LAUKIN: A Multi-jurisdictional Common Law Contract Dataset

Amrita Singh
Computer Science and Engineering
UNSW, Sydney, Australia
amrita.singh1@unsw.edu.au

Aditya Joshi
Computer Science and Engineering
UNSW, Sydney, Australia
aditya.joshi@unsw.edu.au

Jiaojiao Jiang
Computer Science and Engineering
UNSW, Sydney, Australia
jiaojiao.jiang@unsw.edu.au

Hye-young Paik
Computer Science and Engineering
UNSW, Sydney, Australia
h.paik@unsw.edu.au

May Fong Cheong
Law and Justice
UNSW, Sydney, Australia
mf.cheong@unsw.edu.au

Abstract

Multinational companies increasingly require cross-jurisdictional contract review, yet existing legal NLP datasets are largely restricted to a single jurisdiction. We introduce **LAUKIN** (**L**egal equivalence dataset of **A**ustralia, **UK**, and **IN**dia), a dataset of clause pairs (AU-UK, UK-IN, IN-AU) labelled for boolean legal equivalence. We develop a novel multi-stage retrieval and reranking pipeline to construct the initial clause pair mapping, with a subset of clause pairs subsequently annotated by legal experts as Equivalent or Not Equivalent. The dataset comprises 14,727 clause pairs from 204 contracts across 8 agreement types, of which 3,000 are manually labelled: 900 train, 600 dev, and 1,500 test. We evaluate 12 models across 4 techniques, achieving a best macro-F1 of 65.11%, establishing LAUKIN as a challenging benchmark. Results reveal that, despite shared legal heritage, drafting conventions diverge significantly across jurisdictions, making cross-jurisdictional equivalence classification non-trivial. LAUKIN also includes 11,727 unlabelled training pairs to support future semi-supervised learning research in legal NLP.

CCS Concepts

• **Computing methodologies** → **Language resources**.

1 Introduction

Natural language processing (NLP) for the legal domain has witnessed remarkable progress over the past decade, with applications spanning contract review, judgment prediction and summarization, statutory interpretation, and information retrieval from case law [1, 2, 6, 15]. These advancements are reshaping legal practice for scholars, educators, researchers, and practitioners by automating time-intensive processes, cutting operational costs, and reducing human error [2, 7, 8, 15, 17]. However, existing datasets and benchmarks remain largely single-jurisdictional, restricted to specific countries such as the USA, China, and the European Union, limiting their applicability and coverage across other jurisdictions [15]. This is problematic in an increasingly globalised world, where multinational companies not only need to adapt their legal contracts to the distinct terminology, clause structures, and enforceability standards of different jurisdictions, but also perform cross-jurisdictional contract review. The challenge is amplified among jurisdictions such as Australia, the UK, and India, which are rooted in the same English common law tradition, but their contract laws have diverged significantly through independent court systems, distinct legislation, and local legal conventions [3, 10]. Consequently, their clauses differ in lexical choices, syntactic and pragmatic structure, making cross-jurisdictional contract review a complex task. For instance, consider an equivalent force majeure clause pair:

> *India:* Neither Party shall be responsible for delays or failures in performance resulting from acts of god, acts of civil or military authority, fire, flood, strikes, war, epidemics, shortage of power, or other acts or causes reasonably beyond the control of such Party.
> *Australia:* Neither party will be liable for any delay or non-performance of its obligations under this Agreement to the extent that the delay or non-performance is caused or contributed to by a Force Majeure Event, subject to compliance with this clause 16.

This pair illustrates an instance of lexical variation: the Indian clause enumerates disaster events such as floods, epidemics, and shortage of power, whereas the Australian clause adopts the term Force Majeure Event. The absence of multi-jurisdictional datasets for legal NLP impedes the automation of such tasks across common law jurisdictions. To address this gap, we introduce **LAUKIN** (**L**egal equivalence dataset of **A**ustralia, **UK**, and **IN**dia), the **first multi-jurisdictional common law contract dataset** consisting of paired English legal clauses: AU-UK, UK-IN, and IN-AU. LAUKIN is constructed semi-automatically via a novel multi-stage retrieval and reranking pipeline for cross-jurisdictional clause mapping, with a subset of clause pairs annotated by legal experts as Equivalent or Not Equivalent. The dataset comprises 14,727 clause pairs extracted from 204 contracts across 8 agreement types. Of these, 3,000 are manually labelled and split into 900 train, 600 dev, and 1,500 test sets. We evaluate 12 models across 4 techniques, with the best macro-F1 of 65.11%, demonstrating that LAUKIN serves as a challenging benchmark. In addition to the labelled data, we also provide 11,727 unlabelled training clause pairs as a resource for future semi-supervised learning research in legal NLP. LAUKIN enables applications including cross-jurisdictional contract review and clause drafting, legal information retrieval, and LLM benchmarking. The LAUKIN dataset and all associated code (covering both, dataset creation and benchmarking) will be made publicly available upon acceptance.

