Can Large Language Models Reliably Extract Jurisdictional Variations? An Empirical Study on UK Statutory Texts

1st Safia Kanwal
Hillary Rodham Clinton School of Law
Swansea University
Swansea, UK
safia.kanwal@swansea.ac.uk
0000-0003-2041-9696

2nd Livio Robaldo
Hillary Rodham Clinton School of Law
Swansea University
Swansea, UK
livio.robaldo@swansea.ac.uk
0000-0003-4713-8990

3rd Hafsa Dar
Department of Software Engineering
University of Gujrat
Gujrat, Pakistan
hafsa.dar@uog.edu.pk
0000-0003-4538-6632

4th Davide Liga
Department of Computer Science
University of Luxembourg
Luxembourg
davide.liga@uni.lu
0000-0003-1124-0299

5th Joseph Anim
Hillary Rodham Clinton School of Law
Swansea University
Swansea, UK
joseph.anim@swansea.ac.uk
0009-0006-3073-5093

***Abstract*—Structured information extraction is now a common use case for large language models (LLMs), powering applications from document analysis to regulatory search. But as these models are increasingly used in high-stakes domains like law, the challenge of evaluating what they extract—and what they miss—becomes critically important. In legal texts especially, where provisions can differ subtly across jurisdictions, reliable extraction is not just a technical convenience but a prerequisite for trustworthy automation. In this work, we propose a multi-dimensional evaluation framework for assessing LLM-based extraction of territorial distinctions in legislation. Our method captures three key dimensions: how well relevant sections are covered, how accurately territorial mappings are assigned, and how closely the extracted text matches the original law. Together, these form a composite score and status indicator that help identify failures and guide re-extraction. Designed for use across full legislative Acts, our approach is especially valuable in contexts like legal agent development, where overlooked provisions or misclassifications can compromise downstream reasoning. The tool's source code can be accessed through the GitHub repository https://github.com/SafiaK/UKTerritorialDisambiguation**

***Index Terms*—Large Language Model(LLM), LLM extraction validation, UK legislation**

I. Introduction

There are four distinct legal jurisdictions within the United Kingdom: England, Wales, Scotland, and Northern Ireland. Although a single Act of Parliament may cover all of these territories, its provisions often differ in their application across jurisdictions. Therefore, when processing legislation within any legal information system or legal workflow, it is essential to accurately identify which territorial provisions apply, as misinterpretation may lead to incorrect legal reasoning or outcomes. This territorial complexity is particularly pronounced in areas of devolved competence where legislative provisions may establish separate regulatory bodies, define different procedural requirements, or create distinct legal obligations for each jurisdiction. For example, Section 47 of the Education Act 2005[1] defines "denominational education" differently in England and Wales - in England, it is specified as religious education required by the Education Act 2002 but not governed by an agreed syllabus, while in Wales it is defined as teaching related to Religion, Values and Ethics under the Curriculum and Assessment (Wales) Act 2021 and relevant trust deed provisions.

Legal professionals must carefully examine each provision to determine its territorial scope, compare parallel provisions across jurisdictions, and ensure compliance with territory-specific requirements. This process is time-consuming, error-prone, and becomes increasingly complex as the volume of legislation grows and territorial arrangements evolve. The challenge is further compounded by inconsistent drafting patterns, where territorial scope may be indicated through explicit references, implicit jurisdictional boundaries, or cross-references to other provisions.

Moreover, the mere mention of a territory within legislation does not necessarily imply that the provision has a different legal effect. For instance, many sections contain *explicit territorial extent clauses* that formally state where a rule applies without introducing substantive divergence. A common example reads: *"This section extends to England and*

979-8-3315-7480-2/25/$31.00 © 2025 IEEE

[1] https://www.legislation.gov.uk/ukpga/2005/18

Wales"[2] Such clauses clarify applicability but do not signal any difference compared to other regions.

Other provisions include *uniform references* for clarity, even when the operative text remains identical across jurisdictions. For example, Section 114 of the Education Act 2005 defines the term "regulations" separately for England and Wales to avoid ambiguity about which body exercises the power: *"'regulations' means— (a) in relation to England, regulations made by the Secretary of State, and (b) in relation to Wales, regulations made by the Assembly."*[3] Even though the substance of the regulations may be the same, the provision enumerates territorial references to ensure consistency with devolved governance structures.

In some cases, provisions highlight the scope of *reserved powers and devolution*, where a part of the Act formally applies to a devolved territory without modifying the content. For example: *"Inspection of careers services in Wales"*[4]. Such statements acknowledge constitutional arrangements while preserving identical substantive obligations.

