Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 270 (2025) 2096–2105

Procedia Computer Science

www.elsevier.com/locate/procedia

29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)

Lawyering Up: An Evaluation of Large Language Models' Potential for Transforming Business Law Practice

Sascha Schweitzer[a,b,c,*], Navid Sheikh Mouchegani[c], Markus Conrads[c]

[a]*Shenzhen Technology University, 3002 Lantian Road, 518118 Shenzhen, China*
[b]*University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany*
[c]*Reutlingen University, Alteburgstraße 150, 72762 Reutlingen, Germany*

Abstract

Recent developments in Large Language Models (LLMs) and advanced conversational agents (CAs) suggest a rapidly evolving capability to handle complex legal tasks. Building on our previous study—which established a benchmark of 60% accuracy in addressing German business law tasks—this paper presents new findings indicating that current-generation LLMs such as Gemini 2.5 Pro can now successfully solve more than 70% of all cases, and more than 80% in specific domains. Moreover, a longitudinal analysis reveals an average annual improvement rate of approximately 13.68%. To arrive at these findings, we administered a corpus of diverse legal prompts to prominent CAs, including different versions of ChatGPT, Claude, Gemini, and emerging model families such as DeepSeek. Performance and reliability were examined, revealing persistent challenges such as domain-specific knowledge gaps. Despite these constraints, the study investigates whether LLM-based tools are beginning to integrate into legal workflows in a manner that is consistent with the theory of disruptive innovation. From automating routine legal processes to supporting complex contract reviews, these systems exhibit the capacity to boost efficiency and reshape the role of legal professionals. At the same time, their limitations underscore the necessity of expert oversight—particularly for novices relying on AI-generated legal advice. By illuminating both the strengths and shortcomings of today's advanced CAs, this study provides guidance for legal practitioners, policymakers, and researchers aiming to apply LLM technology in business law and beyond.

© 2025 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the KES International.

Keywords: Legal information systems; disruptive innovation; conversational agents; large language models; performance assessment

* Corresponding author.
E-mail address: Sascha.Schweitzer@Reutlingen-University.de

1877-0509 © 2025 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the KES International.
10.1016/j.procs.2025.09.330

1. Introduction

Over the past year, large language models (LLMs) have witnessed rapid and significant advancements—both in terms of raw performance and their application within high-stakes professional contexts [1], [2], [3]. Recent releases such as OpenAI's o1 have extended the capabilities of conversational agents (CAs) far beyond their predecessors. As a result, legal information systems stand at a critical juncture: the shift from experimentation to operational integration is underway. In light of these changes, the question is no longer whether LLMs can process legal information, but rather how reliably they can do so across jurisdictions, use cases, and over time [4].

This paper builds on our prior work [5], [6], which established a benchmark of approximately 60% task-solving accuracy for CAs in the context of German business law. Through the application of a broader set of agents, we now revisit this benchmark to assess the latest generation of LLMs and their current performance trajectory. Our findings indicate a substantial leap in performance—approaching 80% success rates across a certain array of tasks for Google Gemini 2.5 Pro Preview. This improvement suggests not only technical progress but also a growing maturity in how LLMs reason through legal problems and interface with structured legal language.

Recent studies have underscored the growing potential of LLMs within legal and compliance domains. For instance, [7] offer a comprehensive overview of emerging legal LLMs and their applications worldwide, emphasizing both current limitations and future research directions in this rapidly evolving field. Meanwhile, [8] investigate LLMs across critical sectors—including law—highlighting their transformative capacity to refine legal interpretation and compliance strategies, while also calling attention to ethical considerations and the need for robust regulatory frameworks. Yet, systematic and comparable evaluations of CA performance within non-English jurisdictions—especially over time and across agent variants—remain scarce. Our research addresses this gap by analyzing how model quality and reliability have evolved in the German legal domain and what this means for deployment in practice.

To this end, we administered our curated corpus of German legal prompts to evaluate current-generation CAs, including DeepSeek R1, Grok 3 Thinking, ChatGPT-4.5 Preview, Gemini 2.5 Pro Preview, and Llama 4 Maverick. We assessed not only performance metrics but also patterns in typical failure modes, legal reasoning fidelity, and domain coverage gaps. The goal is to offer a longitudinal and comparative understanding of model progression and to identify implications for legal information systems, professional practice, and AI governance.

Our contribution is threefold: (1) we provide empirical evidence of CA performance growth over time in a Central European legal context; (2) we compare domain-specific performance trends—covering the formation and validity of contracts, contract violations, employment law, and other domains of business law—to clarify the scope and consistency of improvements; and (3) we discuss whether these developments align with a trajectory of disruptive innovation, highlighting the potential for LLM-based solutions to influence legal workflows and the need for expert oversight. Ultimately, this study aims to support legal professionals, technologists, and policymakers in navigating the rapidly evolving landscape of legal AI.

