

ISSN 2096-2223

CN 11-6035/N

文献 CSTR：

32001.14.11-6035.csd.2024.0199.zh

文献 DOI：

10.11922/11-6035.csd.2024.0199.zh

数据 DOI：

10.57760/sciencedb.17529

文献分类：计算机科学技术

收稿日期：2024-11-27

录用日期：2025-02-12

发表日期：2025-04-30

* 论文通信作者

戚肖克：qixiaoke@cupl.edu.cn

引　言

在数据科学领域，法律相关数据集的研究和处理具有重要意义。随着司法信息化建设的不断推进，法律文本数据如裁判文书、法律法规、司法解释等迅速增长，为司法人工智能提供了丰富的数据资源[1]。

大语言模型在处理各种任务能力方面已经取得突破性的进展，有效利用模型的通用性进行法律多任务文本处理受到广泛的关注，并且法律数据具有专业性强、结构复杂等难点，此类数据标注困难、成本高昂[2-4]。因此，综合收集多任务法律文本数据对大语言模型测评具有重要意义。

目前已经有大语言模型在法律应用上的评估基准，比如 lawbench[5]，构建了 20 个子任务全面评估大模型在法律领域的理解和应用能力。然而，该方法评测的应用比较宽泛，没有针对某一任务做深入的探究。本研究针对类案检索任务，以最高人民法院对类案的定义作为依据，按照基本事实、争议焦点、法律适用性三个主要的思路，对类案检索任务细分为事件检测、触发词提取、争议焦点识别、裁判文书说理生成、裁判文书事实生成、法律长文本摘要和关键词提取等子任务[6]。其中，事件检测、触发词提取、法律长文本摘要和关键词提取、裁判文书事实生成主要针对的基本事实部分的评估、争议焦点识别针对的是争议焦点的评估、裁判文书说理生成针对法律适用性内容的评估。这些任务的实现，依赖于多个法律判决文本数据集的构建和优化，如 LEVEN[7]、ELAM[8]、EPM[9]、CAIL2024[10]等。

本数据集的利用价值在于评估模型类案检索子任务的处理能力，为类案检索任务大语言模型选择提供评估数据。

1　数据采集和处理方法

1.1　数据采集方法

针对事件检测和触发词提取、焦点识别、裁判文书说理生成、裁判文书事实生成、法律长文本摘要和关键词提取任务，本研究分别收集了如下源数据集。

（1）Ma Y 等人[7]提供的 LEVEN 数据集。LEVEN 是一个大规模的中文法律事件检测数据集，包含丰富的法律事件类型和触发词。

（2）Yao F 等人[8]提出的 ELAM 数据集。此数据集是一个可解释法律案例匹配数据集，提供了法律争议的标注数据。

（3）CAIL2024 判决文书说理数据集[10]。用于裁判文书说理生成任务，包含法院判决说理部分。

（4）CAIL2024 判决文书事实数据集[10]。用于裁判文书事实生成任务，记录了案件的事实情况。

（5）Feng Y 等人[9]提供的 EPM 数据集。用于关键词提取任务，包含法律文档的核心内容。

1.2　数据预处理

数据处理过程如图 1 所示。整个数据的处理过程为：首先去除无用信息。本研究收集的是类案检索及相关子任务数据集，而各个源数据集中包含与任务无关的字段，例如 title、id 等，因此，此步骤主要从源数据中去除无用字段。然后，截取固定数量的信息，如表 1 所示。本数据集主要任务是面向类案检索任务的法律大模型评测，抽样数据测试即可，不需要全量数据。如果原始数据集数据量很大，此步骤截取数据集为固定数量。另外，争议焦点源数据集中没有是、否有争议焦点标注

的，本数据集手动标注了是、否字样，以便于评估。然后，面向子任务分别解析各自任务字段，进行数据处理，将在第 1.3 节详细介绍。之后，对于每个子任务，选用 5 种大语言模型进行测试。最后，标准化文本格式，按照 json 文件格式进行调整。

