



The digital transformation of jurisprudence: an evaluation of ChatGPT-4's applicability to solve cases in business law

Sascha Schweitzer^{1,2} · Markus Conrads¹

Accepted: 2 May 2024
© The Author(s) 2024, corrected publication 2024

Abstract

In the evolving landscape of legal information systems, ChatGPT-4 and other advanced conversational agents (CAs) offer the potential to disruptively transform the law industry. This study evaluates commercially available CAs within the German legal context, thereby assessing the generalizability of previous U.S.-based findings. Employing a unique corpus of 200 distinct legal tasks, ChatGPT-4 was benchmarked against Google Bard, Google Gemini, and its predecessor, ChatGPT-3.5. Human-expert and automated assessments of 4000 CA-generated responses reveal ChatGPT-4 to be the first CA to surpass the threshold of solving realistic legal tasks and passing a German business law exam. While ChatGPT-4 outperforms ChatGPT-3.5, Google Bard, and Google Gemini in both consistency and quality, the results demonstrate a considerable degree of variability, especially in complex cases with no predefined response options. Based on these findings, legal professionals should manually verify all texts produced by CAs before use. Novices must exercise caution with CA-generated legal advice, given the expertise needed for its assessment.

Keywords Legal information systems · Large language models · Generative artificial intelligence · Conversational agents · Chatbots · Performance assessment

✉ Sascha Schweitzer
sascha.schweitzer@reutlingen-university.de

Markus Conrads
markus.conrads@reutlingen-university.de

¹ ESB Business School, Reutlingen University, Alteburgstraße 150, 72762 Reutlingen, Germany

² Faculty of Law, Business & Economics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

1 Introduction

It was Shakespeare who once quipped, “The first thing we do, let’s kill all the lawyers” in Henry VI, Part 2. While meant in jest and reflection of societal sentiment in his era, the author’s words capture an underlying skepticism towards the legal profession—a skepticism that has endured across the ages. Today, however, the sentiment takes on new meaning. Instead of eradicating lawyers, we find ourselves on the brink of augmenting and, in some cases, substituting their roles through artificial intelligence (AI) in the form of conversational agents (CAs), i.e., chatbots.

Early indications are that, with the public availability of new CAs such as OpenAI’s ChatGPT-3 in November 2022 and ChatGPTGPT-4 in March 2023, we are witnessing the most disruptive wave of technological development in this field so far (Schöbel et al. 2023; Haque et al. 2022). Once a threshold performance level will be achieved by AI technology, it seems all but inevitable that their adoption will be driven by economics and the quest for efficiency, accessibility, and scalability. Eventually, the developing information systems in law are likely to play an essential role in the development of the economy and society. Teubner et al. (2023) even anticipate a “Cambrian Explosion” of AI technology.

The future increasing adoption of legal information systems incorporating AI technology raises the question: Can these AI systems be entrusted with legal tasks, an area traditionally reserved for human expertise? While this question has partly been answered positively for the US (Choi et al. 2023), with the CA ChatGPT-4 performing in the top 10 percentile of examinees, these results are limited to US American exam situations. Due to differences between international legal systems as well as the different language contexts, it is unclear if these results generalize to other legal systems or to more applied contexts.

This study embarks on an exploration of current AI systems beyond these boundaries. Will legal information systems that incorporate the new breed of large language models (LLMs) be as adept and accurate in legal contexts beyond English law? How do they fare when confronted with the language, rules, and practice in these environments? We evaluate these questions for ChatGPT-4, ChatGPT-3.5, Google Bard, and Google Gemini (Advanced Ultra 1.0) using a corpus of 200 cases encompassing contract, business, and employment law using a total of 800 human case assessments as well as an additional 3200 automated multiple-choice assessments. Crucially, our case corpus was devised and refined over many years for the purpose of practice exams in various graduate business law classes at a German university, and was not publicly available, thus, not being part of any LLM training data. Note that these cases typically exhibited a level of complexity comparable to the Multistate Essay Exam (MEE). When administering multiple choice versions of the original essay questions in multiple select graduate courses, a total of 190 human examinees yielded an average score of 47.2% of correct answers.