The structure of the paper is as follows: Section 2 reviews related work and situates recent LLM developments within the broader trajectory of legal AI. Section 3 details our methodology and evaluation framework. Section 4 presents empirical results and error analysis. Section 5 concludes with implications for the integration of LLMs into legal practice and outlines avenues for future research.

2. Literature

Recent studies indicate that large language models (LLMs) have made significant progress in performing legal tasks, with some models approaching or even surpassing human-level performance on specific benchmarks. Notably, OpenAI's GPT-4 demonstrated a substantial leap in capability compared to its predecessor, GPT-3.5. According to the GPT-4 Technical Report [9], GPT-4 achieved a score around the 90th percentile on the Uniform Bar Exam (UBE), whereas GPT-3.5 scored approximately in the 10th percentile. This performance gap underscores the rapid evolution in LLM proficiency on complex legal reasoning tasks in the first years since their public release. Further advancements are seen in the LMSYS Chatbot Arena Leaderboard [10]. While GPT-4o consistently still ranks among the top based on human preference comparisons across a variety of tasks, including analytical and writing assistance tasks, at the time of writing, this LLM has been (slightly) outperformed by its successor GPT-4.5-

Preview, as well as Grok-3-Preview and Google-2.0-Flash-Thinking (which is not among the current top-3 performers in this study, as displayed by Figure 1).

Researchers have introduced numerous legal-specific evaluation benchmarks to measure these capabilities. Notably, the collaborative LegalBench suite compiled 162 diverse legal reasoning tasks (covering case analysis, statutory interpretation, etc.) and evaluated 20 different LLMs, revealing varied strengths and weaknesses across models [11]. Results from these and other benchmarks indicate that state-of-the-art LLMs can comprehend legal text, answer legal questions, and perform classification tasks with a high degree of accuracy in zero-shot or few-shot settings. For instance, [12] found a GPT-4 based model achieved 97% accuracy in zero-shot classification in one dataset of annotated privacy policy segments. However, performance varies by jurisdiction and task type, suggesting LLMs are not universally "solved" for all legal contexts. Common law vs. civil law differences are one consideration: models optimized on English common-law data may falter on civil law tasks or other languages [11].

Beyond benchmark performance, researchers are examining how LLMs can be integrated into real-world legal workflows to augment or streamline professional practice. Early evidence suggests that LLM-based tools can dramatically increase efficiency in tasks like document review, legal research, and drafting: [13] directly compared advanced LLMs against junior lawyers in a contract review scenario and expect a "seismic shift in legal practice" – the LLMs matched or exceeded human accuracy in spotting legal issues, completed reviews in seconds (versus hours for humans), and did so at an over 99% cost reduction. In their review, [14] emphasize that successful integration often involves human-AI collaboration, where the LLM handles labor-intensive text processing and the lawyer supervises and corrects as needed. The authors state that when used this way, LLMs can significantly reduce the workload of legal professionals and speed up case handling, without fully replacing human expertise. At the same time, besides the problem of inaccuracies and hallucinations, they caution that there are also data privacy and ethical issues when applying LLMs to confidential legal documents or client data.

The use of LLMs in legal practice has also been investigated in the German business law context, reflecting the country's civil law system and language-specific challenges. Models must handle German-language statutes and court decisions, and align with a codified legal framework. Our own recent studies suggest that advanced LLMs are rising to this challenge [5], [6]. For example in [6], we reported that OpenAI's GPT-4 was the first AI system to successfully pass a German business law exam, demonstrating the model's ability to solve case-based legal questions at a level comparable to business law students. This milestone indicates that LLMs can generalize their legal reasoning ability beyond English common law and into German doctrinal analysis when appropriately prompted. The study [15] explored LLM performance on German legal text classification, specifically extracting norm analysis categories from statutes and regulations. They evaluated several multilingual models and a German-trained LLM (LeoLM) on this task, finding that with carefully engineered prompts (including task descriptions and examples), the models could accurately categorize provisions in German legislation.

Recent advances in legal informatics reach beyond LLM-centric work. Commercial platforms such as ROSS Intelligence's Legal Cortex, LawGeex, and eBrevia automate research, contract analytics and due-diligence tasks. Community challenges such as COLIEE, CAIL and ALQAC supply open corpora and benchmark tasks for retrieval, entailment, and question answering [16], [17], [18]. Other academic work spans transfer-learning summarization for Australian case law [19], contrastive-learning judgment prediction in China [20] and naïve-Bayes dialogue classification in the Canadian Supreme Court [21], illustrating a diverse toolbox that blends classic machine learning with new transformer-based LLM architectures.