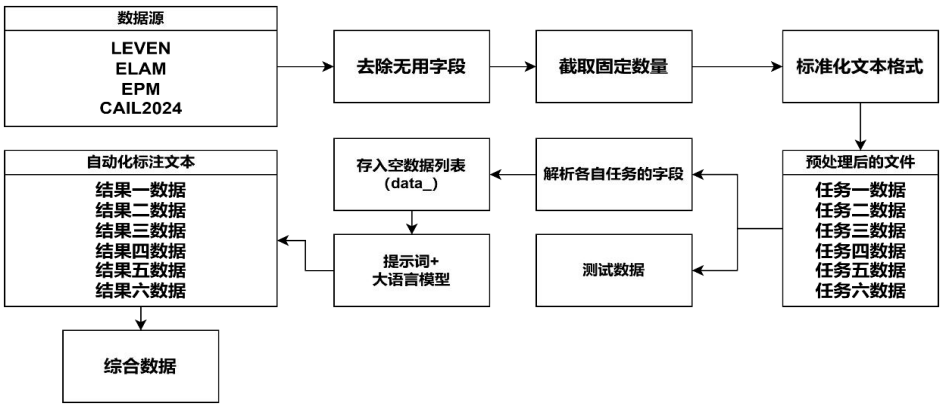


图 1　数据的处理过程

Figure 1　The process of data processing

1.3　数据处理

1.3.1　事件检测和触发词提取

首先初始化一个空的数据列表（data_），这个列表将用来存储处理后的文档数据。之后，打开并读取预处理后的文件，该文件包含了多个法律文本，每个文本都包含法律内容和相关的事件列表。对于文件中的每个文档，解析出其内容（content）和事件列表（events），并遍历文档的内容部分。在这一过程中，创建一个空字典 trigger_dict，用于存储每种事件类型及其对应的触发词。对每种事件类型进行扫描，检查其触发词是否存在于文档内容的列表中。一旦发现触发词，便将其添加到 trigger_dict 中对应的事件类型下。对于每个文档内容，构建一个标签字典 label_dict，用以存储识别出的事件类型和触发词，并将 label_dict 添加到 data_列表中，确保每个文档内容都对应一个字典条目。使用提示词（Prompt）与大语言模型（LLM）交互。对于 data_列表中的每个文档内容，调用不同的 LLM 模型来处理文档内容，并生成事件类别和触发词的识别结果，然后将这些结果存储在数据结构中。最终，将处理后的 data_列表写入文件中保存。

1.3.2　争议焦点识别

争议焦点识别的目标是判断给定的法律案例是否存在争议。首先，从 json 文件中读取预处理后的案例数据，这些数据包含了不同的案例及其标签。初始化一个空列表 data_，用于存储处理后的案例数据。然后，打开并读取包含案例数据的 json 文件，每个案例都包含两个不同的观点及其标签。遍历每个案例，将案例文本以及它们的标签分别添加到 data_列表中。使用 Prompt 与 LLM 模型交互，对 data_列表中的每个案例使用不同的 LLM 模型来生成对争议性的判断。把结果存储在 data_列表中，每个案例对应一个处理后的条目。最终，将处理后的 data_列表写入文件中保存。

1.3.3　裁判文书说理生成

裁判文书说理生成的目标是从法律文档中生成说理（判决原因和判决结果）。从预处理后的文

件中读取数据，文本包含事实描述。首先，初始化一个空列表 data_list，用于存储处理后的文档数据。对于预处理后的文本，使用 Prompt 作为与 LLM 模型交互的提示。对每个文档的内容使用不同的 LLM 模型处理，并生成相应的判决原因和判决文本。生成的判决原因和判决文本被存储在一个字典中，其中包含键 reasoning 和 judgement，分别对应判决原因和判决结果。这个字典随后被添加到 data_list 列表中，每个文档内容对应一个字典条目。处理后的 data_list 列表最终被写入到文件中保存。