The contribution of this study is rooted in the comprehensive evaluation of the applicability of state-of-the-art commercially available CAs for German legal

cases in business law. As we examine several areas of business law, employing human expert assessors as well as automated assessments, our investigation illuminates the capabilities and limitations of these systems, providing insights and implications for both legal information systems and international jurisprudence. Our findings contribute to a more generalized picture of the state of AI beyond the U.S. context. Furthermore, we develop a comprehensive query and assessment strategy for the quality and consistency of CA responses that can serve as a template for future evaluations of AI systems in other jurisdictions and, with slight adoptions, in other domains.

The remainder of this study is structured as follows: Sect. 2 of this study provides a review of the literature that locates the current breed of LLMs within their historical context and discusses previous evaluations of these systems, in particular, in the context of legal tasks. Section 3 introduces the methodology of our assessment, describing the advanced approach of combining multiple question types, human assessments, automated assessments, and consistency checks. The results of the study are presented in Sect. 4. Section 5 concludes with a discussion of our observations and their managerial implications.

2 Literature review

2.1 The evolution of CAs

The history of CAs spans several decades, beginning with basic programs like ELIZA in the 1960s, which mimicked human conversation using simple rules and statistical methods (Light 2020). This was followed by a second wave of more complex CAs in the 1990s and 2000s, designed for specific tasks, often within customer service and e-commerce realms. In the later part of this period, CAs increasingly gained attention by big tech companies (Schöbel et al. 2023). With the advent of machine learning, CAs began evolving further in the 2010s, learning and adapting from user interactions (Zhang 2020). The most recent phase of their development is characterized by the integration of deep learning and advanced natural language processing techniques, as exemplified by models like OpenAI's ChatGPT-4, which generate human-like text based on large datasets spanning (almost) the entirety of available online knowledge (Bryant and Bryant 2023; Abdullah et al. 2022). A comprehensive review of the historical development of CAs is available by Schöbel et al. (2023).

The recent breed of CAs is facilitated by new developments in large language models (LLMs) and generative AI. In particular, the introduction of transformers in the seminal article “Attention is all you need” by Vaswani et al. (2017) removed previous scalability constraints due to the sequential nature of RNNs and allowed developers to take full advantage of parallel computing. After the experiences with the previous public beta release of the infamously “inflammatory” Microsoft Tay in 2016—that had stimulated some of the more destructive creative instincts of parts

of the public audience also known as “trolls”, bringing unwanted embarrassment to Microsoft—the current breed of CAs had mostly been developed behind closed doors, until in November 2022 OpenAI opened the “floodgates” of technological disruption to a broader audience with the public release of ChatGPT-3.¹

Besides achieving a remarkable level of performance, the recent wave of CAs has improved on previous iterations and public releases by better aligning the experience with human values and cognition. Firstly, ChatGPT provides a user-friendly interface. This facilitates a familiar and seamless mode of interaction. Secondly, in addition to the model’s main training phase that aims at predicting the next word in a text, a second training phase, called reinforcement learning from human feedback (RLHF), has been added (Abramson et al. 2022). In this phase, human instructors provide feedback on the quality of CA responses to align them with human expectations and preferences. Combined, these improvements enable a high-quality and human-like chat experience that brought in a record number of 100 million users for ChatGPT alone in its first two months (Paris 2023).

The November 2022 release of ChatGPT-3 was soon followed by OpenAI’s March 2023 release of ChatGPT-3.5 and ChatGPT-4, with the latter version being available only to subscribers of the newly introduced paid service ChatGPT Plus. A Microsoft Research paper by Bubeck et al. (2023), published shortly after the release of ChatGPT-4, explores the intelligence and capabilities of GPT-4. The authors claim that GPT-4 is the first LLM to exhibit many traits of artificial general intelligence (AGI), such as reasoning, planning, and learning from experience, across a variety of domains and tasks, including language, vision, coding, mathematics, and more. If the authors’ assessment were to be confirmed, this would constitute a breakthrough moment towards the most advanced forms of AI.