Disruptive innovation as defined by Clayton Christensen in his seminal book "The Innovator's Dilemma" [22] refers to an innovation that makes products or services more accessible or affordable, initially targeting either low-end customers or nonconsumers (people not served by existing offerings). Such innovations often start with performance that is inferior on traditional metrics but offer other advantages (cost, convenience, simplicity) that appeal to new or less-demanding customers. Over time, the technology improves to meet mainstream needs, potentially displacing incumbents.

Large language models like ChatGPT have emerged as a candidate for disruptive innovation, and recent discourse frequently invokes Christensen's theory to analyze their impact. We see evidence across domains: in education and mental health, ChatGPT is reaching users who had no service before, a classic sign of new-market disruption, as pointed out in a series of articles on the Christensen Institute blog [23], [24]. In healthcare, the authors notes that if LLMs are packaged in low-cost models for underserved patients, they fulfill the disruptive playbook

(enabling tech + new model for nonconsumers). This pattern is playing out in the legal sector as well [25]. LLMs have proven capable of handling many legal tasks, and when leveraged within law firms they chiefly boost efficiency (a sustaining outcome). Yet their more promise lies in expanding legal services to those who traditionally could not afford them, "making [legal aid] available to a larger population" in Christensen's terms [24]. That scenario – AI legal advice for the masses – aligns closely with the disruptive innovation model, as a lower-cost, lower-performance solution for a market (so far) ignored by high-end providers. However, to become disruptive, a technology needs to also keep improving to steadily acquire more of the incumbents' market. The current study will contribute to understanding whether there is such an improvement consistent with Christensen's theory.

3. Research Approach

3.1. Objectives and Approach

This study investigates the capabilities of CAs based on generative AI to navigate and resolve legal scenarios within German business law. Our research is structured around the following questions:

- How would current leading CAs perform in German business law exams?
- What is the time trend of the CA performance in German business law exams?
- Are the findings consistent with Christensen's definition of disruptive innovation?

Our analysis centers on several leading CA families that have been accessible to end users: ChatGPT, Claude, DeepSeek, Google Gemini, Grok, Llama, and Mistral. We introduced 200 cases spanning contract, business, and employment law to evaluate these systems' ability to accurately interpret and respond to real-world legal challenges. These cases are proprietary, ensuring they were not part of the training datasets for the evaluated models.

The performance of CAs was assessed through multiple-choice tasks to remove room for subjective evaluation: there is only one correct answer, and the examiner must not evaluate whether a solution is still acceptable in terms of the subject matter. In Germany, multiple-choice tasks are rarely used in legal examinations. Instead, examinees are presented with a legal case which they must solve in the form of a legal opinion. In the U.S.A., however, multiple-choice tasks are common in legal examinations. In the Multistate Bar Exam (MBE), the U.S. bar exam, examinees are only given multiple-choice questions. A more detailed discussion of the advantages and disadvantages of using multiple-choice tasks can be found in [5].

3.2. Prompting and Assessment Strategy

We conducted our evaluation approach in a single-stage, multiple-choice query format. We executed 2,800 automated multiple-choice question (MCQ) evaluations, each structured to solicit only the CAs' selection from multiple potential answers to a given legal scenario. We required the CAs to respond with their chosen option in a straightforward format without generating a preceding legal opinion or engaging in elaborate reasoning processes, as such methods did not yield advantages in our setting.

The CAs were prompted to identify the correct response from four options, based on the given case details. The prompt provided to the CAs was as follows (translated from the original German language prompt):

> *"Solve the following case of a business law exam in the area of 'Formation and Validity of Contracts' by selecting the correct answer:"*
> *<Case Description>*
> *"Based on the presented case and legal opinion, select the correct answer:"*
> *<Response Options>*
> *"Indicate which option (a, b, c, or d) is the correct answer. Phrase it as 'The correct answer is Option x)'. Write no additional text, only mention the letter of the answer option, not its description."*

Each session was initiated anew to prevent data carryover and to minimize biases. Consistency was maintained by issuing a single, focused prompt, avoiding the possibility of influencing the CAs' answer through prior detailed

