CrossLex: A Source-Grounded Benchmark for Cross-Jurisdictional Legal Reasoning in Large Language Models

Xiaocui Yang[1], Xican Tan[2], Shoujie Chen[1], Shihan Xiao[3], Keke Tong[4], Xinyu Zhou[2]*

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]Institute of Intelligent Computing, University of Electronic Science and Technology of China, Chengdu, China
[3]School of Civil and Commercial Law, Southwest University of Political Science and Law, Chongqing, China
[4]Chongqing Branch of China Unicom, Chongqing, China
yangxiaocui@cse.neu.edu.cn, chenshoujie2003@163.com, 2022214975@stu.cqupt.edu.cn,
202305255416@smail.xtu.edu.cn, 2022214975@stu.cqupt.edu.cn, zhxy@uestc.edu.cn

Abstract

Legal reasoning is inherently jurisdiction-dependent, i.e., the same facts can call for different legal rules and yield different conclusions across legal systems. Yet existing benchmarks rarely evaluate whether Large Language Models (LLMs) can recognize such jurisdiction-specific variation, especially when identical fact patterns lead to divergent legal outcomes. We introduce CrossLex[1], a same-fact, legal-source-grounded benchmark for evaluating Cross-Jurisdictional legal reasoning in LLMs across three jurisdictions covering China, California, and Germany. Built from authoritative legal sources, CrossLex aligns 55 legal issues spanning contract, consumer, criminal, family, and labor law, and constructs jurisdiction-aligned questions paired with answers and supporting citations. In total, CrossLex contains 6,149 instances organized into 385 fact groups, with all legal issues, answers, and cited authorities reviewed by legal professionals. To disentangle basic legal knowledge from cross-jurisdictional reasoning, CrossLex defines three complementary tasks, including **single-jurisdiction reasoning (T1)**, **joint cross-jurisdictional comparison (T2)**, and **fine-grained cross-jurisdictional evaluation (T3)**. We further propose **Grounded Joint**, a metric that jointly assesses answer correctness and legal-source grounding, and provide a unified evaluation for streamlined benchmarking. Extensive experiments on representative LLMs show that, although current models can often answer legal questions correctly, they struggle to provide accurate cross-jurisdictional legal citations. We hope that CrossLex will facilitate future research on source-grounded cross-jurisdictional legal reasoning.

Introduction

Large Language Models (LLMs) have made substantial progress on a range of legal NLP tasks (Guha et al. 2023; Fei et al. 2024; Fan et al. 2025). Existing legal benchmarks are typically organized around a single jurisdiction, a fixed body of law, or loosely connected collections of independent legal tasks (Chalkidis et al. 2022; Guha et al. 2023; Fei et al. 2024). Consequently, they provide limited insight into a fundamental question in comparative legal reasoning, i.e., *whether LLMs can derive jurisdiction-specific legal conclusions when the factual scenario remains unchanged but the legal system varies*.

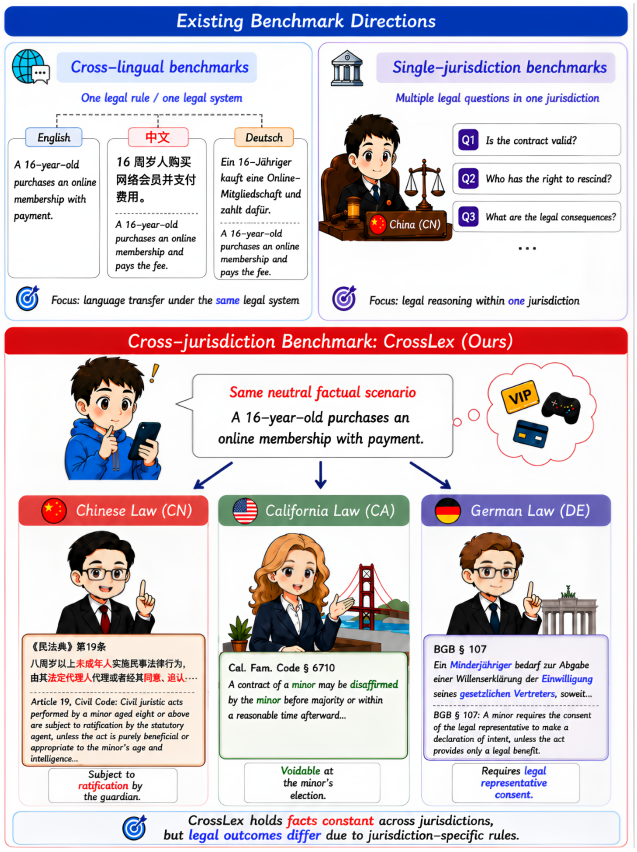

Figure 1: The same factual scenario may lead to different legal conclusions across jurisdictions. CrossLex evaluates whether LLMs can distinguish jurisdiction-specific conclusions under neutral fact patterns.

Legal reasoning is inherently jurisdiction-dependent, i.e., the same factual scenario may yield different legal conclusions under diverse legal systems. As illustrated in Figure 1,

[1]Relevant resources will be released upon paper acceptance.
*Corresponding author.