1.3.4 裁判文书事实生成

裁判文书事实生成的目标是从法律案例中总结事实。流程类似，首先，初始化一个空列表 data_list，这个列表将用于存储处理后的案例数据。从预处理后的文件中获取案例数据，每个案例都包含有起诉陈述、辩护观点和证据列表等。对于每个案例，先构建一个综合文本，这个文本通过结合 Prompt 提示，用大语言模型将起诉、辩护和证据等信息融合在一起。这个综合文本将作为输入，用于生成每个案例的事实总结。对综合文本使用不同的 LLM 模型进行处理，并生成事实总结。生成的事实总结被存储在一个字典中，这个字典随后被添加到 data_list 列表中，每个案例对应一个字典条目。

1.3.5 法律文本摘要和关键词提取

法律文本摘要和关键词提取任务将从法律文档中提取与基本事实相关的摘要和关键词。首先，打开并读取预处理后的文件，从中获取文档数据，每个文档包含事实描述。对于每个文档，使用 Prompt 作为与 LLM 模型交互的提示。使用不同的 LLM 模型来处理文档内容，并生成相应的摘要和关键词。生成的摘要和关键词被存储在一个字典中，其中包含键 summary 和 keywords，分别对应摘要和关键词。这个字典随后被添加到预先定义的空列表中，每个文档内容对应一个字典条目。

1.3.6 大语言模型预测输出数据

本研究选取了 5 种大语言模型进行全面的数据评测，分别是 Deepseek-chat[11]、qwen-plus[12]两种国产大语言模型，Gpt-4o-mini[13]、Gpt-3.5-turbo[14]两种通用大语言模型和 Farui-plus[15]一种法律专用大语言模型。在 Prompt 的设计方面，根据具体任务，使用准确的语言与大语言模型进行交互，给模型适当的样本和输出格式作为参考。最终模型生成预测数据。

1.3.7 类案检索子任务数据整合

以上处理过程包含源文件中本研究需要的字段和大语言模型预测结果，为了确保数据的一致性和可读性，将这些信息按照标准化的格式进行整理。使用 Python 中的 json 模块，将对象编码 json 格式，每个任务数据集按照事件和触发词检测、争议焦点识别、说理生成、事实生成、摘要生成、关键词抽取命名并保存在同一个 json 文件中。最终构建了面向类案检索的大语言模型多子任务评估数据集，以便于后续的分析和应用。

2 数据样本描述

本数据集包括两个 json 文件，第一个是用于类案检索相关子任务的大语言模型评测数据集 Case_Retrieval_Datasets.json，详细信息如表 1 所示。源数据量比较多时，会先进行数据截取，再进

行后续处理。但也有源数据量本身就很少，比如本研究中的事实生成数据，此源数据虽然样本量少，但是文本提供的信息丰富，并且分点提供，如图 2 所示。在生成事实方面，此源数据作为评测数据在引导模型输出时具有优势。但如果进行模型训练、微调等，在实际操作过程中，需要继续数据增强，增加数据量。

表 1　大语言模型类案检索相关任务评估数据集详细信息

Table 1　Details of the dataset of evaluation for legal case retrieval tasks using large language models

任务	数据	处理过程	数据量
事件检测	LEVEN	截取数量、标准化文本格式	500
触发次抽取	LEVEN	截取数量、标准化文本格式	500
争议焦点识别	ELAM	截取数量、标注、标准化文本格式	500
说理生成	CAIL2024	标准化文本格式	100
事实生成	CAIL2024	标准化文本格式	10
摘要生成	EPM	标准化文本格式	407
关键词提取	EPM	标准化文本格式	407

用于类案检索相关子任务的大语言模型评测数据集数据样例如下所示：

```
[
    {
      "task": "具体任务",
      "content": [
        {
          "text": "法律文本内容",
          "label": "标签内容",
        }
]
```

每个样例详细记录了子任务的类型（task）和内容（content），内容中包含法律文本（text）和标签（label）。在评测过程中，输入 text 的内容，通过 Prompt 与大语言模型交互，使大语言模型输出任务预测结果。将预测结果与 label 进行对比，从而评估大语言模型对这个任务的处理能力。