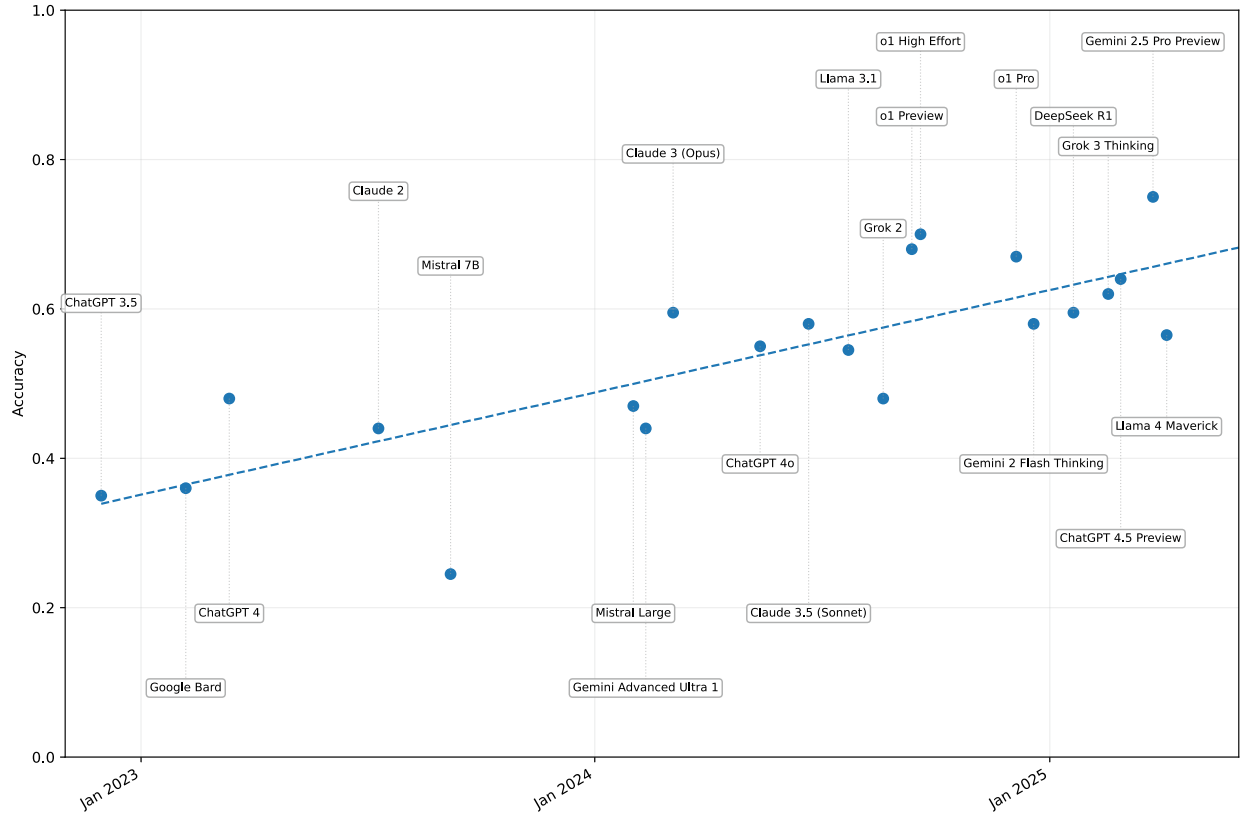
analyses, in contrast to the sequential approach used previously. The results of this method will contribute to understanding the direct decision-making capabilities of CAs in legal multiple-choice assessments and their ability to deduce the correct answers from complex legal scenarios without additional contextual prompting.

The multiple-choice responses were then assessed automatically using an automated script. The results of the script were then double checked by a human assessor. To facilitate comparability of outcomes with the MBE, the CAs were provided with four answer choices. For MCQs with four options, an average score of 25% would be achieved even by random guessing. To exceed this score would be the minimum performance for any useful application. Further details on the implementation of the prompting and assessment procedures can be found in [5].

4. Results

4.1. Overall Model Performance Over Time

Figure 1 depicts the accuracy of the CAs in this study. There is a clear upward trajectory in model performance across the study period, starting from around 25–45% accuracy in early 2023 and reaching levels above 70% for certain models by late 2024 / early 2025.



Notes: The dotted line represents the linear regression.

Fig. 1. Overall Model Performance Over Time.

The performance progression is reflected in the OLS regression reported in Table 1, where the overall annual improvement rate stands at 13.68% ($p < 0.0001$, $R^2 = 0.63$). The current data does not hint at a plateauing in the performance. It will be seen in the remaining 1 to 2 years, whether models performance will increase straight up towards 100% as predicted by the linear model, converge more slowly, or hit an earlier barrier.