The current NLP tools and legal systems fall short in differentiating between territorial references, that signal substantive legal differences with those that only presents procedural purposes. Existing legal information systems and NLP approaches provide limited support for automatically identifying and interpreting these nuances. Our work aims to address this gap by developing a robust, multi-dimensional evaluation framework capable of systematically capturing and validating territorial distinctions across entire legislative acts.

A. Research Objectives

The primary research objectives are:

- To develop a systematic framework for identifying and extracting territorial differences from legislative text.
- To adapt and implement validation methodologies that enable quality assessment without requiring extensive manual annotation. Drawing inspiration from the MINEA framework, we develop a cross-validation approach that leverages pre-identified territorial information as ground truth.
- To demonstrate the practical application of this methodology through a comprehensive case study of the different UK acts.

B. Research Contributions

The research makes several key contributions to legal NLP and information extraction methodology. We introduce a novel approach to territorial difference extraction that combines domain-specific prompt engineering with systematic validation techniques. Our validation framework adapts recent advances in LLM evaluation to legal domain applications, demonstrating how synthetic validation concepts can be applied using actual legislative content rather than artificially generated data.

[2]https://www.legislation.gov.uk/ukpga/2005/18/section/127
[3]https://www.legislation.gov.uk/ukpga/2005/18/section/114
[4]https://www.legislation.gov.uk/ukpga/2005/18/section/55

The remainder of this paper is structured as follows. Section II reviews related work. Section III presents our comprehensive methodology. Section IV draws the details of experiments and state the results applying this methodology to different UK acts, including detailed validation outcomes and error analysis. Section V presents validation of the work, and Section VI concludes this study.

II. Related Work

Artificial intelligence-based data extraction has been consistently shown to fall short of manual methods, particularly when applied to nuanced or multi-component content in high-stakes domains. Recent legal research highlights these limitations in professional contexts where accuracy is paramount. [1] conducted the first preregistered empirical evaluation of AI-driven legal research tools, testing claims by major legal research providers that their tools "eliminate" or "avoid" hallucinations. The study found that AI research tools made by LexisNexis (Lexis+ AI) and Thomson Reuters (Westlaw AI-Assisted Research and Ask Practical Law AI) each hallucinate between 17% and 33% of the time, demonstrating that providers' claims are overstated. This pattern underscores the difficulty of reliably capturing intricate legal details even in specialized, professionally-developed systems. Similar validation studies in legal AI confirm these limitations where a comprehensive analysis of AI [2] was performed by contract review systems. The findings showed that while specialized legal AI tools achieved 94% accuracy in spotting risks in non-disclosure agreements (NDAs), this performance was significantly higher than the 85% accuracy rate achieved by experienced lawyers, yet lack variability in legal contexts.

The MINEA (Multiple Infused Needle Extraction Accuracy) framework introduced a novel approach for evaluating extraction tasks without pre-labeled data by embedding synthetic "needles" into source text and measuring retrieval performance, according to [3]. This technique achieved 0.780 accuracy across 695 entities and has proven valuable in domains like law, where gold-standard annotations are scarce. The AutoNuggetizer framework tackled similar evaluation challenges in multi-document synthesis, demonstrating strong correlation ($\tau = 0.887$–0.901) between automated and human judgments, proposed by [4]. Such strategies are particularly relevant for retrieval-augmented generation (RAG) pipelines in legal research.

Entity-centric evaluation metrics also represent an important advancement. [5] introduced the AESOP (Approximate Entity Set Overlap) metric to capture multi-property entity extraction performance more comprehensively. Human evaluators consistently preferred AESOP-based assessment over simpler measures of precision and recall, emphasizing the importance of completeness and correctness. [6] further demonstrated that word-level F1 severely underestimates model performance compared to expert evaluation, especially in specialized domains such as law.

Chang et al. [7] identified terminology gaps, limited knowledge depth, regulatory compliance constraints, and context

sensitivity as persistent challenges in legal NLP. For instance, Martin et al. found GPT-4 matched junior lawyer accuracy (F1=0.87) in contract review while achieving drastic cost reductions and faster completion times, stated by [8]. These findings illustrate both the potential efficiency gains and the heightened risk of critical errors in legal applications.