Benchmark	Multi-Domain	Multi-lingual	Cross-Jurisdiction	Aligned Facts	Citation Grounding
COLIEE (2024)	∼	✗	∼	✗	∼
LexGLUE (2022)	✓	✗	∼	✗	✗
LegalBench (2023)	✓	✗	✗	✗	✗
LawBench (2024)	✓	✗	✗	✗	∼
MultiEURLEX (2021)	✓	✓	✗	✗	✗
LEXTREME (2023)	✓	✓	∼	✗	✗
LEXam (2025)	✓	✓	∼	✗	✗
MultiLegalBench (2026)	∼	✓	✓	✗	✗
CrossLex	✓	✓	✓	✓	✓

Table 1: Comparison with representative legal benchmarks. Multi-domain denotes coverage of multiple legal areas. Cross-jurisdiction evaluates multiple legal systems, while aligned facts requires the same neutral scenario across jurisdictions. Citation grounding evaluates supporting legal sources. ✓, ✗, and ∼ denote full, no, and partial support, respectively.

a 16-year-old's purchase of a paid online membership may have different contractual consequences across jurisdictions. For example, under Chinese law, validity may depend on ratification by the minor's guardian; under California law, the contract may be voidable at the minor's election; and under German law, it may require prior consent or subsequent approval from the minor's legal representative. It highlights a challenge for LLMs in cross-jurisdictional legal reasoning, where models must not only infer a plausible legal conclusion, but also associate it with the correct governing jurisdiction. Otherwise, they may overgeneralize familiar doctrines, conflate similar legal concepts, or incorrectly transfer rules across legal systems. However, existing studies provide limited support for diagnosing jurisdictional confusion, as shown in Table 1. Even when they cover multiple legal systems or languages, evaluation instances are typically considered in isolation, with each question asking for a conclusion under a single specified jurisdiction (Chalkidis, Fergadiotis, and Androutsopoulos 2021; Chalkidis et al. 2022; Niklaus et al. 2024; Fan et al. 2025). They may assess single-jurisdiction legal knowledge, but they do not directly test controlled cross-jurisdictional comparison. Whether an LLM can reason over the same fact, distinguish divergent jurisdiction-specific conclusions, and correctly associate each conclusion with its governing legal.

To address this gap, we introduce **CrossLex**, a source-grounded cross-jurisdictional legal benchmark constructed through a six-step pipeline, as illustrated in Figure 2. CrossLex evaluates LLMs on jurisdiction-aware legal reasoning over aligned, jurisdiction-neutral fact patterns. It spans three representative jurisdictions, including China, California (U.S.), and Germany, and five legal domains, such as contract law, consumer protection law, criminal law, family law, and labor law. Each fact pattern is written independently of any particular legal system and paired with jurisdiction-specific legal conclusions and supporting legal sources, enabling controlled comparison across legal systems under the same factual scenario. CrossLex consists of three task formats with increasing difficulty. **T1: Single-Jurisdiction Judgment** evaluates whether a model can answer a legal question given a specified jurisdiction. **T2: Multi-Jurisdiction Joint Judgment** requires the model to select the correct combination of conclusions for the same fact pattern under different jurisdictions. **T3: Fine-Grained Multi-Jurisdiction Matching** asks the model to match each jurisdiction to its corresponding conclusion from a set of candidates. These tasks form a progression from jurisdiction-specific competence to joint cross-jurisdictional reasoning and fine-grained diagnosis of jurisdiction–conclusion alignment. We evaluate a diverse set of representative LLMs on CrossLex. Our results show that strong performance on T2 can conceal substantial failures on T3, i.e., models can often select the correct set of jurisdictional conclusions, but still fail to align each conclusion with its governing legal system. We further fine-tune LLMs at multiple parameter scales and show that adapted lightweight models can approach, and in some cases surpass, much larger proprietary models in source-grounded legal answering, particularly in citation grounding. Our contributions are threefold:

- We introduce CrossLex, a source-grounded cross-jurisdictional legal benchmark spanning three jurisdictions, China, California (U.S.), and Germany, across five legal domains, paving the way for jurisdiction-aware legal reasoning.
- We design three increasingly challenging tasks over aligned, jurisdiction-neutral fact patterns, evaluating LLMs from single-jurisdiction judgment to multi-jurisdiction joint judgment and fine-grained jurisdiction–conclusion matching.
- We conduct extensive experiments with diverse LLMs, revealing persistent challenges in cross-jurisdictional legal reasoning. These findings highlight the need for further research on jurisdiction-aware legal models.

CrossLex

Construction Overview

To construct CrossLex, we design an end-to-end benchmark construction pipeline, as illustrated in Figure 2. The pipeline comprises six stages: (1) defining the legal scope and cross-jurisdictional issue schemas; (2) constructing and validating controlled fact groups; (3) generating task-specific candidate QA instances; (4) verifying answer–citation consistency through multi-model consensus; (5) conducting professional legal validation and final auditing; and (6) assembling the validated instances into the final benchmark.

The pipeline begins with authoritative legal sources from China, California, and Germany. We identify functionally comparable legal issues across the three jurisdictions and encode each issue as a structured schema containing jurisdiction-specific rules, source evidence, controlled factual variables, distractor templates, and generation guardrails. Each issue schema is then expanded into seven controlled fact groups, which are instantiated into T1, T2, and T3 candidate questions. Candidate instances must pass both multi-model consensus checking and professional legal review before being included in the final benchmark.