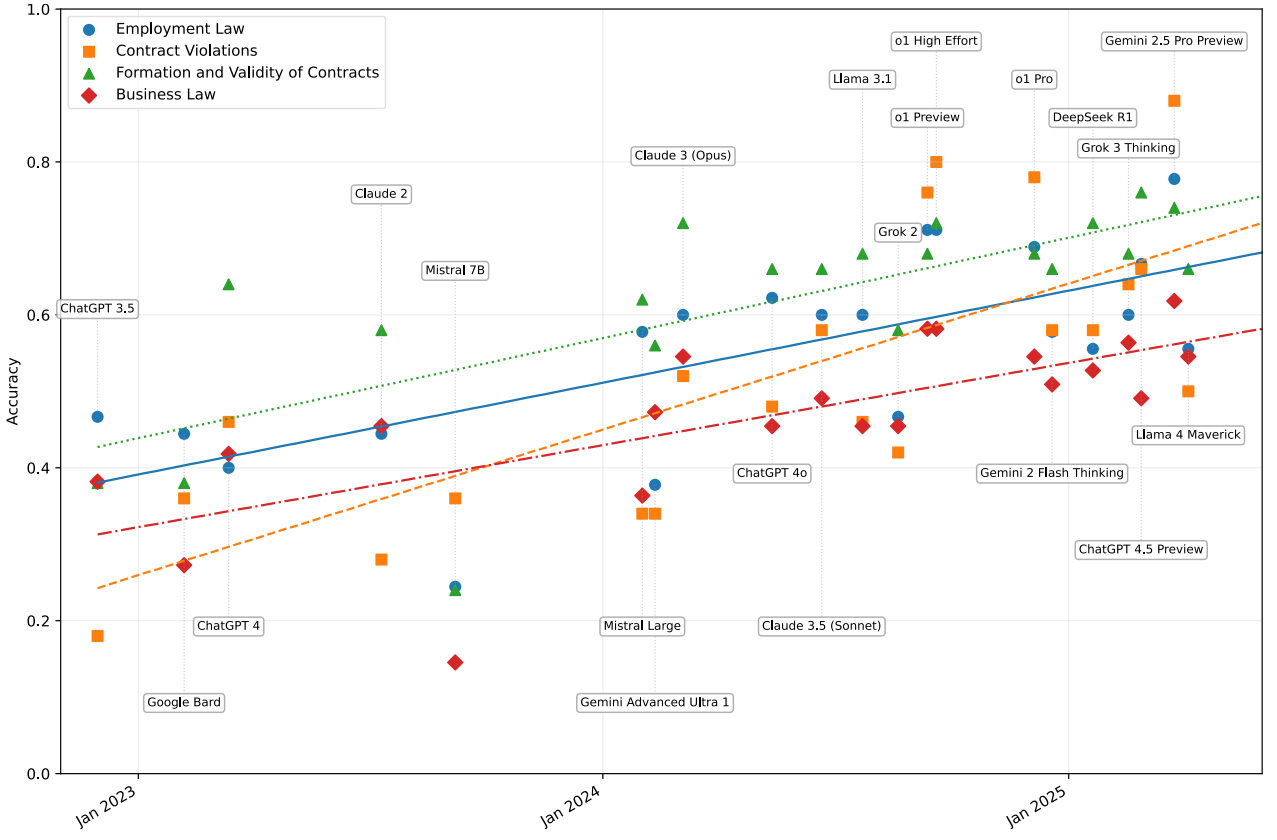
Table 1. Ordinary Least Squares (OLS) Regression of LLM Performance Over Time.

Legal Domain	Case Number	Annual Increase (%)	Std. Error (%)	p-value	R^2
Formation and Validity of Contracts	50	13.09	2.87	0.0002	0.52
Contract Violations	50	19.04	3.82	0.0001	0.57
Business Law	55	10.72	2.45	0.0003	0.50
Employment Law	45	12.01	2.97	0.0007	0.46
Overall	200	13.68	2.42	<0.0001	0.63

Notes: This table shows the annual rate of improvement in LLM performance on legal questions across different domains, based on linear regression analysis. The slope coefficient represents the estimated annual percentage point increase in accuracy. For each regression model, the number of observations is n=21.

4.2. Model Performance by Legal Domain Over Time

Figure 2 disaggregates model results into four legal domains—employment law, contract violations, formation and validity of contracts, and business law. Each subdomain similarly exhibits statistically significant improvement over time, as highlighted in Table 1. The patterns across domains generally mirror the overall trend, with a moderate to strong annual gain in each subfield.



Notes: The four lines represent the linear regression lines for the four legal domains.

Fig. 2. Model Performance by Legal Domain Over Time.

Notably, several current models—including o1 High Effort and Gemini 2.5 Pro Preview—consistently achieve results above 60–80%, whereas others, such as Llama 4 Maverick, still score below 50% for specific tasks. For employment law, models initially clustered between 35% and 60% accuracy in early 2023, but many by now exceed 70%. The regression estimates in Table 1 indicate a 12.01% improvement per year ($p = 0.0007$, $R^2 = 0.46$). For contract violations, accuracy exhibits both the widest spread and the steepest annual gain, at 19.04% ($p = 0.0001$, $R^2 = 0.57$). Several models still show mid-range performance, around 40–50%, but the best-performing systems now reach or exceed 80%. For formation and validity of contracts, we observe an 13.09% yearly improvement ($p = 0.0002$, $R^2 = 0.52$). Models tested early in the timeline show accuracy levels around 35–50%, whereas current releases often reach the mid-70% range. Finally, for business law, Table 1 indicates a 10.72% annual improvement rate ($p = 0.0003$, $R^2 = 0.50$). Accuracy varies from roughly 25% at the low end to over 60% at the upper end, reflecting differences in how each system processes and synthesizes the multifaceted business-law problems of this study.