第二个是加入大语言模型预测结果的综合数据集 Case_Retrieval_Datasets_llm.json。综合数据包含对源数据预处理后的字段和大语言模型的预测结果，即在第一个数据集的基础上在内容中加了 prediction 键值对。以下是整合后的数据格式样例：

```
[
    {
      "task": "具体任务",
      "content": [
        {
          "text": "法律文本内容",
          "label": "标签内容",
```

```
        "prediction":{模型标注内容}
    }
]
```

加入 5 种大语言模型的预测结果之后的综合数据有较大的应用价值。比如，大模型在理解类案检索任务上的能力较高，可以将输出内容作为标注数据，对小模型进行微调，提高模型在类案检索任务上的性能。此外，大模型预测结果可以作为评估其他模型的基线，以便与更多模型进行一致的比较。

截取整合后的评测大语言模型的数据样本，各部分任务数据样本展示如图 2 所示。

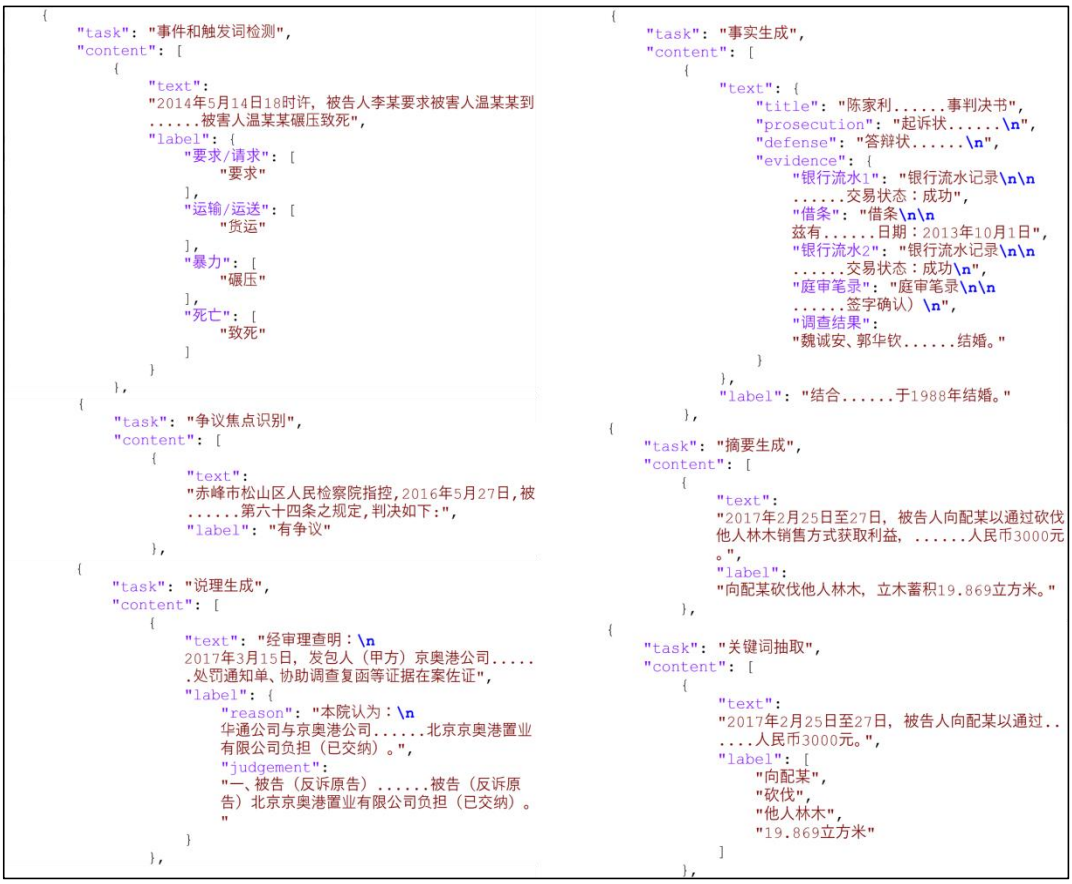


图 2　测试样本数据展示

Figure 2　Illustration of the test sample data

3　数据质量控制和评估

本研究使用 Deepseek 模型、Gpt-3.5 模型对整理的评测数据集进行测试评估。以下评估方法对于其他模型的结果评估同样适用。

3.1　评估指标

评估指标涵盖了多个方面，包括精确度（Precision，P）、召回率（Recall，R）、F1 分数（F1 Score，F）、文本生成任务评估（Rouge）等。针对事件检测、触发词提取、争议焦点检测、关键词生成、说理生成和事实生成等关键任务进行了详细的评估。