2 Creation of LAUKIN

Figure 1 illustrates the three-stage creation workflow: Legal Clause Collection, Initial Pair Selection, and Manual Annotation.

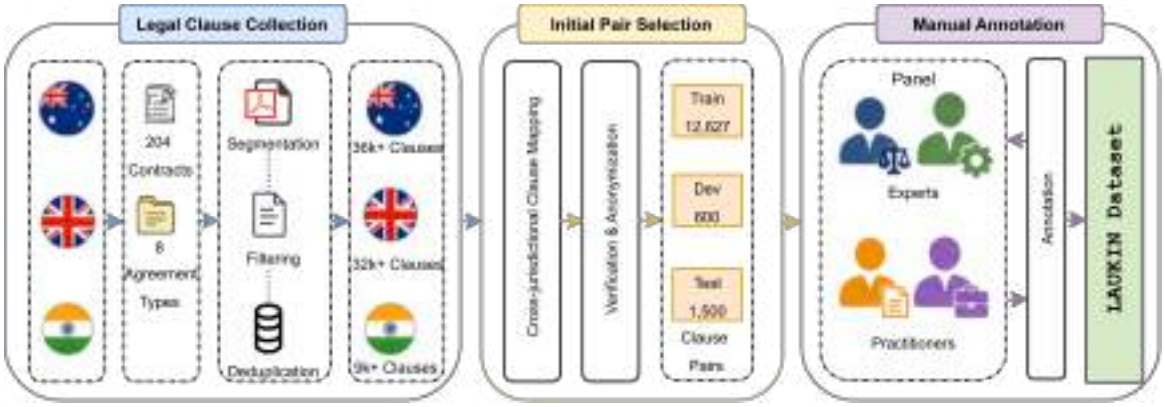

Figure 1: Creation of LAUKIN: Legal Clause Collection (§2.1), Initial Pair Selection (§2.2), and Manual Annotation (§2.3)

2.1 Legal Clause Collection

Contracts for Australia are sourced from AusTender[1], for the UK from Contract Finder[2], and for India from various publicly accessible central and state government portals. All extracted contracts are freely available for public use; the majority are sample or model contracts serving as agreement templates, along with a few expired contracts. The collection process involves manual verification of source correctness and reliability, review of licensing policies across over 100 government state and territory websites, and removal of duplicate contracts, incurring 7 person-days of effort. Finally, a total of 204 contracts are selected across Australia (67), the UK (83), and India (54), covering 8 agreement types as detailed in Table 1. These contracts vary widely in length, ranging from a few pages to 900 pages. UK and Australia contracts reach up to 900 and 750 pages respectively, whereas India contracts are comparatively shorter, with a maximum length of 250 pages. Clauses are automatically

Table 1: Distribution of Contracts and Resultant Clauses

Contract/Agreement Type	Australia	UK	India
Service	7	8	5
Rental & Lease	7	11	10
Franchise	6	7	2
Loan	4	2	4
Marketing & Consultancy	8	9	1
Partnership, NDA & Employment	7	20	16
Construction & Maintenance	17	12	3
Hosting & License	11	14	13
Total Contracts (204)	**67**	**83**	**54**
Total Clauses (w/o Post-processing)	**50,111**	**47,058**	**13,844**
Total Clauses (w Post-processing)	**36,942**	**32,716**	**9,183**

extracted from each contract PDF using `PyMuPDF` (`fitz`). Raw text is parsed page by page and segmented into sentences via regex-based boundary detection. Sentences shorter than 10 words are discarded to eliminate headers, footers, and fragments. This yields a raw corpus of over 50K, 47K, and 13K clauses for Australia, the UK, and India, respectively. Since contracts within the same jurisdiction frequently reuse identical or near-identical clauses, deduplication is applied independently within each jurisdiction using `rapidfuzz`, where a character 4-gram blocking index is constructed to avoid an $O(n^2)$ all-pairs comparison. Pairwise similarity is computed via `fuzz.ratio`, and any two clauses scoring $\geq 90\%$ are grouped into a duplicate cluster. Within each cluster, only the longest clause is retained. This per-jurisdiction deduplication reduces the final corpus to 36K+, 32K+, and 9K+ clauses for Australia, the UK, and India, respectively.