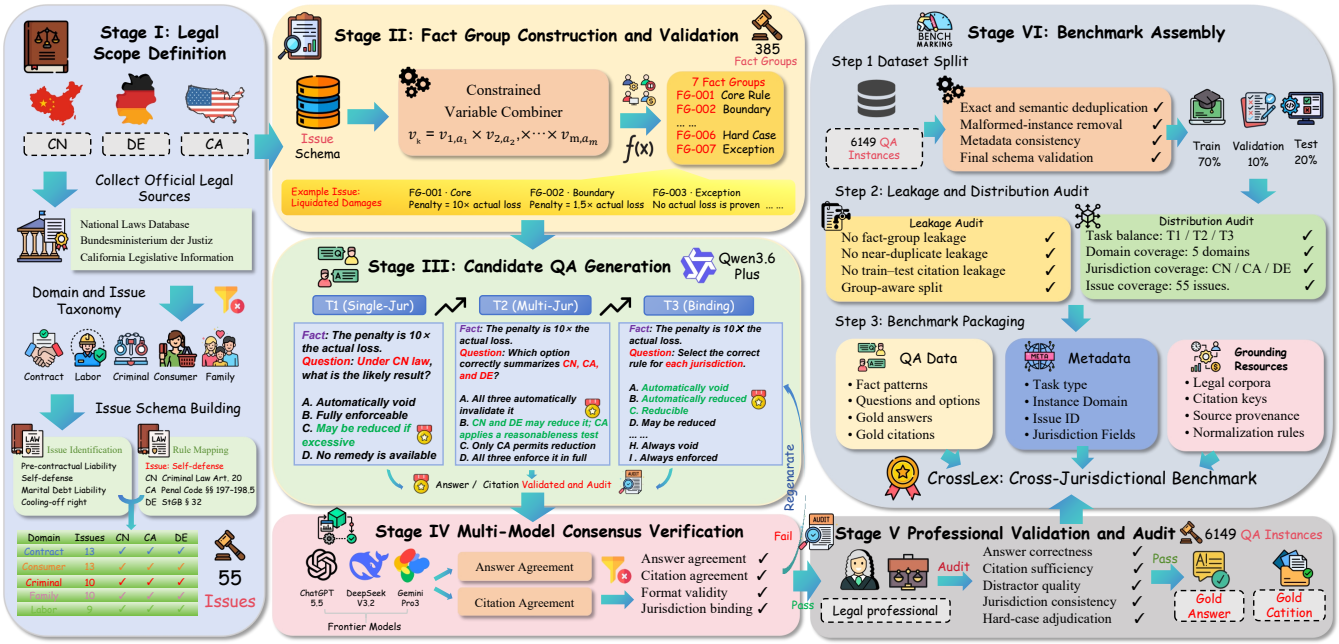

Figure 2: Overview of the CrossLex construction framework. Neutral fact groups are aligned with jurisdiction-specific rules and source evidence, then instantiated into T1, T2, and T3 tasks.

Stage I: Legal Scope and Issue Schema Definition

We define the legal scope of CrossLex over representative legal jurisdictions, i.e., China(CN), California (U.S. CA), and Germany(DE), and five legal domains, such as contract law, consumer protection law, criminal law, family law, and labor law. Let

$$\mathcal{J} = \{\mathrm{CN}, \mathrm{CA}, \mathrm{DE}\} \tag{1}$$

denote the set of jurisdictions considered in CrossLex, where CN, CA, and DE refer to China, California (U.S.), and Germany, respectively.

For each domain, we identify functionally comparable legal issues that can be meaningfully interpreted across all jurisdictions in $\mathcal{J}$. The resulting taxonomy contains 55 legal issues, as shown in Figure 3. Issue alignment is functional rather than literal: although the corresponding rules may differ in legal terminology, doctrinal structure, or legal consequences, they address a shared legal problem.

We construct an issue schema for each issue i.

$$\mathcal{S}_i = (R_{i,\mathrm{CN}}, R_{i,\mathrm{CA}}, R_{i,\mathrm{DE}}, F_i, D_i, B_i, U_i), \tag{2}$$

where $R_{i,j}$ denotes the jurisdiction-specific rule record for jurisdiction $j \in \mathcal{J}$, F_i denotes the fact-specific slots, D_i denotes adversarial distractor templates, B_i denotes jurisdiction-specific forbidden terms and leakage constraints, and U_i denotes answer-uniqueness and issue-scope guardrails.

Each jurisdiction-specific rule record contains a concise rule summary, $s_{i,j}$, an expected legal direction, $y_{i,j}$, and supporting legal authorities, $E_{i,j}$.

$$R_{i,j} = (s_{i,j}, y_{i,j}, E_{i,j}). \tag{3}$$

To prevent models from relying on superficial pattern matching and to assess their ability to distinguish jurisdiction-specific legal rules, we introduce adversarial error templates at the issue-schema level. These templates define recurring failure modes, including jurisdiction confusion, rule swapping, overgeneralization, and incorrect legal citation, which guide later candidate option construction. For each issue schema, we define at least three jurisdiction-specific error templates covering these failure modes. We further use forbidden-term lists to prevent explicit jurisdictional leakage and apply uniqueness guardrails to ensure that each multiple-choice item has exactly one legally defensible answer. Detailed distractor examples and generation prompts are included in the supplementary materials.

Stage II: Fact Group Construction and Validation

Each locked issue schema is expanded into multiple controlled fact groups. For issue i, let

$$F_i = \{V_{i,\ell}\}_{\ell=1}^{m_i}$$

denote its m_i controllable factual slots, where $V_{i,\ell}$ is the set of candidate values for slot ℓ. The k-th valid variable combination is

$$\begin{aligned} \mathbf{v}_{i,k} &= \left(v_{i,1,a_1^{(k)}}, \ldots, v_{i,m_i,a_{m_i}^{(k)}}\right), \\ \text{s.t.} \quad & v_{i,\ell,a_\ell^{(k)}} \in V_{i,\ell}, \qquad \mathbf{v}_{i,k} \models \mathcal{G}_i, \end{aligned} \tag{4}$$

where $a_\ell^{(k)}$ is the index of the value selected for slot ℓ, and $\mathbf{v}_{i,k} \models \mathcal{G}_i$ indicates that the combination satisfies all issue-specific compatibility, legality, and uniqueness constraints.

Each valid combination is instantiated into a fact group $g_{i,k}$, where $k \in \{1, \ldots, K\}$ and $K = 7$. Its jurisdiction-specific answer directions are represented as

$$Y_{i,k} = (y_{i,k,\mathrm{CN}}, y_{i,k,\mathrm{CA}}, y_{i,k,\mathrm{DE}}). \tag{5}$$

Here, $y_{i,k,j}$ denotes the final legal direction for jurisdiction j after instantiating the factual variables of $g_{i,k}$, whereas $y_{i,j}$ denotes the issue-level default direction defined in Stage I.

Most fact groups preserve the issue-level rule direction, while selected divergence-focused groups intentionally produce different outcomes across jurisdictions. These groups are particularly important for constructing T3 jurisdiction-binding instances. Before locking, each fact group is audited for fact specificity, constraint completeness, issue consistency, cross-issue risk, and human-review requirements. Failed groups are regenerated or revised. This process yields $385 = 55 \times 7$ validated fact groups from 55 issue schemas.