$$P=\frac{TP}{TP+FP} \tag{1}$$

TP 是正确识别的正例数量。在评估模型时，这指的是模型正确预测为正类（相关内容）的实例数量。*FP* 是错误识别的正例数量。这指的是模型错误预测为正类的负类（不相关内容）实例数量。

$$R=\frac{TP}{TP+FN} \tag{2}$$

TP 如上所述，正确识别的正例数量。*FN* 是错误排除的正例数量。这指的是模型未能预测为正类的正类实例数量。

$$F=2*\frac{P*R}{P+R} \tag{3}$$

P 和 *R* 如上所述，是精确度和召回率。*F* 分数是精确度和召回率的调和平均数，它试图平衡两者。

Rouge 用以衡量生成文本与参考文本之间的相似性。它主要包括 Rouge-1（基于单词重叠）、Rouge-2（基于双词组合重叠）和 Rouge-L（基于最长公共子序列的连续匹配），其中每个指标通过计 P、R 以及两者的调和平均 F1 分数，综合评估生成文本在准确性和完整性方面的表现。

3.2 评估结果

Deepseek 模型预测进行指标计算结果如表 2–表 3 所示。可以发现，在少样本提示下，大语言模型在法律类案检索多任务处理方面有巨大潜力。在模型表现较好的任务上，如争议焦点检测，P、R、F 指标分别达到 69%、68%、69%，关键词生成任务 R 指标达到 71.3%，判决原因和判决生成任务 P 指标分别达到 71.2%和 81.6%。Gpt-3.5 模型的预测评测结果如表 4–表 5 所示。可以发现 Gpt-3.5 模型在大多数的结果上差于 Deepseek 模型，评测结果验证了 Deepseek 模型在类案检索任务上的优势。因此，面向类案检索的法律大模型评测数据集能有效评估模型的相关能力。

表 2　数据质量控制和评估结果一

Table 2　Data quality control and assessment results (I)

任务	P	R	F
事件检测	0.604	0.578	0.5917
触发词抽取	0.508	0.465	0.485
争议焦点检测	0.690	0.680	0.69
关键词生成	0.542	0.713	0.616

表 3　数据质量控制和评估结果二

Table 3　Data quality control and assessment results (II)

任务		Rouge-1			Rouge-2			Rouge-L		
		R	P	F	R	P	F	R	P	F
说理生成	判决原因	0.623	0.712	0.650	0.329	0.377	0.333	0.494	0.556	0.513
	判决结果	0.522	0.816	0.595	0.351	0.575	0.393	0.479	0.730	0.543
事实生成		0.575	0.389	0.453	0.266	0.208	0.226	0.407	0.294	0.332

表 4　数据质量控制和评估结果四

Table 4　Data quality control and assessment results (IV)

任务	P	R	F
事件检测	0.340	0.515	0.409
触发词抽取	0.287	0.372	0.324
争议焦点检测	0.550	0.420	0.360
关键词生成	0.392	0.645	0.487

表 5　数据质量控制和评估结果五

Table 5　Data quality control and assessment results (V)

任务		Rouge-1			Rouge-2			Rouge-L		
		R	P	F	R	P	F	R	P	F
说理生成	判决原因	0.642	0.529	0.567	0.272	0.200	0.218	0.430	0.349	0.376
	判决结果	0.471	0.761	0.542	0.239	0.437	0.281	0.387	0.614	0.443
事实生成		0.573	0.395	0.455	0.249	0.155	0.180	0.438	0.300	0.345