According to [9], the emergence of LLMs offers promising opportunities for addressing these challenges through automated information extraction capabilities. Recent work by [10] and [11] has demonstrated LLMs' effectiveness in legal domain tasks, including regulatory compliance analysis and legislative summarization. However, the application of LLMs to territorial difference extraction faces significant methodological challenges, particularly in validation and quality assurance.

A critical gap exists in evaluation methodologies for specialized legal information extraction tasks where traditional benchmark datasets are unavailable. Most legal NLP evaluation relies on manually annotated datasets or established benchmarks given by [12], but territorial difference extraction represents a highly specialized task for which no gold-standard annotations exist. This creates a validation challenge: How can the accuracy and completeness of territorial difference extraction be assessed without extensive manual annotation by legal experts?

III. Research Methodology

As shown in figure 1,methodology is given in five different steps. We have detailed each stage in the following subsections.

A. Data Acquisition and Preprocessing

We retrieved the official XML representations of relevant Acts from the UK National Archives' legislation.gov.uk API, which are un the LegalDocML(XML) format. The UK Publication Office (The National Archives, TNA) is the single institution in the world providing all its legislation and case law in LegalDocML format. Therefore, the tool is not currently applicable to other jurisdictions.

An overview of methodology is shown in the figure 1

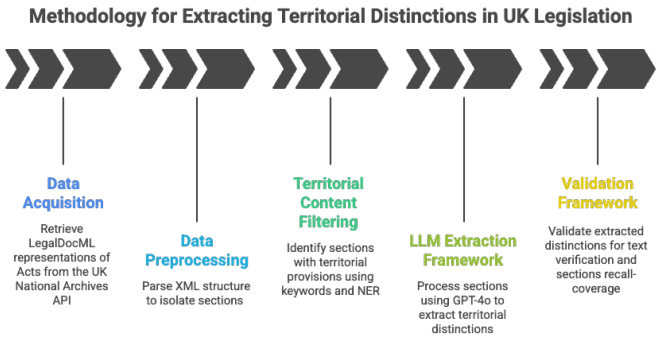

Fig. 1. Overview of the methodology pipeline used to extract and validate territorial distinctions in UK legislation.

Following best practices in legal document processing [13], the raw text each `<section>` was preserved while removing formatting artifacts to ensure downstream processing accuracy.

B. Territorial Content Filtering

We developed an automated filter using both rule-based pattern matching and Named Entity Recognition (NER) recognition. The filter scanned sections for explicit mentions of territorial keywords including "England," "Wales," "Scotland," and "Northern Ireland," as well as compound references such as "England and Wales." The filtering strategy employed two complementary approaches: (i) regex-based detection of territorial mentions within legal contexts using patterns such as `(in|for|applies to|extends to|except in) ([A-Za-z ]+)`, and (ii) spaCy's GPE (Geo-Political Entity) [14] recognition to capture territorial references that might be missed by rule-based approaches. After this phase each section have its corresponding territory/territories information.

C. Large Language Model Extraction Framework

Each candidate section was processed using GPT-4o with carefully engineered prompt designed to output structured JSON capturing: (i) provisions with explicit territorial scope mapped to specific territories, (ii) difference types categorized as content, scope, penalty, or procedure , (iii) concise explanations of territorial divergence, and (iv) verbatim text excerpts supporting each territorial claim. The prompt engineering process was informed by recent advances in legal information extraction, stated by [15] and incorporated domain-specific terminology and context. Prompts were iteratively refined through multiple rounds of testing and the final prompt was use a structured extraction output strategy. To support systematic extraction and classification, It is prompted to identify each territorial distinction with a `difference_type` label indicating the nature of the variation. These include *content* (substantive legal differences in wording or meaning), *scope* (differentiated territorial applicability), *penalty* (territory-specific sanctions), *procedure* (variation in legal or administrative process), and *other* for residual cases that do not fit standard categories. These distinctions reflect the types of divergence illustrated in Section I, where provisions may vary across jurisdictions due to devolution, policy differences, or legal system structure.

For sections where no substantive difference is found, we group them using `category` labels under `other_sections`. These categories include: *explicit_extent_clause* (where territorial applicability is stated without altering content), *uniform_provision_with_territorial_reference* (where identical rules mention territories for drafting clarity), *reserved_powers_devolution* (where references acknowledge devolution without legal divergence), and *other* for provisions that do not fall into these classes. These categories are essential for disambiguating structural references from genuine legal divergence in the UK's multi-jurisdictional legislative framework. For sections referencing multiple

territories or containing complex territorial relationships, we implemented a multi-section comparative analysis approach. Related provisions were grouped together and processed in a single prompt to enable the LLM to identify nuanced cross-territorial patterns and dependencies.