Stage III: Candidate QA Generation

Each validated fact group is transformed into three task-specific QA formats. Let $\mathcal{T} = \{\mathrm{T1}, \mathrm{T2}, \mathrm{T3}\}$ denote the task set. For each $g_{i,k}$ and $t \in \mathcal{T}$,

$$\mathcal{Q}^{\mathrm{cand}}_{i,k,t} = G_\theta\left(f_{i,k}, Y_{i,k}, \mathcal{S}_i, t\right), \tag{6}$$

where $\mathcal{S}_i$ is the corresponding issue schema and G_θ denotes the candidate-generation LLM. In CrossLex, we use Qwen-3.6-Plus to generate task-specific questions, answer options, and distractors from the validated fact groups, while all legal directions and final annotations are determined through source-grounded validation.

We design three task formats with increasing difficulty, as shown in Stage III of Figure 3. For T1, the generator produces a single-jurisdiction multiple-choice question for one specified jurisdiction. For T2, it constructs a joint multiple-choice item whose correct option encodes the complete CN–CA–DE conclusion tuple. For T3, it produces a candidate-conclusion pool and requires separate jurisdiction-to-conclusion assignments. Correct answer candidates are instantiated from the expected answer directions. During candidate QA generation, the predefined adversarial error templates are instantiated into concrete distractor options. These distractors simulate realistic legal reasoning conditions, where similar doctrines, related legal issues, and cross-jurisdictional rules may lead to plausible but incorrect conclusions. They prevent models from relying on superficial patterns and enable evaluation of fine-grained jurisdiction-aware reasoning. Distractors are generated from the locked adversarial templates in the issue schema. They cover direction reversal, jurisdiction-rule swapping, overgeneralization, and citation-related confusion. Candidate-level guardrails remove items containing forbidden jurisdictional terms, duplicated options, multiple defensible answers, or unintended adjacent legal issues.

Stage IV: Multi-Model Consensus Verification

Candidate QA instances are independently evaluated by three LLM validators: GPT-5.5 (OpenAI 2026a), DeepSeek-V3.2 (DeepSeek-AI et al. 2025), and Gemini 3 Pro (Google DeepMind 2026a). Each validator receives the same fact pattern, question, candidate answers, and task instructions, but does not observe the provisional gold label from Stage III. Each model independently predicts the answer and provides supporting legal authorities for all applicable jurisdictions. Unlike conventional answer-only validation, our validation process jointly examines answer correctness and legal-source grounding: a candidate is retained only when the predicted conclusion is consistent and the cited authorities match the corresponding jurisdiction-specific evidence after citation normalization.

For each candidate $x \in \mathcal{X}^{\mathrm{cand}}$, three validators $V^{(m)}$, where $m \in \{1, 2, 3\}$, independently produce

$$o^{(m)}_x = V^{(m)}(x) = \left(\hat{y}^{(m)}_x, \{\hat{E}^{(m)}_{x,j}\}_{j \in \mathcal{J}_x}\right), \tag{7}$$

where $\mathcal{J}_x \subseteq \mathcal{J}$ is the set of jurisdictions involved in item x, $\hat{y}^{(m)}_x$ is the predicted answer, and $\hat{E}^{(m)}_{x,j}$ is the supporting legal evidence predicted for jurisdiction j. Raw citations are normalized into canonical citation identifiers before comparison. An item passes model verification only when all validators agree on the answer and their normalized citations are consistent with the allowed source evidence.

$$C_{\mathrm{model}}(x) = C_{\mathrm{ans}}(x) \wedge C_{\mathrm{cite}}(x) \wedge C_{\mathrm{fmt}}(x) \wedge C_{\mathrm{jur}}(x), \tag{8}$$

where the four terms denote answer agreement, citation agreement, format validity, and jurisdiction consistency. Detailed validator prompts, output schemas, and verification criteria are provided in the supplementary materials. Items that fail answer or citation consensus are rejected or returned to the candidate-generation stage. Items that pass are treated as model-verified QA candidates and forwarded to professional validation.

Stage V: Professional Validation and Final Audit

Model consensus provides a large-scale consistency check over both answers and cited legal authorities. Since agreement among LLMs does not guarantee legal validity, all consensus-passed candidates are further reviewed by legal professionals for final verification. Legal reviewers with formal legal training participated in the final validation stage. Each item was independently reviewed by two legally trained annotators. Reviewers were assigned according to jurisdictional expertise, and disagreements were resolved through adjudication with reference to the official legal sources. They manually audit each item along five dimensions:

- **Answer correctness**: whether the selected option or jurisdiction mapping follows from the relevant law;
- **Citation sufficiency**: whether the cited authority directly supports the legal conclusion;
- **Distractor quality**: whether distractors are plausible but legally distinguishable;
- **Jurisdiction consistency**: whether each rule and citation is associated with the correct legal system;
- **Hard-case adjudication**: whether ambiguous or borderline instances require revision or exclusion.

Reviewers may accept, revise, or reject an item. Revised items are returned for revalidation. Only items that pass the professional audit receive final gold answers and gold citations. This process produces 6,149 expert-validated QA instances. Detailed review guidelines and annotation instructions are provided in the supplementary materials.