4　数据价值

随着司法信息化的深入，裁判文书、法律法规等文本数据的快速增长，利用大语言模型进行法律多任务文本处理成为研究热点。本研究收集并构建了多个大模型评测法律任务数据集，通过数据处理一系列步骤构建了面向类案检索子任务的大语言模型评估数据集。本数据集的应用价值显著，它能够支持法律文本摘要、类案检索等多种应用的大语言模型评估，促进类案检索司法智能领域的发展。

数据作者分工职责

王怡文（2000—），女，山东省烟台市人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据集的采集、预处理和整合、论文撰写。

戚肖克（1985—），女，山东省菏泽市人，博士，副教授，研究方向为语音信号处理、自然语言处理。主要承担工作：数据集的采集、数据质量控制与综合管理。

赵小兵（1967—），女，内蒙古自治区呼和浩特市人，博士，教授，研究方向为自然语言处理。主要承担工作：数据质量控制与综合管理。

马传连（1978—），男，山东省临邑市人，博士研究生，研究方向为计算机语言学。主要承担工作：数据质量控制与综合管理。

陈波（1988—），男，湖南省长沙市人，博士，副教授，研究方向为自然语言处理。主要承担工作：数据质量控制与综合管理。

参考文献

[1] FENG Y, LI C Y, NG V. Legal case retrieval: a survey of the state of the art[C]. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand. Stroudsburg, PA, USA: ACL, 2024: 6472–6485.. DOI: 10.18653/v1/2024.acl-long.350.

[2] 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. Virtual Event Canada. ACM, 2021: 2342–2348.. DOI: 10.1145/3404835.3463250.

[3] LI H T, AI Q Y, CHEN J, et al. SAILER: structure-aware pre-trained language model for legal case retrieval[C]. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei, China. ACM, 2023: 1035–1044. DOI: 10.1145/3539618.3591761.

[4] MA Y X, WU Y Y, SU W H, et al. CaseEncoder: a knowledge-enhanced pre-trained model for legal case encoding[C]. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore. Stroudsburg, PA, USA: ACL, 2023: 7134–7143.. DOI: 10.18653/v1/2023.emnlp-main.441.

[5] FEI Z W, SHEN X Y, ZHU D W, et al. LawBench: benchmarking legal knowledge of large language models[EB/OL]. (2023-09-28) [2024-10-27]. https://arxiv.org/abs/2309.16289v1.

[6] DENG C L, MAO K L, DOU Z C. Learning interpretable legal case retrieval via knowledge-guided case reformulation[EB/OL]. (2024-06-28) [2024-10-27]. https://arxiv.org/abs/2406.19760v1.

[7] YAO F, XIAO C J, WANG X Z, et al. LEVEN: a large-scale Chinese legal event detection dataset[EB/OL]. (2022-03-16) [2024-10-27]. https://arxiv.org/abs/2203.08556v1.

[8] YU W J, SUN Z X, XU J, et al. Explainable legal case matching via inverse optimal transport-based rationale extraction[C]. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid Spain. ACM, 2022: 657-668.. DOI: 10.1145/3477495.3531974.

[9] FENG Y, LI C Y, NG V. Legal judgment prediction via event extraction with constraints[C]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland. Stroudsburg, PA, USA: ACL, 2022: 648-664.. DOI: 10.18653/v1/2022.acl-long.48.

[10] CAIL Challenge of ai in Law. Official website[EB/OL]. (2024-10-28) [2024-11-02]. http://cail.cipsc.org.cn/task_summit.html?raceID=0&cail_tag=2024.

[11] DEEPSEEK-AI,: , BI X, et al. DeepSeek LLM: scaling open-source language models with longtermism[EB/OL]. (2024-06-05) [2024-10-27]. https://arxiv.org/abs/2401.02954v1.

[12] BAI J Z, BAI S, CHU Y F, et al. Qwen technical report[EB/OL]. (2023-09-28) [2024-10-27]. https://arxiv.org/abs/2309.16609v1.