D. Validation Framework

To check how reliable the extractions were, we built a custom validation process designed specifically for UK legislation. We developed a domain-adapted validation framework tailored [16] to the unique requirements of UK legislative text. The framework combines deterministic text verification and similarity text matching techniques to address the absence of gold-standard annotations. For each extracted territorial difference, a custom Python validation script performed systematic checks against the original statutory text. Specifically, the script implemented robust text preprocessing routines that normalized case, removed ellipsis markers and non-alphanumeric characters, and standardized whitespace to reduce superficial discrepancies. Validation comprised three main stages: (i) territorial presence verification; (ii) text alignment verification; and (iii) consistency checks. The code with the prompt is in the github repository and is not put here due to the confined number of pages.

IV. Experiments and Results

A. Legislative Corpus Selection

Our experimental evaluation focused on six diverse pieces of UK legislation to test the robustness and generalization of our territorial difference extraction methodology. The selected Acts represent different legislative domains and territorial complexity patterns including Wildlife and Countryside Act 1981, Education Act 2005, Housing (Wales) Act 2014, Housing and Planning Act 2016, Housing and Regeneration Act 2008, and Social Housing (Regulation) Act 2023.

These Acts collectively provide 245 territorial sections spanning different legislative domains, temporal periods, and territorial complexity levels, enabling comprehensive evaluation of our methodology across diverse legal contexts.

B. Validation Strategy Implementation

Our validation strategy operationalizes a systematic mapping between the structured extraction schema and quantitative performance metrics. Each metric leverages specific fields in the model output JSON to assess the completeness and correctness of extraction.

Coverage Verification (C): Coverage quantifies whether the extraction system processes all relevant input sections, regardless of extraction quality or label accuracy. For every legislative Act, we first computed the set of unique input sections (S_{in}) supplied to the LLM. Next, we extracted the union of all sections listed in the `provisions` field across both `territorial_differences` and `other_sections` in the output. This set of matched sections is denoted S_{out}. Coverage is computed as:

$$C = \frac{|S_{out}|}{|S_{in}|} \times 100\%.$$

High coverage ($C = 100\%$) indicates that no sections were omitted in the output JSON. For example, if an Act includes 50 sections in S_{in} but the LLM output lists provisions for only 45, coverage would be 90%. This metric ensures that even sections without substantive territorial differences are accounted for via their inclusion in `other_sections`.

Section–Territory Mapping Accuracy (MA): This metric assesses whether the system correctly identifies which territories apply to each section. Specifically, for each provision dictionary entry in `territorial_differences` and `other_sections`, the validator script compares the assigned list of territories to the set of explicit territory mentions in the original legislative text. A provision is considered correct if all claimed territories were verifiably mentioned in the section. The mapping accuracy is computed as the proportion of provisions with correct territory assignments:

$$MA = \frac{\text{Number of provisions with correct TE}}{\text{Total provisions extracted}} \times 100\%.$$

where TE = territorial mapping

Text Extraction Accuracy (TE): This metric measures fidelity between the extracted text snippets and the source content. For every entry in the `territory_texts` field of each territorial difference, the script checks whether the snippet either occurs as a clean substring of the section text or achieves a similarity text match above a similarity threshold ($\geq 90\%$). A record is counted as accurate if all its territory-specific snippets satisfy this criterion. Text extraction accuracy is thus computed as:

$$TE = \frac{\text{Number of TD with accurate text evidence}}{\text{Total TD extracted}} \times 100\%.$$

where TD = territorial differences **Overall Score (OS)**: To compute an overall score reflecting completeness and correctness, we define a composite metric combining these three components. Let:

$$w_C, w_M, w_T \in [0, 1]$$

be weights (uniformly set to 1 for equal contribution), then the Overall Score (OS) is computed as:

$$OS = \frac{C + M_A + T_E}{3}.$$

This formulation ensures that both section inclusion (coverage) and per-record precision (mapping and text extraction) contribute equally to the aggregate measure.

Metric Interpretation and Schema References:

- *Coverage* derives from all entries in the `provisions` field across both `territorial_differences` and `other_sections`.

- *Mapping Accuracy* evaluates correctness of the assigned territories per provision (referencing `provisions` in comparison with actual section text).
- *Text Extraction Accuracy* focuses exclusively on `territory_texts` fields within `territorial_differences`, as `other_sections` do not require text snippet evidence.