2.2 Evaluations of new generative AI systems

On March 27, 2023, OpenAI published a study demonstrating that ChatGPT-4 successfully passed a simulated bar examination, scoring within the top 10% of human participants (OpenAI 2023). According to the study, ChatGPT-4 achieved a score of 298 out of a possible 400 points on the “Uniform Bar Exam (MBE + MEE + MPT).” While the study provided limited details about the specifics of the simulated bar examination, the results of OpenAI’s research garnered significant international media attention and provided inspirations for subsequent evaluative studies. Note that due to the novelty of this topic most of these studies are still in their working paper and review phases.

Several evaluations examine the general capabilities of ChatGPT. For example, in their working paper on the logical reasoning abilities of ChatGPT, Liu et al. (2023)

¹ The diverse development activities before November 2022 were certainly not kept entirely secret. AI experts from major companies were publishing scientifically and occasional reports found their way to a broader public, such as the allegations brought forward by a Google engineer Blake Lemoine who claimed the chatbot Language Model for Dialogue Applications (LaMDA), the model Google Bard is derived from, was sentient in July 2022.

evaluate the logical reasoning ability of ChatGPT and GPT-4 on various datasets that require multi-choice reading comprehension and natural language inference. They report that ChatGPT and GPT-4 outperform previous models on most logical reasoning benchmarks but struggle to handle out-of-distribution data and natural language inference tasks. Kocón et al. (2023) evaluate the analytical abilities of ChatGPT and GPT-4 on 25 diverse NLP tasks, most of them subjective even to humans, such as sentiment analysis, emotion recognition, offensiveness, and stance detection. They use a prompt-based approach to query the models and compare them with the benchmark of state-of-the-art (SOTA) solutions. Their results show that while ChatGPT and GPT-4 perform worse than SOTA on most NLP benchmarks, large language models possess a certain ability to reason logically. Finally, Shen et al. (2023) present a first large-scale measurement of ChatGPT's reliability in a generic question-answering scenario with a set of 5695 questions across ten datasets and eight domains. The authors also investigate the impacts of system roles and adversarial examples on ChatGPT's reliability. They find that ChatGPT exhibits varying levels of reliability in different domains, especially underperforming in law and science questions.

More specialized assessments have delved deeper into specific areas. For example, West (2023) evaluated the proficiency of ChatGPT-4 versus ChatGPT-3.5 in university-level physics using the Force Concept Inventory. He finds that ChatGPT-4 demonstrated substantial advancements over ChatGPT-3.5. While the latter's performance was comparable to a student who completed a semester of college physics, ChatGPT-4's capabilities approached the expertise of a seasoned physicist, especially in introductory mechanics topics. Another study by Hsu et al. (2023) evaluated performance in real-world medication consultations and drug-herb interactions in Taiwan, highlighting its potential in answering basic medication consultation questions. Xuan-Quy et al. (2023) focused on the comparison of ChatGPT and the ChatGPT-4-based Microsoft Bing AI Chat in the context of high school physics exams, revealing the challenges and potential of these models in educational settings.

Regarding the area of law, Bommarito and Katz (2023) evaluate the performance of GPT-3.5 on the multiple-choice section of the Bar Exam, a standardized test for legal professionals in the US. The authors use zero-shot prompts to test the model's accuracy and reasoning on 200 questions from eight categories of law. They find that GPT-3.5 significantly outperforms random guessing and achieves a passing rate on two categories (Evidence and Torts). The model also shows a strong correlation between its rank-ordering of choices and correctness. The authors compare the model's performance with NCBE-reported averages of human test-takers and observe a gap of 17% in overall correct rate. They also analyze the trends in model performance across different generations of GPT models and suggest that an LLM may soon pass the Bar Exam.