These domain-specific trends underscore both the breadth and depth of the observed performance gains. Across all models and tasks, the clear year-over-year improvement signals that LLMs are evolving beyond mere novelty toward increasingly reliable tools for legal analysis. Certain subdomains, particularly contract violations, appear to benefit from more consistent patterns of factual reasoning, whereas areas like business law present a wider range of edge cases requiring nuanced understanding. Future releases of LLM-based CAs may further consolidate these gains if they incorporate additional technologies. Nevertheless, even at current levels, the leading systems consistently outperform earlier baselines from 2023, lending credence to the theory that LLMs are advancing toward professional-grade performance.

5. Discussion and Conclusion

5.1. Trajectory of LLM performance

The empirical results show significant improvements in the ability of large language models (LLMs) to tackle German business law tasks, with annual accuracy gains averaging around 13.68%. The data further reveal notable progress across all examined subdomains — employment law, contract violations, formation and validity of contracts, and general business law—though the magnitude of improvement and absolute performance levels vary by domain and model. These findings underscore both the promise of LLM-based tools for legal practice and the importance of continued scrutiny regarding their capabilities and limitations.

A first key implication concerns practical utility. The observed performance gains suggest that advanced LLMs are steadily moving from proof-of-concept experimentation toward deployable solutions. Tasks once requiring considerable effort—such as identifying breaches of contract, analyzing validity issues, or interpreting statutory provisions—are becoming more feasible for automated systems. Although many models still exhibit weaknesses (e. g., below 40% on some tasks), the leaders in this field are beginning to reach accuracy levels that can meaningfully assist legal professionals in research, document drafting, and preliminary risk assessments.

At the same time, reliability gaps remain. Despite top models reaching or exceeding 70% accuracy in specific categories, their performance is not uniform; some tasks continue to produce considerably lower success rates. Problems such as hallucinations (fabricated references or statutes) and domain-specific blind spots (e. g., specialized employment statutes or less common contractual doctrines) persist. Moreover, while an annual improvement rate of approximately 10 – 20% is striking, it should not be assumed to continue indefinitely. Future model evolutions may encounter diminishing returns or necessitate more specialized fine-tuning and more advanced reasoning steps to sustain current growth trajectories.

5.2. Criteria of Technological Disruption in Legal Tasks

Disruptive innovation theory, as conceptualized by Christensen and described in Section 2, frames technology adoption as an often non-linear process in which new entrants initially serve underserved or low-end segments

before advancing to displace incumbent solutions. The evidence from this study suggests that LLM-based CAs exhibit several characteristics of potential disruption within legal practice.

Many CAs can be accessed at relatively low cost and operate in natural language, offering potential for broader legal support—particularly to individuals or small businesses that previously lacked access to advanced legal services. Consistent performance increases, as observed here, support the argument that LLMs may soon meet or exceed the needs of mid-level and perhaps even higher-level use cases. By automating time-intensive tasks, LLMs free practitioners to focus on higher-value advisory roles—a pattern also seen in other industries disrupted by technology.

Real-world deployments of AI in the legal domain reported by [26] already reflect the early stages of a disruptive trajectory by focusing on simpler tasks. For instance, in Germany, the "OLGA" and "FRAUKE" systems expedite case processing by categorizing routine claims and assisting in drafting documents; in China, the "206 System" in Shanghai automates evidence analysis; and in Argentina, "Prometea" accelerates preliminary reviews of housing requests. Colombia has begun exploring AI through its Constitutional Court's "Pretoria," which rapidly prioritizes and summarizes thousands of cases. These applications illustrate how LLM-based CAs, as a first stage of disruptive innovation, are initially used for straightforward legal services, thereby lowering costs and broadening access. Over time, they could evolve to handle more complex tasks, converging with the broader implications discussed above—namely, reshaping the legal services landscape by automating labor-intensive activities and enabling practitioners to shift toward higher-value, advisory roles.

Nonetheless, it remains uncertain how far these technologies will penetrate mainstream legal service markets, particularly given regulatory, ethical, and liability considerations. Traditional law firms may integrate LLMs as a sustaining innovation to improve efficiency rather than a truly disruptive force that fundamentally reshapes market structures. On the other hand, new entrants offering LLM-centric legal solutions to individual consumers and small-scale enterprises could significantly broaden access to legal services, consistent with a disruptive innovation trajectory.