Stage VI: Benchmark Assembly

The 6,149 expert-validated QA instances are curated and assembled into the final CrossLex benchmark. We first perform exact and semantic deduplication, remove malformed

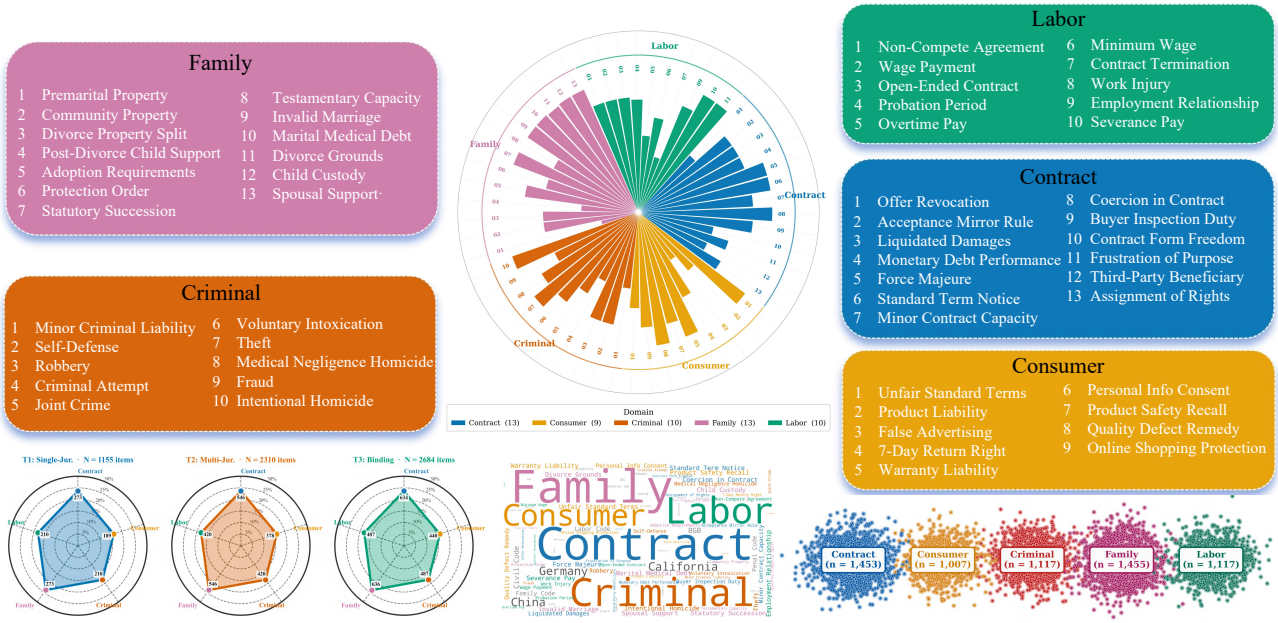

Figure 3: Distribution of legal issues across the five domains in CrossLex. The central radial bar chart shows the number and relative distribution of issues in contract, consumer protection, criminal, family, and labor law, while the surrounding panels list the fine-grained legal issues included in each domain.

instances, verify metadata consistency, and conduct final schema validation. To prevent leakage, all QA instances derived from the same fact group are assigned to the same split. In particular, T1, T2, and T3 variants that share the same underlying fact group cannot be separated across training and evaluation splits. We further audit near-duplicate questions and ensure that retrieval resources and query-expansion lexicons are constructed without using development or test information. The data is partitioned into training, validation, and test sets using a 70%/10%/20% split. We audit the distributions of task types, legal domains, jurisdictions, and issue coverage across the three splits. The final benchmark contains three components: **QA data**: neutral fact patterns, questions, options, gold answers, and gold citations; **Metadata**: task type, domain, issue identifier, fact-group identifier, jurisdiction fields, and split information; **Grounding resources**: legal corpora, normalized citation keys, source provenance, and citation-normalization rules.

Experiments

Dataset Statistics

CrossLex contains 55 aligned legal issues, 385 fact groups, and 6,149 professionally validated QA instances across five domains and three jurisdictions. Each instance includes a gold answer, jurisdiction-specific citations, normalized citation identifiers, task metadata, and source provenance.

Experimental Setup

We evaluate CrossLex under a progressive setting: T1 evaluates single-jurisdiction judgment, T2 requires joint prediction of the CN–CA–DE conclusion tuple under aligned facts, and T3 evaluates jurisdiction–conclusion binding.

Baselines

Representative LLMs We evaluate representative LLMs across different scales and access settings. Closed-source and API-accessible models include GPT-5.6 (OpenAI 2026b), Claude-Opus-4.8 (Anthropic 2026), Gemini-3.1-Pro (Google DeepMind 2026b), and Qwen3.7-Max (Alibaba Cloud 2026). Large open-weight models include DeepSeek-V4-Pro (DeepSeek-AI 2026), GLM-5.2 (GLM-5 Team 2026), and Llama-3.3-70B-Instruct (Meta 2024). Medium-scale models include Qwen3.5-35B-A3B and Qwen3.5-27B (Qwen Team 2026), and DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI 2025). Small-scale models include GLM-4-9B (GLM Team 2024), DeepSeek-R1-Distill-8B (DeepSeek-AI 2025), Qwen3-14B and Qwen3-8B (Qwen Team 2025), and Qwen3.5-2B, Qwen3.5-4B, and Qwen3.5-9B (Qwen Team 2026).

Multilingual BM25-RAG Because CrossLex involves heterogeneous languages and citation conventions across Chinese, California, and German law, conventional monolingual retrieval is not well suited to retrieving relevant legal authorities across jurisdictions. We therefore construct a multilingual BM25-based RAG baseline by combining BM25 retrieval (Robertson and Zaragoza 2009) with retrieval-augmented generation (Lewis et al. 2020), and evaluate whether external legal evidence improves source-grounded reasoning. Given a legal query, the retriever first identifies the target jurisdiction and retrieves relevant authorities from the corresponding legal corpus. The retrieved evidence is then provided to the model for answer generation. Details of these models are provided in the supplementary materials.