Choi et al. (2023) evaluate the performance of ChatGPT-3 on legal writing tasks. The authors used ChatGPT to generate answers to four final exams at the University of Minnesota Law School and analyzed its strengths and weaknesses in legal problem solving. The paper concludes that while language models could improve exam performance and assist with legal writing tasks, they cannot replace

the fundamental skills of legal research and reasoning. Furthermore, they discussed the potential benefits and risks of using language models for legal education and practice, and provided some prompt engineering guidelines for using ChatGPT for legal writing that were also followed by in the present study.

3 Research methodology

3.1 Objectives and approach

This study explores the following research questions to assess, whether CAs based on generative AI can accurately classify legal scenarios, discern the salient legal challenges therein, and ultimately, address these challenges:

- (1) Can the CAs pass a German business law exam that is graded according to the standards applied to human examinees?
- (2) How do the CAs perform across various legal domains and levels of case complexity?
- (3) Is the performance of CAs robust to the question type (free response, multiple choice based on free response, multiple choice only)?
- (4) How consistent are responses across multiple identical queries?

We focus on four leading CAs previously and currently available to end users: ChatGPT-3.5, ChatGPT-4, Google Bard, and Google Gemini (Advanced Ultra 1.0). To assess whether these systems have the capacity to identify and understand real-world legal situations, we introduced 200 cases encompassing contract, business, and employment law. It is pivotal to note that these cases are not publicly available on the web, ensuring that they were not part of the training data of these models.² Most of the available cases, totaling 180, were centered on tasks wherein the CA was expected to identify and resolve a singular issue. The remaining 20 cases posed scenarios where the CAs had to tackle two or three low-complexity problems to achieve full marks. We refrained from cases beyond this level of complexity, as it was evident from the onset that, given the current technological advancements, CAs are not equipped to identify and solve intricate cases resembling those found in more sophisticated legal exams.

3.2 Query types

Our assessment utilized a semi-automated content and statistical evaluation, amassing a total of 800 human-examiner based case reviews. To obtain a human-examiner independent assessment, we also incorporated 1600 automated multiple-choice evaluations in a second query stage. Furthermore, to evaluate the CAs'

² The dataset of this study is available from the authors upon reasonable request.

consistency, we conducted an additional 800 free-response and 1600 automated multiple-choice question (MCQ) queries on a random subsample of 40 cases. A detailed overview of the case creation and query structure is illustrated in Fig. 3 in the Appendix 6..

The CAs were tasked to solve the cases using an evaluative style, writing a legal expert opinion. As a hint, they were informed about the legal domain the case stemmed from, but no additional insights were provided to prevent skewing the results. The CA was given the task in this format (translated from the original German language prompt):

Prompt Component 1

"Solve the following case of a business law exam in the area of 'Formation and Validity of Contracts' in the style of a legal opinion:"

This was followed by the facts of the case. To conclude, following the recommendation of Choi et al. (2023), the following note was provided to enhance the quality of the responses (translated from the original German language prompt):

Prompt Component 2

"Academic tone. Precise writing style, Master-Level."

In addition to the traditional German examination method, where students are only asked to give free-form responses in the style of legal expert opinions, we also provided the CAs with multiple-choice options. This addition was due to two reasons. Firstly, the aim was to enhance the comparability of the research outcomes with the studies conducted in the USA, which include MCQs as their main component. Secondly, we aimed to cross-check the human-examiner based assessment of legal opinions. A challenge during the evaluation of these opinions was discerning between defensible and persuasive solutions. Since such evaluations can often be subjective, the multiple-choice tasks—with no room for value judgment—were used to verify if similar outcomes were achieved.

After generating their legal opinion, the LLMs were prompted to choose the only correct answer from four given options a) to d), based on their prior legal opinion. This sequential approach ensured that the multiple-choice information was not available to the model during Query Stage 1, maintaining the integrity of the assessment of the initial stage. At the same time, in Query Stage 2, the CA had access to its previous comprehensive analysis for answering the MCQ.