[13] OPENAI. Gpt-4o-mini: advancing cost-efficient intelligence[EB/OL]. (2024-03-03) [2024-10-27]. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

[14] OPENAI. API reference for completions[EB/OL]. (2023-07) [2024-10-27]. https://platform.openai.com/docs/api-reference/completions.

[15] FARUI-PLUS. Tongyi Farui API[EB/OL]. (2025-03-31) [2024-10-27]. https://help.aliyun.com/zh/model-studio/developer-reference/tongyi-farui-quick-start.

论文引用格式

王怡文，戚肖克，赵小兵，等. 面向类案检索的法律大模型评测数据集[J/OL]. 中国科学数据, 2026, 11(1). (2025-04-30). DOI:10.11922/11-6035.csd.2024.0199.zh.

数据引用格式

王怡文，戚肖克，赵小兵，等. 面向类案检索的法律大模型评测数据集[DS/OL]. V1. Science Data Bank, 2024. (2024-12-13). DOI:10.57760/sciencedb.17529.

A dataset of evaluation for legal case retrieval tasks using large language models

WANG Yiwen[1,2], QI Xiaoke[3*], ZHAO Xiaobing[1,2] , MA Chuanlian[2,4], CHEN Bo[1,2]

1. School of Information Engineering, Minzu University of China, Beijing 100081, P. R. China

2. National Language Resource Monitoring & Research Center of Minority Languages, Beijing 100081, P. R. China

3. School of Information Management for Law, China University of Political Science and Law, Beijing 102249, P. R. China

4. School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, P. R. China

*Email: qixiaoke@cupl.edu.cn

Abstract: In recent years, the application evaluation of large language models (LLMs) in the legal domain has become a significant focus. Similar case retrieval is one of the most important tasks in this field. However, existing legal LLM evaluation methods often lack in-depth assessment of similar case retrieval tasks and fail to decomposed the task into smaller subtasks. To fully explore the performance of LLMs on tasks in similar case retrieval, this study draws on the definition of similar cases as outlined by the Supreme People's Court of China, breaks down the evaluation into seven subtasks: event detection, fact generation in judgment documents, trigger word extraction, keyword extraction, long-text legal summarization, dispute focus identification, and reasoning generation in judgment documents. For each subtask, open-source datasets were collected and through preprocessing, reconstruction, and integration, a comprehensive evaluation dataset for LLMs on similar case retrieval tasks has been collected and processed through steps including preprocessing, reconstruction, and integration. The resulting dataset—designed specifically for evaluating legal LLMs on similar case retrieval tasks—contains information such as the subtask name, corresponding legal texts (extracted from relevant sections of criminal and civil judgment documents), and labels (the intended output content of the task). This dataset integrates seven types of case retrieval subtasks, totaling 2,424 data entries. To ensure data quality and diversity, five different LLMs were evaluated, with experiment results indicating that the top-performing models achieved accuracy rates exceeding 70%. This dataset offers valuable potential for assessing the multi-task processing capabilities of legal LLMs, and

provides a standardized benchmark for identifying models that better suited to handling similar case retrieval tasks.

Keywords: large language models; legal application evaluation; legal multi-task datasets; legal case retrieval

Dataset Profile

Title	A dataset of evaluation for legal case retrieval tasks using large language models
Data corresponding author	QI Xiaoke (qixiaoke@cupl.edu.cn)
Data authors	WANG Yiwen, QI Xiaoke, ZHAO Xiaobing
Time range	2022–2024
Data volume	11.3MB
Data format	*.json
Data service system	<https://doi.org/10.57760/sciencedb.17529>
Source of funding	The National Social Science Foundation of China (22&ZD035)
Dataset composition	The dataset comprises two data files. The first file, Case_Retrieval_Datasets.json, is a collection of multi-task datasets that have been preprocessed for the evaluation and testing of large language models, with a data size of 3.73 MB. The second file, Case_Retrieval_Datasets_llm.json, is an extended version of the dataset that includes the prediction results generated by large language models, with a data size of 7.66MB.