5.3. Considerations for CA's Disruptive Potential in Legal Practice

Extrapolating alone of the linear trend of reported annual improvements in accuracy and current overall performance levels—in some cases exceeding 80%—does not allow for reliable predictions as to whether and when AI might replace human lawyers today or in the near future. Identifying legal issues and generating proposed solutions in legal cases represents only a fraction of the challenges legal professionals face in everyday practice. Legal analysis typically requires identifying and assessing multiple interrelated issues in the correct order and in their mutual context. If a language model solves individual subproblems with an accuracy of around 70%, the overall probability of a correct solution decreases rapidly once several interconnected legal questions must be resolved. In a case involving three logically dependent issues, for example, a model would need to solve each subproblem with more than 96% accuracy to achieve an overall 90% likelihood of reaching a correct decision—a standard that current AI models clearly do not meet.

Furthermore, human lawyers continue to hold a decisive advantage wherever legal reasoning requires value judgments: It is the human who evaluates which interests should prevail in a given situation, how vague legal terms should be interpreted, and which line of argument is most convincing in light of statutory purpose and case law. AI is limited in making such assessments independently. While language models may simulate legal evaluation linguistically and (possibly) also logically, they are currently overwhelmed when expected to perform such evaluations autonomously. Their method of legal reasoning lacks a normative compass, as they are not conscient and do not bear the consequences for their output.

But if AI lacks a normative compass, how is it that these models still performed so well in the evaluations? The answer lies in the nature of the tasks presented: the models were asked to solve exam-style legal problems. The facts were fully defined, the issues clearly delineated, and a legally "correct" solution is available in principle. In such settings, language models can play to their strengths: recognizing patterns, reconstructing relevant norms and common reasoning structures, and selecting the most likely correct answer. The challenge was not to structure the case independently or resolve normative conflicts, but to choose a solution from among already familiar options. Here, AI models benefit from the standardized format of legal exams—formats that map closely to the structure of

their training data. Strong performance in such tasks, however, says little about how these systems would fare in real-world legal practice—where problems are not pre-formulated, information may be incomplete or contradictory, and legal uncertainty opens genuine decision-making room. In this sense, AI performs well in a "closed" format, but not (yet) in the open-ended reality of practice.

Nonetheless, lawyers might be well-advised to use AI models in practice. While they lack a normative compass, they already outperform human professionals in several areas: they can process large volumes of text, extract relevant information, and formulate standardized content quickly and fluently. AI is not a competitor to the lawyer—it is a tool: an intelligent assistant that handles routine work, accelerates research, and provides valuable intellectual impulses. It creates time for what only humans can deliver—and for what is often neglected without AI: careful evaluation, strategy, and personal client interaction. Used properly, AI becomes a productive sparring partner—a digital research assistant, a drafting helper, and a generator of ideas. The lawyer remains in charge: verifying, deciding, and weighing. The AI supports—quickly, systematically, and reliably—but it does not take the helm. It is not the captain—but an excellent first officer.

5.4. Conclusion

In conclusion, the growing capabilities of LLMs in German business law underscore the potential for meaningful, and possibly disruptive, impact on legal practice. Practical outcomes—such as faster contract review, cost savings in repetitive tasks, and improved access for small-scale clients—reinforce the argument that LLM-based CAs may shift standard operations in both law firms and legal aid contexts. However, the technology's broader adoption hinges on robust validation, careful risk mitigation, and clear regulatory guidance. Addressing challenges such as hallucinations, domain gaps, and confidentiality is crucial if these systems are to progress from promising tools to core components of future legal services.

References

[1] H. Hsu, K. Hsu, S. Hou, C. Wu, Y. Hsieh, and Y.-D. Cheng, "Examining Real-world Medication Consultations and Drug-Herb Interactions in Taiwan: An Evaluation of ChatGPT's Performance," *JMIR Med. Educ. 2023*, doi: 10.2196/48433.

[2] C. G. West, "AI and the FCI: Can ChatGPT Project an Understanding of Introductory Physics?," pp. 1–13, 2023, *arXiv.* doi:10.48550/arXiv.2303.01067.

[3] D. Xuan-Quy, L. Ngoc-Bich, P. Xuan-Dung, N. Bac-Bien, and V. The-Duy, "Evaluation of ChatGPT and Microsoft Bing AI Chat Performances on Physics Exams of Vietnamese National High School Graduation Examination," 2023, *arXiv*. doi: 10.48550/arXiv.2306.04538.

[4] M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches," *IEEE Access*, vol. 13, pp. 18253–18276, 2025, doi: 10.1109/ACCESS.2025.3533217.

[5] S. Schweitzer, M. Conrads, and J. Naeve, "Claude Rules: An Evaluation of Large Language Models' Applicability to Solve Cases in German Business Law," *Procedia Comput. Sci.*, vol. 246, pp. 2675–2683, 2024, doi: 10.1016/j.procs.2024.09.406.