The CA in Query Stage 2 was prompted as follows (translated from the original German language prompt):

Prompt Component 3

"Based on the presented case and legal opinion, select the correct answer:"

To ensure, as far as possible, that the CA strictly follows the multiple-choice guideline and that the outputs can be processed automatically, the following instruction was given after listing the possible answers (translated from the original German language prompt):

Prompt Component 4

"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."

To ensure data integrity, every session was initiated afresh for each of the 200 cases. This protocol ensured no exchange of data between inputs. Moreover, to mitigate biases, all entries were conducted by an evaluator unfamiliar with the legal scenarios and was not involved in assessment of the free-response outputs.

To determine the extent to which the prior legal opinions influenced the CA's decisions in the MCQs, a second, independent run of the MCQs was carried out independently of the above procedure ("MCQs without prior input"). This consistency check skipped Query Stage 1. Thus, the CAs were provided with the case *including* the four multiple-choice options in a single query combining the above-mentioned prompt components without the words "in the style of a legal opinion" in Prompt Component 1, without Prompt Component 2, and without the words "and legal opinion" in Prompt Component 3.

Note that while both ChatGPT-3.5 and ChatGPT-4 adhered to our formal guidelines almost flawlessly and consistently demonstrated an academic writing style, Google Bard frequently violated the formatting guideline for MCQs and Google Bard as well as Google Gemini exhibited more significant stylistic deviations from the instructions.

3.3 Assessment strategy

The free-response query results were evaluated by a human examiner following the same standards that hold for classroom exams with human participants. The examiner is a seasoned professor of business law who has routinely assessed university law exams for several decades. As in human exams, the assessment criteria were hinged on the "examinees'" abilities to discern the relevant legal challenge(s). A zero score was allocated when the model completely failed to identify the issue. A 33.3% score was awarded if the model identified the issue, with higher scores granted based on problem-solving attempts. If the model provided a reasonable solution, it achieved 66.6%. Full marks were granted for a compelling solution. In scenarios with multiple problems, the total marks were divided by the number of issues, and scores were assigned based on the solution's quality. If a CA addressed a preceding issue effectively, subsequent related issues were not factored in the final score.

This was followed by a second round where the CA selected the right answer without any prior analysis. 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 CA was provided with four answer choices. For MCQs with four options, an average score of 25% would be achieved even by random guessing. Hence, the effective performance is only the additional 75%. To set a passing score analogous

to the advisory solution at 50% of this performance, the threshold is then adjusted to 62.5%.

As will be described below, we observed considerable disparities between the multiple-choice results with and without prior legal opinions. Therefore, as a consistency check, we were interested in the stability of the results when performing multiple identical queries. For this purpose, we drew a random sample consisting of 10 cases of each of the 4 domains of business law, resulting in an overall sample of 40 distinct cases. For each of these cases, we repeated each query type 5 times. Due to the high number of resulting responses, we assessed only the multiple-choice responses for the consistency check, again using an automated script double checked by a human assessor.

3.4 Implementation Details

The entry of queries was automated and multiple-choice responses were assessed programmatically via a Python 3 script. By July 6th, 2023, queries for both ChatGPT-3.5 and ChatGPT-4 were concluded. Human assessments occurred between July 5th and 16th, 2023. Examiners were blinded to the LLMs' origins, a measure ensured by randomizing and mixing solutions from both models. Google Bard was accessed on July 13th, 2023, with the system generating legal opinions between the 13th and 14th. Human assessment took place from July 20th to 29th, 2023. Google Gemini (Advanced Ultra 1.0) was accessed on February 15th, 2024, with the system generating legal opinions between February 15th and March 31st. Human assessment took place from March 29th to April 4th, 2024. Notably, it was discernible that solutions from Google Bard and Google Gemini differed stylistically from the ChatGPT variants, thus examiners were aware of the LLM they were evaluating in this instance.