Metric Notation: IS = Input Sections, C = Coverage, MA = Mapping Accuracy, TE = Text Extraction Accuracy, OS = Overall Score, P = Passed, NR = Needs Review.

As shown in Table I, the framework provides granular visibility into system performance across Acts and metrics.

TABLE I
TERRITORIAL LLM VERIFIER METHOD 2 EVALUATION RESULTS ACROSS MULTIPLE LEGISLATIVE ACTS

Legislation	IS	C(%)	MA(%)	TE(%)	OS(%)	Status
Wildlife and Countryside Act 1981	57	100.0	100.0	100.0	100.0%	P
Education Act 2005	45	100.0	93.33	100.0	96.94	P
Housing (Wales) Act 2014	32	96.88	100.0	100.0	98.46	P
Housing and Planning Act 2016	71	88.73	100.0	100.0	94.29	P
Housing and Regeneration Act 2008 (Run 1)	35	57.14	100.0	100.0	75.0	NR
Housing and Regeneration Act 2008 (Run 2)	35	91.43	100.0	100.0	95.95	P
Social Housing (Regulation) Act 2023	5	100.0	100.0	0.0*	100.0	P

Note: * For the Social Housing (Regulation) Act 2023, text extraction accuracy (TE) is marked as 0.0% because no territorial differences requiring snippet extraction were identified. All sections were classified under `other_sections`, where no `territory_texts` fields are present, resulting in an effective TE denominator of zero. As such, the 0.0% value does not indicate an error but reflects the absence of extractable text evidence in this specific case.

The results demonstrate strong overall performance, with most Acts achieving `PASSED` status and overall scores exceeding 90%. Notably, text extraction accuracy reached 100% in all cases, confirming that when territorial provisions were identified, the extracted snippets reliably matched the original legislative content.

Performance Analysis: Coverage varied between 57.14% and 100%, reflecting the difficulty of comprehensively identifying territorial references in more complex legislation.

As LLM hallucinate sometimes but works well in other scenarios. For instance, in above case when we didn't get a good score for "Housing and Regeneration Act 2008" in first run - we rerun it - If same things goes on multiple times, we can eventually change our methodology of extraction.

Error Analysis:

Missing sections typically involved complex provisions where territorial scope was implied through cross-references or embedded within procedural specifications rather than explicitly stated.

Mapping errors occurred primarily in provisions where territorial scope extended beyond explicitly mentioned jurisdictions, requiring inference about implicit territorial application.

V. MANUAL EXPERT VALIDATION

We conducted manual validation to assess classification accuracy between "territorial differences" and "other sections" categories, which cannot be automatically validated. We examined 45 sections from the Education Act 2005 and 35 sections from the Housing and Regeneration Act 2008. The classification results are as follows.

TABLE II
MANUAL CLASSIFICATION VALIDATION RESULTS

Act	Total Sections	Correctly Classified	Wrongly Classified	Accuracy (%)
Education Act 2005	45	45	0	100.0
Housing Act 2008	35	34	1	97.1
Combined	**80**	**79**	**1**	**98.8**

In Table II, Education Act 2005 and Housing Act 2008 are manually classified with coverage of total sections 45 and 35 respectively. One section out of 35 in Housing Act was wrongly classified which indicates 97.1% accuracy in Housing Act sections and 100% in Education Act. Overall 98.8% accuracy is attained during validation.

A. Correctly Classified Examples

Sections 48/50 (Education Act) - Procedural Difference:

- England (Section 48): "An inspection under this section is to be conducted by a person chosen by the governing body"
- Wales (Section 50): "An inspection under this section is to be conducted by a person chosen... The person chosen need not be registered as an inspector under section 25"

Sections 60-61 (Housing Act) - Content Difference:

- England: "This Part replaces the system of 'registered social landlords' under Part 1 of the Housing Act 1996"
- Wales: "That Part will continue to apply in relation to Wales"

1) Other Sections: **Section 19 (Education Act) - Devolved Administration**: Correctly classified as "other section" containing Wales-specific inspector appointment procedures without creating substantive legal differences.

Section 55 (Housing Act) - Uniform Provision: Correctly classified as "other section" applying notice procedures uniformly across the United Kingdom with territorial reference for jurisdictional clarity.

B. Misclassification Analysis

1) Education Act 2005: No sections were misclassified between territorial differences and other sections categories. All mapping errors involved correct identification of territorial sections but incorrect territorial assignments (e.g., assigning "England" instead of "England and Wales").