Evaluation Metrics

We report answer correctness using Accuracy and citation quality using Citation-F1. Citation-F1 is computed over normalized legal-source identifiers independently of answer correctness. For T1, we measure whether the model selects the correct legal conclusion under the specified jurisdiction. For

Model	T1						T2			T3						GJ-Avg
	All	CN	CA	DE	Citation-F1	GJ	Joint	Citation-F1	GJ	All	CN	CA	DE	Citation-F1	GJ	
Closed-source LLMs																
GPT-5.6	99.6	**100.0**	**100.0**	98.7	60.0	**59.8**	98.1	39.3	38.6	43.7	47.6	46.6	50.7	18.7	16.0	38.2
Claude-4.8	98.3	98.7	98.7	97.4	**60.6**	59.6	98.7	**52.1**	**51.5**	**86.9**	**91.8**	**95.3**	**95.5**	**51.5**	**45.2**	**52.1**
Gemini-3.1	**100.0**	**100.0**	**100.0**	**100.0**	46.1	46.1	**99.1**	49.7	49.5	40.5	45.3	42.4	45.3	24.4	21.5	39.0
Qwen-3.7	**97.8**	**100.0**	**100.0**	93.5	60.6	59.1	**99.6**	**49.8**	**49.6**	**71.8**	**79.5**	**89.9**	**94.4**	48.2	**34.8**	47.8
Open-source Large LLMs																
GLM-5.2	**97.8**	98.7	**100.0**	**94.8**	**67.9**	**66.6**	99.1	46.5	46.2	62.7	78.0	88.1	82.6	**50.1**	32.1	**48.3**
DS-V4	84.8	96.1	85.7	72.7	55.0	48.5	86.6	41.2	39.6	30.4	39.4	45.5	47.9	27.4	14.3	34.1
Llama-3.3-70B	**96.1**	94.8	**97.4**	**96.1**	52.2	50.4	97.0	29.7	29.3	27.6	42.4	77.8	66.0	25.8	8.0	29.3
Open-source Medium LLMs																
Qwen3.5-35B	94.4	**98.7**	**97.4**	87.0	58.7	56.0	98.3	**49.8**	**49.3**	37.3	56.7	75.4	70.5	45.5	16.2	40.5
Qwen3.5-27B	**96.1**	**98.7**	**97.4**	92.2	**63.1**	**61.0**	**99.4**	49.2	49.1	**51.3**	**65.7**	**84.7**	**82.8**	**46.4**	**24.7**	**44.9**
DS-R1-14B	87.4	93.5	96.1	72.7	48.0	46.2	93.9	25.0	23.8	16.0	47.9	60.4	44.0	19.6	3.5	24.5
Qwen3-14B	82.3	94.8	90.9	61.0	52.9	46.7	90.7	32.0	30.8	21.8	42.7	58.2	43.8	21.7	5.1	27.5
Open-source Small LLMs																
GLM-4-9B	78.8	85.7	88.3	62.3	**53.5**	46.4	85.9	27.0	24.6	5.2	31.0	38.1	30.4	22.8	1.0	24.0
DS-R1-8B	84.0	**98.7**	87.0	66.2	45.7	43.6	89.4	21.7	20.5	17.7	50.7	**64.9**	43.8	15.6	3.2	22.4
Qwen3-8B	80.5	94.8	89.6	57.1	50.3	45.4	89.2	24.6	22.7	12.1	37.9	41.2	34.3	18.7	2.3	23.4
Qwen3.5-9B	**92.5**	98.4	**94.7**	**91.8**	37.2	36.6	**94.6**	25.8	25.1	15.5	**60.1**	47.6	**43.9**	**23.9**	3.7	21.8
Qwen3.5-4B	77.4	98.4	87.3	64.8	37.1	36.0	79.0	17.4	16.1	10.6	51.6	44.0	42.1	17.9	2.3	18.1
Qwen3.5-2B	46.2	89.3	53.7	37.8	29.1	24.4	70.7	17.4	13.7	1.2	36.0	28.0	26.1	9.6	0.0	12.7

Table 2: Main zero-shot results of diverse LLMs on the test set of CrossLex (%). Apart from Citation-F1 and GJ, other columns are Accuracy, covering three jurisdictions (CN, CA, DE) across three tasks (T1, T2, T3).T3 All denotes exact jurisdiction-binding accuracy, and GJ-Avg is the macro-average of the T1, T2, and T3 GJ scores. Best results within each model group are shown in **bold**.

T2, Joint Accuracy requires the complete CN–CA–DE conclusion tuple to be correct. For T3, Jurisdiction Binding Accuracy measures whether each predicted conclusion is assigned to the correct jurisdiction. Since a legally grounded response requires both a correct conclusion and appropriate supporting authority, we define **Grounded Joint (GJ)** as

$$\mathrm{GJ} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left[\hat{y}_n = y_n\right] \cdot \mathrm{F1}_n^{\mathrm{cite}}, \tag{9}$$

where N is the number of evaluated instances, $\hat{y}_n$ and y_n are the predicted and gold answers, and $\mathrm{F1}_n^{\mathrm{cite}}$ is the Citation-F1 score between the predicted and gold normalized citation sets for instance n.

Main Results

Table 2 reports the overall results on CrossLex. Performance generally improves with model scale, and larger models tend to achieve stronger overall results. However, even the best-performing models remain unreliable in cross-jurisdictional legal citation grounding.

Models perform strongly on T1, suggesting that current LLMs possess substantial single-jurisdiction legal knowledge. This capability, however, does not fully transfer to cross-jurisdictional reasoning. In particular, T3 reveals a pronounced **binding gap**: models may identify plausible legal conclusions under the same fact pattern, yet fail to associate each conclusion with the correct governing jurisdiction. This indicates that recognizing applicable legal rules and preserving jurisdiction-specific boundaries are distinct capabilities.

Grounded Joint (GJ) scores are also consistently lower than answer accuracy, exposing a substantial **grounding gap**. Models may select the correct answer while providing unsupported, incomplete, or jurisdiction-inconsistent citations. As shown in Fig. 4, citation grounding further varies across CN, CA, and DE, as well as across legal domains. Models tend to struggle more in domains with stronger jurisdictional variation, reflecting differences in legal language, citation conventions, and source structures. Overall, these findings show that legal model evaluation should move beyond answer accuracy toward jurisdiction-aware and source-grounded assessment.