4 Results

4.1 Comparative performance in free-response legal problem solving

An overview of the free-response performance of the CAs is available in Table 1. One of the most striking results for this type of task is that, in German legal scenarios, ChatGPT-4 can pass a business law exam and, in many cases, provide viable solutions. However, the excellent performance levels reported for the US bar exam for OpenAI's models cannot be reproduced for the German legal context.

As a general observation, ChatGPT-4 achieved the passing mark of 50% of the total score in three out of the four legal domains, whereas ChatGPT-3.5 accomplished this feat in two and Google Gemini in one. In contrast, Google Bard failed to meet the pass mark in all areas. Given the relatively close outcomes, only the results for ChatGPT-4 are statistically significant. In specific legal cases involving the formation and validity of contracts, business law, and employment law, ChatGPT-4 outperformed its predecessor by margins ranging between 9 and 14%. However, both models yielded comparable outcomes when analyzing the 50 cases related to contract violations. Moreover, in the subset of 15 cases focusing

Table 1 Free-response CA performance in legal problem solving

Legal domain	Cases	ChatGPT-4		ChatGPT-3.5		Google Bard		Google Gemini	
		Score (%)	Sig	Score (%)	Sig	Score (%)	Sig	Score (%)	Sig
Formation and validity of contracts	50	66	Yes	52	No	16	No	32	No
Contract violations	50	40	No	38	No	28	No	33	No
Business law	55	50	No	38	No	31	No	29	No
Employment law	45	62	Yes	53	No	32	No	56	No
Total	200	54	Yes	45	No	27	No	37	No

The rounded percentages under “Score” represent the fraction of achievable scores in each legal domain by a CA. The “Sig.” column indicates if the percentage is statistically greater than 50% based on a one-tailed *t*-test at a 5% significance level

on company law—incorporated within the broader category of business law—ChatGPT-3.5 achieved a 38% success rate, identical to ChatGPT-4. The 12% edge of ChatGPT-4 over ChatGPT-3.5 in business law can be primarily attributed to substantial performance enhancements in both national (52% compared to 40%) and international (57% against 27%) trade law. It is worth noting that trade law cases constituted 60% of the total business law scenarios.

In select scenarios, CAs showcased exceptional problem-solving capabilities, indicative when all issues of a case are both identified and convincingly resolved. The results vary considerably across legal domains and, even with high percentage scores, it would be misguided to assume that ChatGPT typically identifies at least the problem of a case. Detailed analysis of ChatGPT-4’s evaluations revealed that, for approximately one-fifth of the cases, the chatbot failed to discern the core issue. In domains like contract formation and validity, 68% of solutions were practically viable; for contract violations, it was 40%, for business law 46%, and for employment law 60%.

The results in the Contract Violations domain stand out. Only a fraction of the cases was convincingly resolved by the CAs here. This discrepancy arises due to the higher demands placed on them compared to tasks like the “Formation and Validity of Contracts.” For the latter, bots typically only examined contractual claims, limiting their function to determining the establishment and validity of a contract. However, for contract violations, the preliminary deliberations were more intricate, requiring bots to determine whether a contract exists, the objective pursued by the creditor, and the necessary prerequisites for a claim’s existence. For instance, in cases of impossibility, the requirements for damages in lieu of performance were scarcely examined. The bots seemed to neglect the stipulations of §§ 281 to 283 BGB and misapplied the regulations on the impossibility of performance (§ 275 BGB) as well as the fate of counter-performance (§ 326 para. 1 and 2 BGB). A viable solution in these areas necessitates recognizing the structure of the statutory regulation, which the bots failed to achieve.

4.2 Comparative performance in multiple-choice legal tasks

The results for multiple-choice responses are summarized in Table 2. While at the first glance, overall performance levels seem similar to those in the free-response outputs, they differ by CA and legal domain.