2) Housing Act 2008: One section was misclassified into the wrong category (97.1

Section 149 - Exempted disposals: Incorrectly classified as a territorial difference when it should have been classified as "other section." This section contains administrative exemptions for different tenancy types rather than substantive territorial differences in legal obligations.

The remaining errors were coverage errors (3 sections missed entirely) and mapping errors (incorrect territorial assignments), not classification errors between categories.

VI. Conclusion

In this paper, a novel methodology for the systematic extraction and validation of territorial differences in UK legislation is presented using LLMs. We addressed a critical gap in legal information systems - the ability to automatically differentiate provisions that apply differently across the UK's complex devolved jurisdictions. Our validation framework, demonstrated that quantitative assessment of extraction performance is feasible even in the absence of gold-standard annotations. Through comprehensive evaluation across multiple acts, we showed that the approach consistently achieved high mapping and text extraction accuracy, while coverage improved substantially with iterative prompt refinement. The capacity to rerun the validation and track incremental improvements ensures that this methodology can be systematically evolved over time. Furthermore, this work contributes tools and insights that can be directly applied by legal practitioners, researchers, and policymakers seeking to navigate multi-jurisdictional statutory obligations.

Future research focuses on scaling the approach to larger legislative corpora, incorporating advanced techniques for detecting implicit territorial references, and integrating retrieval-augmented generation workflows to further improve coverage. Additionally, the creation of domain-specific benchmark datasets for territorial difference extraction remains an important direction to support broader comparative evaluation.

Acknowledgments

This work is part of the *Odyssey Project* (https://www.liviorobaldo.com/odyssey.html), sponsored by Innovate UK under the "Professional & Financial Services Data Access Demonstrators: ESG" competition, which provided crucial funding. We also acknowledge the National Archives(https://www.nationalarchives.gov.uk/) efforts to maintain the legislation in LegalDocML which we used for experiments.

References

[1] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, "Hallucination-free? assessing the reliability of leading ai legal research tools," *arXiv preprint arXiv:2405.20362*, 2024, forthcoming in Journal of Empirical Legal Studies.

[2] "Ai in contract drafting: Transforming legal practice," *Richmond Journal of Law and Technology*, October 2024, available at: https://jolt.richmond.edu/2024/10/22/ai-in-contract-drafting-transforming-legal-practice/.

[3] D. Seitl, T. Kovárík, S. Mirshahi, J. Kryštufek, R. Dujava, M. Ondreicka, H. Ullrich, and P. Gronat, "Assessing the quality of information extraction," *arXiv preprint arXiv:2404.04068*, 2024.

[4] R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, and J. Lin, "The great nugget recall: Automating fact extraction and rag evaluation with large language models," 2025. [Online]. Available: https://arxiv.org/abs/2504.15068

[5] H. Wu, Y. Yuan, L. Mikaelyan, A. Meulemans, X. Liu, J. Hensman, and B. Mitra, "Structured entity extraction using large language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, USA: Association for Computational Linguistics, Nov. 2024.

[6] J. Dagdelen *et al.*, "Structured information extraction in specialized domains," *Computational Linguistics*, 2024.

[7] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, Mar. 2024. [Online]. Available: https://doi.org/10.1145/3641289

[8] L. Martin, N. Whitehouse, S. Yiu, L. Catterson, and R. Perera, "Better call gpt, comparing large language models against lawyers," *arXiv preprint arXiv:2401.16212*, 2024, aI Centre of Excellence, Onit Inc., Auckland, New Zealand. [Online]. Available: https://arxiv.org/abs/2401.16212

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, S. Girish, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[10] S. J. Blair and M. K. Chen, "Large language models for legal compliance analysis: Opportunities and challenges," *Stanford Technology Law Review*, vol. 26, no. 1, pp. 45–78, 2023.

[11] A. Kornilova and V. Eidelman, "Bilingual experiments on automatic generation of legal text summaries," in *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, 2018, pp. 112–121.

[12] N. Guha *et al.*, "Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models," in *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

[13] D. M. Katz *et al.*, "Natural language processing in the legal domain," *Journal of Legal Technology*, vol. 12, no. 3, pp. 45–72, 2023.

[14] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[15] I. Chalkidis *et al.*, "Lexglue: A benchmark dataset for legal language understanding in english," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 4310–4330.

[16] D. Seitl *et al.*, "Assessing the quality of information extraction," *arXiv preprint arXiv:2404.04068*, 2024.