CrossLex-guided Legal Adaptation

We examine whether retrieval-augmented generation (RAG) and CrossLex supervision improve source-grounded legal reasoning, as reported in Fig. 5. Compared with zero-shot inference, multilingual BM25-RAG supplies additional legal evidence and improves answer selection in some settings. Nevertheless, its gains in GJ remain limited, indicating that retrieval alone is insufficient for jurisdiction-aware legal reasoning. Models must also associate retrieved authorities with the correct jurisdiction-specific conclusions.

We then fine-tune lightweight models on the CrossLex training set using Low-Rank Adaptation (LoRA) (Hu et al. 2022). LoRA substantially improves citation quality and Grounded Joint scores while maintaining high answer ac-

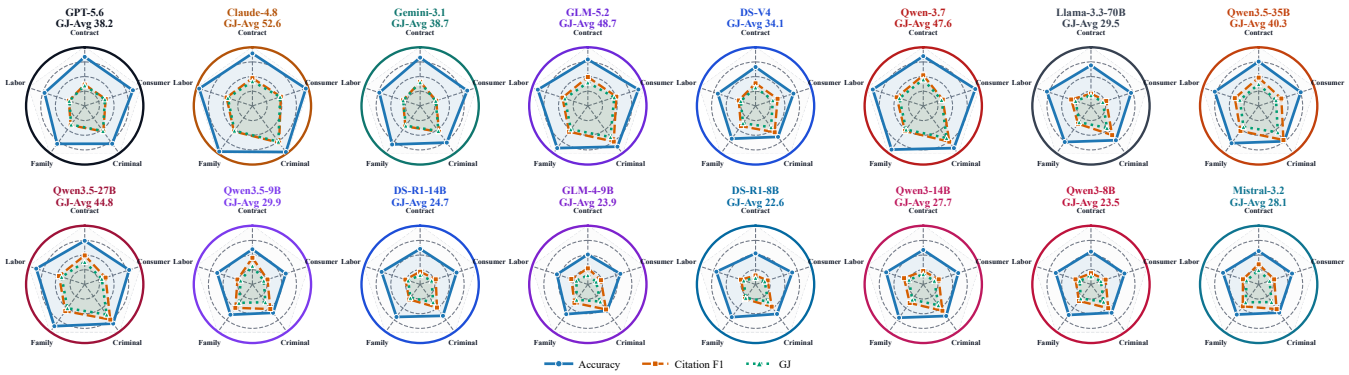

Figure 4: Per-model multi-metric radar profiles on CrossLex (T1/T2/T3 Accuracy, Citation-F1, and GJ) across five law domains (Contract, Consumer protection, Criminal, Family, and Labor law). Each panel shows one model under unified exact-set scoring with multiple denominators.

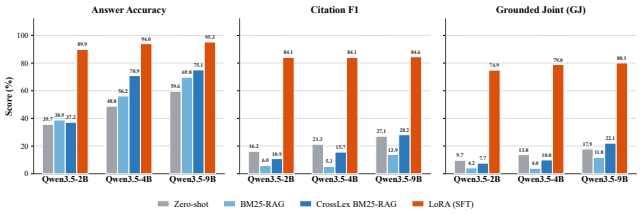

Figure 5: Adaptation results on CrossLex across model scales.

curacy. These results suggest that CrossLex is useful not only for evaluation, but also as supervision for learning jurisdiction-aware and source-grounded legal reasoning.

Related Works

Legal Benchmarks

Existing legal benchmarks cover judgment prediction, question answering, case holding identification, contract review, statutory reasoning, legal retrieval, and general legal reasoning. Representative examples include CAIL for Chinese legal judgment prediction, JEC-QA for judicial-exam question answering, CaseHOLD for U.S. case holding identification, CUAD for contract review, and recent LLM-oriented benchmarks such as LegalBench, LawBench, and LEXam (Xiao et al. 2018; Zhong et al. 2020; Zheng et al. 2021; Hendrycks et al. 2021; Guha et al. 2023; Fei et al. 2024; Fan et al. 2025).

Another line of work studies multilingual and multi-system legal understanding through benchmarks such as MultiEURLEX, LexGLUE, LEXTREME, and MultiLegalBench (Chalkidis, Fergadiotis, and Androutsopoulos 2021; Chalkidis et al. 2022; Niklaus et al. 2023; Ovcharov 2026), as well as multilingual legal corpora and pretrained models such as MultiLegalPile, LEGAL-BERT, and Lawformer (Niklaus et al. 2024; Chalkidis et al. 2020; Xiao et al. 2021).

These resources are valuable, but they generally evaluate jurisdictions, languages, or tasks independently. Even when multiple legal systems or languages are covered, they do not align identical factual scenarios across jurisdictions or test whether models bind each legal conclusion to the correct legal system. In contrast, CrossLex focuses on cross-jurisdictional reasoning under aligned neutral fact patterns, where the same facts are paired with jurisdiction-specific rules and legal outcomes.

Legal Reasoning

Legal reasoning differs from general question answering because legal conclusions should be supported by authoritative statutes, cases, or regulatory sources. Prior work has therefore studied legal retrieval, legal entailment, contract clause extraction, and citation-sensitive evaluation (Hendrycks et al. 2021; Goebel et al. 2024; Lewis et al. 2020; Karpukhin et al. 2020; Izacard and Grave 2021). Retrieval-augmented generation and dense retrieval methods are widely used for knowledge-intensive tasks, while legal retrieval benchmarks such as COLIEE emphasize the need to connect legal questions with relevant legal texts. However, source grounding alone does not resolve the cross-jurisdictional problem. i.e., a model may retrieve or cite plausible legal sources but still assign the resulting conclusion to the wrong legal system. CrossLex therefore pairs each jurisdiction-specific conclusion with supporting legal evidence and further evaluates fine-grained jurisdiction-to-conclusion matching, making it possible to diagnose whether models preserve legal-system boundaries under aligned facts.