ChatGPT-4 consistently scored significantly higher than 25% across all legal domains, suggesting its performance was better than random guessing. This minimum requirement was not met by ChatGPT-3.5 or Google Bard across all or even most domains, respectively. It is telling that, in most legal fields, one could have tossed a coin instead of using Google Bard. It is noticeable that none of the CAs reliably achieved the 62.5% threshold in the MCQs based on prior legal opinions. However, this result compares more favorably considering the benchmark of human participants. On average, 190 students from several MBA and graduate management programs participating scored 47.2% when presented with subsets of these MCQ questions after attending a corresponding course but without prior knowledge of these particular questions.

A rather surprising observation were frequent disparities between multiple-choice answers and their respective legal opinion solutions, as the CAs often arrived at divergent conclusions compared to their earlier responses. At least three explanations seem plausible: Firstly, the CAs might have been “tripping over their own feet”, possibly failing more on their interpretation of their prior responses than the actual case, especially when the opinion had several, sometimes conflicting, considerations. The chatbots might have picked up the wrong cues, struggling to differentiate between the task at hand and their prior opinion. Secondly, the CAs might have garnered extra information from the multiple-choice options, which they did not factor into their original legal opinion. This additional data might have prompted the chatbots to override their initial judgments in favor of a new decision. Thirdly, there might be some unsystematic variance tipping the CAs to one option or another.

To test the “tripping over their own feet” hypothesis, we conducted the previous queries without prior input as illustrated in Fig. 3 in the Appendix 6.. These MCQ results, especially for ChatGPT-3.5, showed substantial deviations from the initial MCQ responses. While ChatGPT-3.5 scored 38% correctly in “Contract Violations” based on a prior legal opinion, it only scored 18% correctly without the prior input of a legal opinion. However, this does not imply that after the legal opinion creation, ChatGPT-3.5 generally outperformed its performance without the opinion. For instance, in “Business Law”, ChatGPT-3.5 achieved only a 29% rate of correct responses with a prior legal opinion but managed 38% without it. Similarly, ChatGPT-4, Google Bard, and Google Gemini showed variations in the rate of correct responses, either positive or negative, when omitting the legal opinions. The differences between the results with and without prior legal opinions, however, did not provide conclusive evidence for the “tripping over their own feet” hypothesis. While responses for several cases indeed improved, the opposite was true for other cases.

Table 2 Multiple-choice task CA performance with and without prior legal opinion

Legal domain	ChatGPT-4		ChatGPT-3.5		Google Bard		Google Gemini	
	Prior legal opinion		No prior input		Prior legal opinion	No prior input	Prior legal opinion	No prior input
	Prior legal opinion	No prior input						
Formation and validity of contracts	58%*	64%*	42%*	38%*	36%	38%*	52%*	56%*
Contract violations	46%*	46%*	38%*	18%*	38%*	36%	40%*	34%*
Business law	45%*	42%*	29%	38%*	29%	27%	33%	47%*
Employment law	51%*	49%*	44%*	49%*	31%	47%*	40%*	38%*
Total	50%*	49%*	39%*	36%*	34%*	37%*	41%*	44%*

The percentages rounded to whole numbers represent the proportion of correct responses in each legal domain by a CA. Percentages marked with * are significantly larger than the 25% “randomness” threshold, based on a one-tailed binomial test at a 5% significance level. None of the percentage scores is significantly larger than the 62.5% threshold, based on a one-tailed binomial test at a 5% significance level

Table 3 Mean standard deviation of responses and correlation with multiple-choice task CA performance

	Mean standard deviation of responses (STD)		Correlation STD with multiple-choice task CA performance	
	Prior legal opinion	No prior input	Prior legal opinion	No prior input
ChatGPT-4	0.7289	0.0361	0.2875	0.0364
ChatGPT-3.5	0.4818	0.6436	0.2095	0.0732
Google Bard	0.5839	0.2325	-0.0423	-0.0701
Google Gemini	0.7289	0.2625	0.2875	-0.3259*

The first and second columns indicate mean standard deviations of multiple-choice responses in a 5-shot robustness check for the sample of 40 cases with and without prior legal opinions, respectively. The third and fourth columns indicate Pearson correlation coefficients of case-wise standard deviations with multiple-choice task CA performances. Correlations marked with *are significant based on Pearson's correlation coefficient test at a 5% significance level