[6] S. Schweitzer and M. Conrads, "The digital transformation of jurisprudence: an evaluation of ChatGPT-4's applicability to solve cases in business law," *Artificial Intelligence and Law*, Jul. 2024, doi: 10.1007/s10506-024-09406-w.

[7] B. Padiu, R. Iacob, T. Rebedea, and M. Dascalu, "To What Extent Have LLMs Reshaped the Legal Domain So Far? A Scoping Literature Review," *Information*, vol. 15, no. 11, p. 662, Oct. 2024, doi: 10.3390/info15110662.

[8] Z. Z. Chen *et al.*, "A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law," 2024, *arXiv*. doi: 10.48550/ARXIV.2405.01769.

[9] OpenAI *et al.*, "GPT-4 Technical Report," 2023, *arXiv*. doi: 10.48550/ARXIV.2303.08774.

[10] W.-L. Chiang *et al.*, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," 2024, *arXiv*. doi: 10.48550/ARXIV.2403.04132.

[11] Z. Fei *et al.*, “LawBench: Benchmarking Legal Knowledge of Large Language Models,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 7933–7962. doi: 10.18653/v1/2024.emnlp-main.452.
[12] C. Tang *et al.*, “PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models,” 2023, *arXiv*. doi: 10.48550/ARXIV.2309.10238.
[13] L. Martin, N. Whitehouse, S. Yiu, L. Catterson, and R. Perera, “Better Call GPT, Comparing Large Language Models Against Lawyers,” 2024, *arXiv*. doi: 10.48550/ARXIV.2401.16212.
[14] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, “Large Language Models in Law: A Survey,” Nov. 26, 2023, *arXiv*. doi: 10.48550/arXiv.2312.03718.
[15] S. T. Bachinger *et al.*, “GerPS-Compare: Comparing NER methods for legal norm analysis,” 2024, *arXiv*. doi: 10.48550/ARXIV.2412.02427.
[16] R. Goebel, Y. Kano, M. Y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka, "Summary of the competition on legal information, extraction/entailment COLIEE 2023," in Proceedings of the nineteenth international conference on artificial intelligence and law, pp. 472–480, 2023, doi: 10.1145/3594536.3595176.
[17] C. Xiao, et al., "A large-scale legal dataset for judgment prediction", 2018, *arXiv*. doi: 10.48550/arXiv.1807.02478.
[18] T. M. Nguyen, et al., "NOWJ1@ ALQAC 2023: Enhancing legal task performance with classic statistical models and pre-trained language models," 2023, *arXiv*. doi: 10.48550/arXiv.2309.09070.
[19] G. Moro, N. Piscaglia, L. Ragazzi, and P. Italiani, "Multi-language transfer learning for low-resource legal case summarization," *Artificial Intelligence and Law*, vol. 32, pp. 1111–1139, 2024, doi: 10.1007/s10506-023-09373-8.
[20] H. Zhang, Z. Dou, Y. Zhu, and J. R. Wen, "Contrastive learning for legal judgment prediction," *ACM Transactions on Information Systems*, vol. 41(4), 2023, doi: 10.1145/3580489.
[21] J. Tomin and V. Adiyillam, "Supreme court dialogue classification using machine learning models," *International Journal of Electrical and Computer Engineering*, vol. 13(2), pp. 2350–2355, 2023, doi: 10.11591/ijece.v13i2.pp2350-2355.
[22] C. M. Christensen, *The Innovator’s Dilemma*. Cambridge, Mass: Harvard Business School Press, 1997.
[23] T. Arnett, “Will chatbot therapy, VR schooling, and AI-aided writing instruction become the new normal in K–12 education?,” Christensen Institute Blog. Accessed: Mar. 21, 2025. [Online]. Available: https://www.christenseninstitute.org/blog/will-chatbot-therapy-vr-schooling-and-ai-aided-writing-instruction-become-the-new-normal-in-k-12-education/
[24] J. Plante, “ChatGPT and health care: Disruptive? Or just the same old thing?,” Christensen Institute Blog. Accessed: Mar. 21, 2025. [Online]. Available: https://www.christenseninstitute.org/blog/chatgpt-and-health-care-disruptive-or-just-the-same-old-thing/
[25] P. Weichbroth, “AI and the Law: Evaluating ChatGPT’s Performance in Legal Classification,” 2025, *arXiv*. doi: 10.48550/ARXIV.2502.12193.
[26] L. Carrasquilla-Díaz, A. De Luque-Pisciotti, and E. Lagos-González, “AI Adoption in Colombian Legal Practice: Challenges and Opportunities.,” *Procedia Comput. Sci.*, vol. 241, pp. 508–513, 2024, doi: 10.1016/j.procs.2024.08.072.