[1]https://www.tenders.gov.au/, CC BY 4.0 AU, accessed 30 Nov 2025.
[2]https://www.contractsfinder.service.gov.uk/, OGL v3.0, accessed 30 Nov 2025.

2.2 Initial Pair Selection

We propose a novel multi-stage retrieval and reranking pipeline for cross-jurisdictional clause mapping. The clause corpus for each jurisdiction, collected in the previous step, serves as input to the pipeline, as shown in Figure 2. As India has the fewest clauses (9,183) compared to Australia and the UK, each Indian clause serves as a query and is matched against the most lexically and semantically equivalent clause from both the Australian and UK corpora. For each corpus, the top-100 candidates are fetched by three retrievers in parallel: BM25 for lexical matching, and MPNet (`all-mpnet-base-v2`) and GTR (`gtr-t5-base`) for semantic retrieval. The resulting ranked lists are merged using Reciprocal Rank Fusion (RRF) [5] into a single fused candidate list per corpus. This results in the top-10 Australian and top-10 UK candidate clauses for each Indian query clause. These candidates are then reranked using a Cross-Encoder (`ms-marco-MiniLM-L-6-v2`), where Australian candidates serve as queries and UK candidates serve as documents. The Australian-UK pair with the highest reranking score is selected as the best match for the Indian query clause, yielding a clause triplet (IN-AU-UK). Once a triplet is formed, each matched clause is removed from its respective corpus, ensuring unique triplet mapping. The final output is a set of 9,183 IN-AU-UK clause triplets. For ethical and legal reasons, all personally identifiable information is anonymised, replacing organisation names, places, and email addresses with `<Organization>`, `<Place>`, and `<Email>` respectively. Next, a qualitative check is performed, as contracts vary in structure, formatting, and drafting style across agreement types and jurisdictions, and may contain incomplete or non-clause segments that convey general information rather than contractual obligations. Any triplet

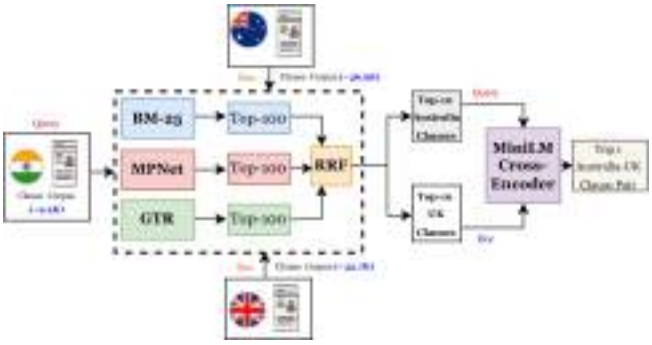

Figure 2: Cross-jurisdictional Clause Mapping Pipeline

containing such segments is removed entirely to ensure data quality. The anonymisation and qualitative check are carried out manually by an annotator, requiring 5 person-days. Finally, 4,909 high-quality triplets remain and are split into three sets: train (4,209 triplets), dev (200 triplets), and test (500 triplets). Each triplet in these sets is decomposed into three clause pairs (AU-UK, UK-IN, IN-AU), resulting in 12,627 train, 600 dev, and 1,500 test clause pairs, totalling 14,727 clause pairs.

2.3 Manual Annotation

For the dataset annotation and task formulation, a panel of two experts and two practitioners is formed. The law expert has 30+ years of experience specialising in contract and commercial law. The NLP expert has 14+ years of experience in dataset creation and methodology development for NLP. Both experts guide the labelling instructions and task design. The two practitioners are a Legal NLP researcher with 4+ years of experience in contract annotation and requirement implementation, and an Associate Lawyer at a law firm specialising in corporate and commercial law. As labelling the entire dataset requires both legal expertise and significant time, a semi-supervised annotation setting is adopted, reflecting the real-world challenge where fully annotated datasets are rarely available. A clause pair is labelled as Equivalent if both clauses share the same core legal function and clause type, irrespective of lexical or syntactic variation. A clause pair is labelled as Not Equivalent if the clauses differ in legal function or clause type, resulting in different legal obligations or outcomes. The two practitioners independently annotate the 600 dev clause pairs. Inter-Annotator Agreement (IAA) is computed, yielding Cohen's $\kappa = 0.75$ (substantial agreement [11]). Disagreements are resolved through iterative rounds of discussion between the experts and practitioners. Following completion of the dev set, the 1,500 test and 900 train clause pairs are annotated by the second practitioner, resulting in a high-quality labelled dataset. The total annotation effort for LAUKIN amounts to 25 person-days.