Conclusion

We introduce CrossLex, a source-grounded benchmark for cross-jurisdictional legal reasoning over aligned neutral facts from Chinese, California, and German law across five domains. It defines three tasks covering single-jurisdiction judgment (T1), joint cross-jurisdictional comparison (T2), and fine-grained jurisdiction–conclusion binding (T3). Experiments show that frontier models approach ceiling performance on T2, while T3 reveals substantial failures in assigning legal conclusions to the correct jurisdiction, highlighting the limitations of conventional single-jurisdiction and multilingual legal QA evaluation.

Ethics Statement

This work involves human participants serving as legal reviewers. They review synthesized neutral fact patterns and publicly available legal sources and do not provide private or personally identifiable data. CrossLex is intended for research evaluation only and should not replace professional legal judgment. Real-world use should involve qualified legal professionals and appropriate human oversight.

References

Alibaba Cloud. 2026. Qwen3.7-Max in Alibaba Cloud Model Studio. https://www.alibabacloud.com/help/en/model-studio/models. Model ID: `qwen3.7-max`; accessed: 2026-07-29.

Anthropic. 2026. Claude Opus 4.8 System Card. https://www.anthropic.com/system-cards. System card published in May 2026; accessed: 2026-07-29.

Chalkidis, I.; Fergadiotis, M.; and Androutsopoulos, I. 2021. MultiEURLEX: A Multi-lingual and Multi-label Legal Document Classification Dataset for Zero-shot Cross-lingual Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6974–6996. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; and Androutsopoulos, I. 2020. LEGAL-BERT: The Muppets Straight Out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. Online: Association for Computational Linguistics.

Chalkidis, I.; Jana, A.; Hartung, D.; Bommarito, M.; Androutsopoulos, I.; Katz, D.; and Aletras, N. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4310–4330. Dublin, Ireland: Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in Large Language Models via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

DeepSeek-AI. 2026. DeepSeek-V4 Technical Report. Technical report, DeepSeek-AI.

DeepSeek-AI; Liu, A.; Mei, A.; Lin, B.; et al. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv preprint arXiv:2512.02556*.

Fan, Y.; Ni, J.; Merane, J.; Salimbeni, E.; Tian, Y.; Hermstrüwer, Y.; Huang, Y.; Akhtar, M.; Geering, F.; Dreyer, O.; Brunner, D.; Leippold, M.; Sachan, M.; Stremitzer, A.; Engel, C.; Ash, E.; and Niklaus, J. 2025. LEXam: Benchmarking Legal Reasoning on 340 Law Exams. *arXiv preprint arXiv:2505.12864*.

Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Huang, A.; Zhang, S.; Chen, K.; Yin, Z.; Shen, Z.; Ge, J.; and Ng, V. 2024. LawBench: Benchmarking Legal Knowledge of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7933–7962. Miami, Florida, USA: Association for Computational Linguistics.

GLM-5 Team. 2026. GLM-5: From Vibe Coding to Agentic Engineering. *arXiv preprint arXiv:2602.15763*.

GLM Team. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*.

Goebel, R.; Kano, Y.; Kim, M.-Y.; Rabelo, J.; Satoh, K.; and Yoshioka, M. 2024. Overview and Discussion of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. *The Review of Socionetwork Strategies*, 18: 27–47.

Google DeepMind. 2026a. Gemini 3 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-pro/. Model released November 2025; last updated May 2026; accessed July 29, 2026.

Google DeepMind. 2026b. Gemini 3.1 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Published: 2026-02-19; accessed: 2026-07-29.

Guha, N.; Nyarko, J.; Ho, D. E.; Ré, C.; Chilton, A.; Narayana, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D. N.; Zambrano, D.; Talisman, D.; Hoque, E.; Surani, F.; Fagan, F.; Sarfaty, G.; Dickinson, G. M.; Porat, H.; Hegland, J.; Wu, J.; Nudell, J.; Niklaus, J.; Nay, J.; Choi, J. H.; Tobia, K.; Hagan, M.; Ma, M.; Livermore, M.; Rasumov-Rahe, N.; Holzenberger, N.; Kolt, N.; Henderson, P.; Rehaag, S.; Goel, S.; Gao, S.; Williams, S.; Gandhi, S.; Zur, T.; Iyer, V.; and Li, Z. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36.

Hendrycks, D.; Burns, C.; Chen, A.; and Ball, S. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Izacard, G.; and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880. Online: Association for Computational Linguistics.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781. Association for Computational Linguistics.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33.

Meta. 2024. Llama-3.3-70B-Instruct Model Card. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: 2026-07-29.

Niklaus, J.; Matoshi, V.; Rani, P.; Galassi, A.; Stürmer, M.; and Chalkidis, I. 2023. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3016–3054. Singapore: Association for Computational Linguistics.

Niklaus, J.; Matoshi, V.; Stürmer, M.; Chalkidis, I.; and Ho, D. 2024. MultiLegalPile: A 689GB Multilingual Legal Corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15077–15094. Bangkok, Thailand: Association for Computational Linguistics.

OpenAI. 2026a. GPT-5.5 System Card. https://openai.com/index/gpt-5-5-system-card/. Published April 23, 2026; accessed July 29, 2026.

OpenAI. 2026b. GPT-5.6 System Card. https://deploymentsafety.openai.com/gpt-5-6. Published: 2026-07-09; accessed: 2026-07-29.

Ovcharov, V. 2026. Multi-Legal-Bench: Evaluating LLMs on Legal Reasoning Across Jurisdictions, Languages, and Legal Traditions. *arXiv preprint arXiv:2605.29738*.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-07-29.

Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Xiao, C.; Hu, X.; Liu, Z.; Tu, C.; and Sun, M. 2021. Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents. *AI Open*, 2: 79–84.

Xiao, C.; Zhong, H.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; and Xu, J. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv preprint arXiv:1807.02478*.

Zheng, L.; Guha, N.; Anderson, B. R.; Henderson, P.; and Ho, D. E. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 159–168. Association for Computing Machinery.

Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020. JEC-QA: A Legal-Domain Question Answering Dataset. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, 9701–9708. AAAI Press.