3 LAUKIN Statistics

LAUKIN consists of 14,727 clause pairs. The clause pairs are split into 12,627 train (900 labelled, 11,727 unlabelled), 600 labelled development, and 1,500 labelled test sets, as shown in Table 2. The dataset is imbalanced, with Equivalent clause pairs forming the minority class, accounting for only 10.6%, 9.7%, and 26.7% of the train, dev, and test splits respectively. This reflects a real-world challenge, where clauses across jurisdictions differ more than they align due to each country's distinct legal system and requirements [10, 14]. We further analyse the word overlap between clause pairs using Jaccard similarity (J). Across all splits, the vast majority of pairs exhibit low word overlap ($J < 0.25$), and even among Equivalent pairs, only 18.9%, 25.9%, and 15.0% of train, dev, and test pairs have $J \geq 0.25$. This suggests that lexical overlap alone is insufficient to determine equivalence. Next, we present examples from LAUKIN

Table 2: LAUKIN Dataset Statistics. Lab. = Labelled; Unlab. = Unlabelled; %HS = percentage of pairs with Jaccard ≥ 0.25

	Train			Dev		Test	
Class	**Unlab.**	**Lab.**	**%HS**	**Lab.**	**%HS**	**Lab.**	**%HS**
Equivalent	–	95	18.9%	58	25.9%	400	15.0%
Not Equivalent	–	805	6.3%	542	6.8%	1100	4.5%
Total (14,727)	**11,727**	**900**	**7.7%**	**600**	**8.7%**	**1,500**	**7.3%**

to illustrate both classes. The Equivalent clause pair in Section 1 shows an Indian and an Australian force majeure clause that exempt parties from liability for delays beyond their control, despite lexical and syntactic differences. In contrast, the Not Equivalent clause pair shown below shares 36% word overlap but differs in legal function: the Australian clause governs direct service delivery by the Consultant, while the UK clause governs personnel provision by the Consultant Company to the Council.

> *Australia:* The Consultant has agreed to provide the Services and any Additional Services in accordance with the terms and conditions of this agreement.
> *UK:* The Consultant Company shall make available to the Council the Individual to provide the Services on the terms of this agreement.

4 Experimental Setup

We introduce the legal equivalence classification task using LAUKIN. Given a clause pair, a model predicts whether the pair is Equivalent or Not Equivalent. We evaluate 12 models across four baseline techniques: Zero-Shot [9], Few-Shot [4], CoT prompting [18], and SFT [13]. For prompting-based techniques, 8 decoder-based models are used via OpenRouter[12]: claude-sonnet-4.6, gemini-3.1-flash-lite-preview, gpt-4.1-mini, gemma-4-26b-a4b-it, llama-4-scout, mistral-small-3.2-24b-instruct, qwen3-6-flash, and phi-4, with temperature set to 0.0 for deterministic output. For SFT, 4 encoder-based models reported in recent legal NLP work [16] are used: two general-purpose models, BERT and RoBERTa, and two legal-specific models, Legal-BERT and Contracts-BERT. Models are trained with 5 epochs, learning rate 2e-5, batch size 64, and maximum sequence length 256. The SFT results are reported as mean ± std over 3 seeds. Macro-F1 is used as the evaluation metric, as the dataset is imbalanced and Equivalent is the minority class.

5 Results and Analysis

Table 3 reports macro-F1 scores across four settings: AU-UK, UK-IN, IN-AU, and Overall. Country-specific scores represent models trained and evaluated on their respective country pairs, while the

overall score represents models trained on all labelled country pairs and evaluated on the whole test set. For AU-UK, the best performance is achieved by CoT using Claude Sonnet 4.6, for UK-IN by Few-Shot using Qwen 3.6 Flash, and for IN-AU by Zero-Shot using Mistral 3.2. Different models and techniques peak on different country pairs, which is expected given the three distinct jurisdictions in LAUKIN. In country-specific settings, prompting-based techniques outperform SFT. However in the overall setting, SFT with BERT achieves the best performance across all techniques and models, suggesting that training on clause pairs from other jurisdictions improves both country-specific and overall performance. Macro-F1 scores vary across country pairs, ranging from 57.75-68.57 for UK-IN, 56.44-66.07 for AU-UK, and 56.21-61.77 for IN-AU. These differences reflect the degree of legal language divergence across jurisdictions. The UK-India pair consistently achieves the highest scores across all models and techniques, as India inherited English legal traditions through British colonial influence, preserving similar clause structures and terminology, making equivalent clauses easier to classify. Although Australia and the UK share a common law heritage, their modern drafting conventions have diverged, resulting in moderate performance on the Australia-UK pair. India-Australia achieves the lowest scores, as both legal systems have evolved independently from their shared colonial origin. Table 4 presents full test set performance stratified by word-overlap range, reporting only the best-performing overall model per technique due to space constraints. Two ranges are considered: low (0.0-0.25) and high (0.25-0.63), where 0.63 is the maximum word overlap in the test set (representing 63% overlap). All best-performing models achieve higher macro-F1 on high word-overlap pairs compared to low word-overlap pairs. SFT with BERT achieves the best overall macro-F1 of 65.11% (Table 3), while achieving 71.12% on high word-overlap pairs and beating all other models and techniques (Table 4), but scoring 62.07% on low word-overlap pairs. This indicates that most errors occur on low word-overlap pairs, making equivalence classification difficult. These results demonstrate that LAUKIN is a challenging and competitive benchmark for LLMs.

Table 3: LAUKIN test set performance across country pairs and overall. For models trained on the training set, results are reported as mean ± std. Bold indicates the best-performing model per technique and setting.

Technique	Model	AU–UK	UK–IN	IN–AU	Overall
		Macro-F1 (500)	Macro-F1 (500)	Macro-F1 (500)	Macro-F1 (1500)
Zero-Shot	Claude Sonnet 4.6	**65.41**	64.61	61.28	63.9
	Gemini 3.1 Flash Lite	62.97	62.12	60.47	62.01
	GPT-4.1 Mini	62.44	61.39	59.55	61.29
	Gemma 4	63.30	65.21	59.53	62.82
	Llama 4 Scout	63.84	60.93	60.85	62.02
	Mistral 3.2	64.84	**65.68**	**61.77**	**64.25**
	Qwen 3.6 Flash	63.54	64.57	59.57	62.65
	Phi-4	56.44	57.75	56.21	56.89
Few-Shot	Claude Sonnet 4.6	64.05	65.58	60.70	63.53
	Gemini 3.1 Flash Lite	63.05	63.74	60.98	62.72
	GPT-4.1 Mini	**65.52**	63.30	**61.52**	63.68
	Gemma 4	65.07	65.00	60.46	63.75
	Llama 4 Scout	61.97	65.52	60.44	62.75
	Mistral 3.2	61.02	64.14	58.06	61.19
	Qwen 3.6 Flash	64.32	**68.57**	60.07	**64.35**
	Phi-4	57.11	61.74	59.05	59.39
CoT	Claude Sonnet 4.6	**66.07**	64.37	60.11	**63.66**
	Gemini 3.1 Flash Lite	62.30	62.70	59.00	61.38
	GPT-4.1 Mini	61.66	**65.48**	58.98	62.12
	Gemma 4	60.04	64.72	59.21	61.29
	Llama 4 Scout	60.81	62.26	56.51	59.95
	Mistral 3.2	63.13	61.45	59.59	61.55
	Qwen 3.6 Flash	62.58	65.26	**61.20**	63.04
	Phi-4	60.49	60.78	57.45	59.69
SFT	BERT	$59.50_{\pm 4.92}$	$\mathbf{64.43}_{\pm 3.37}$	$\mathbf{59.00}_{\pm 3.87}$	$\mathbf{65.11}_{\pm 0.20}$
	RoBERTa	$53.74_{\pm 4.86}$	$56.07_{\pm 6.15}$	$53.61_{\pm 6.66}$	$61.31_{\pm 0.24}$
	Legal BERT	$\mathbf{60.52}_{\pm 4.30}$	$56.60_{\pm 3.06}$	$56.18_{\pm 3.80}$	$62.80_{\pm 0.95}$
	Contracts BERT	$59.13_{\pm 5.36}$	$60.59_{\pm 1.83}$	$58.61_{\pm 3.62}$	$64.18_{\pm 0.24}$

Table 4: LAUKIN test set performance by word-overlap range for the best model per technique

Technique	Best Model	0.0-0.25	0.25-0.63
		Macro-F1 (1391)	Macro-F1 (109)
Zero-Shot	Mistral 3.2	62.60	64.33
Few-Shot	Qwen 3.6 Flash	62.53	65.64
CoT	Claude Sonnet 4.6	61.40	67.40
SFT	BERT	$62.07_{\pm 0.25}$	$71.12_{\pm 3.45}$

6 Conclusion

We introduce LAUKIN, the first multi-jurisdictional common law contract dataset. It comprises 14,727 clause pairs (AU-UK, UK-IN, and IN-AU) from 204 contracts across 8 agreement types. We construct the dataset via a novel multi-stage retrieval and reranking pipeline, with a subset of clause pairs subsequently annotated by legal experts as Equivalent or Not Equivalent. We evaluate 12 models across four techniques. The best macro-F1 of 65.11% establishes LAUKIN as a challenging benchmark. Prompting-based techniques outperform SFT in country-specific settings; however, SFT with BERT achieves the best overall performance, suggesting that training on clause pairs from other jurisdictions improves performance. Low word-overlap clause pairs remain the primary source of error, posing a significant challenge for legal equivalence classification. LAUKIN also facilitates future semi-supervised learning research. The dataset enables applications including cross-jurisdictional contract review and clause drafting, legal information retrieval, and LLM benchmarking. LAUKIN advances legal NLP beyond single-jurisdiction benchmarks toward real-world cross-jurisdictional legal contract understanding.

Acknowledgment

We thank Praharsh Singh of LexStart Partners for his contributions to the LAUKIN annotation and for his discussions on the legal aspects of this work.

GenAI Usage Disclosure

This paper is written entirely by the authors without any generative AI tools for text generation or writing assistance. All manuscript content, figures, and experimental results are human-produced. AI-based writing assistants are used solely for optional spelling and grammar checks.

References

[1] Mousumi Akter, Erion Çano, Erik Weber, Dennis Dobler, and Ivan Habernal. 2025. A comprehensive survey on legal summarization: Challenges and future directions. *Comput. Surveys* 58, 7 (2025), 1–32.

[2] Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *Comput. Surveys* 58, 6 (2025), 1–37.

[3] Katy Barnett. 2024. Reflections on the principles of remoteness in contract in comparative law. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique* 37, 5 (2024), 1587–1616.

[4] 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 (2020), 1877–1901.

[5] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 758–759.

[6] Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6472–6485.

[7] Jia-Hong Huang, Chao-Chun Yang, Yixian Shen, Alessio M Pacces, and Evangelos Kanoulas. 2024. Optimizing numerical estimation and operational efficiency in the legal domain through large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4554–4562.

[8] Daniel N Kluttz and Deirdre K Mulligan. 2019. Automated decision support technologies and the legal profession. *Berkeley Technology Law Journal* 34, 3 (2019), 853–890.

[9] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*.

[10] Manasi Kumar and Maren Heidemann. 2022. Contract law in common law countries: A study in divergence. *Liverpool Law Review* 43, 2 (2022), 133–147.

[11] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[12] OpenRouter. 2023. OpenRouter: The Unified Interface For LLMs. https://openrouter.ai. Accessed: 2026-05-26.

[13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[14] Madhav Saxena. 2023. Comparative Study between Contract of Indemnity and Guarantee vis-a-vis Provisions of UK and India. *Issue 3 Int'l JL Mgmt. & Human.* 6 (2023), 1366.

[15] Amrita Singh, Aditya Joshi, Jiaojiao Jiang, and Hye-young Paik. 2025. A survey of classification tasks and approaches for legal contracts. *Artificial Intelligence Review* 58, 12 (2025), 380.

[16] Amrita Singh, H Suhan Karaca, Aditya Joshi, Hye-young Paik, and Jiaojiao Jiang. 2026. Evaluating Customized vs. Generalist Transformer-based Models for Legal Contract Classification. In *Proceedings of the 2nd Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*.

[17] Onyinye Ucheagwu-Okoye and Chidimma Stella Nwakoby. 2025. LEGAL PRACTICE AND LEGAL EDUCATION IN THE 21ST CENTURY: HARNESSING AI FOR EFFICIENCY, SPEED AND ACCURACY. *Chukwuemeka Odumegwu Ojukwu University Law Journal* 9, 1 (2025).

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.