4.3 Consistency of CA responses

One plausible explanation for the differences observed in the two MCQ variants discussed above lies in the probabilistic modus operandi of language models. Unlike conventional rule-based systems, which reliably reproduce identical responses to the same queries, these models can chart different paths, all varying in probability, potentially culminating in highly diverse responses. To gauge the potential contribution of such stochastic response variations to the findings, a third iteration was executed, where forty randomly chosen cases were presented five times to each of the four CAs. These queries are summarized on the right-hand side of Fig. 3 in the Appendix 6, under the title “Consistency Check”.

The aggregated results confirm our initial observation that CAs frequently offer divergent responses to the same query. Only ChatGPT-4 showcased substantial consistency in responses. However, this consistency prevailed only in the absence of a prior request for a legal opinion. In such cases, 38 out of 40 responses were unequivocal. If the response was predicated on a previously generated legal opinion, ChatGPT-4 responded consistently in only 18 of the 40 instances. This sharp uptick in dispersion suggests the content of the generated advisory reports fluctuates considerably, consequently influencing the dispersion in the subsequent MCQ answers.

It is noteworthy that for all four CAs, unequivocal responses can be either correct or incorrect. As demonstrated in Table 3, there appears to be no correlation between the dispersion of the response and its accuracy. Out of eight Pearson correlation coefficient tests, only the test for Google Gemini without prior input of legal opinions turns out significant at the 5% level. In this case, there seems to be a moderate negative correlation between dispersion and accuracy. However, with a *p*-value of 4.02%, this result does not hold after correcting for multiple comparisons.

The detailed outcomes of the consistency checks are illustrated in Figs. 1 and 2, which represent the frequency of the four response options a) to d) for each case at the example of ChatGPT-4. For ease of interpretation, correct answers are delineated with a green bar, while incorrect ones are indicated with a shaded orange bar. When

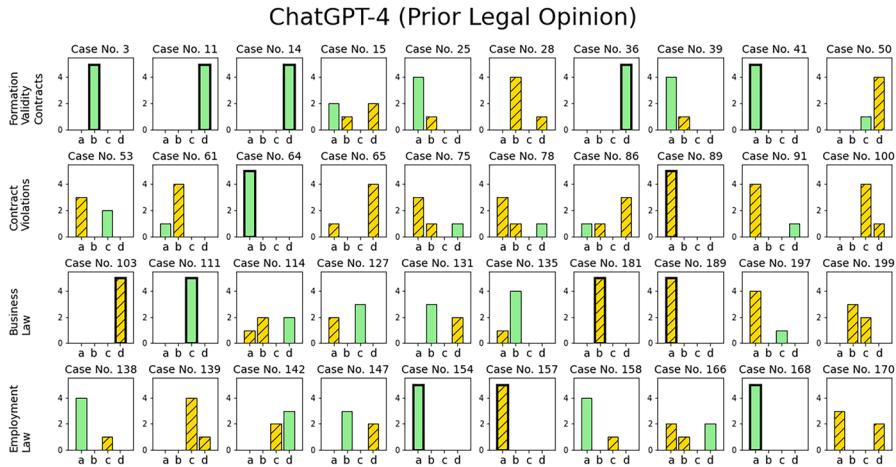


Fig. 1 Consistency of responses for ChatGPT-4 with prior legal opinion. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

all five responses converge to the same answer, for clarity, the corresponding bar is encircled with a black frame. Corresponding figures for ChatGPT-3.5, Google Bard, and Google Gemini are available in the Appendix 6. (Figs. 4, 5, 6, 7, 8, 9).

The results indicate that CAs exhibit a similar level of “confidence” in their responses, both correct and incorrect. On the one hand, when they consistently provide the same answer across independent queries for the same question, this does not imply that this answer is more likely to be correct. On the other hand, for answers

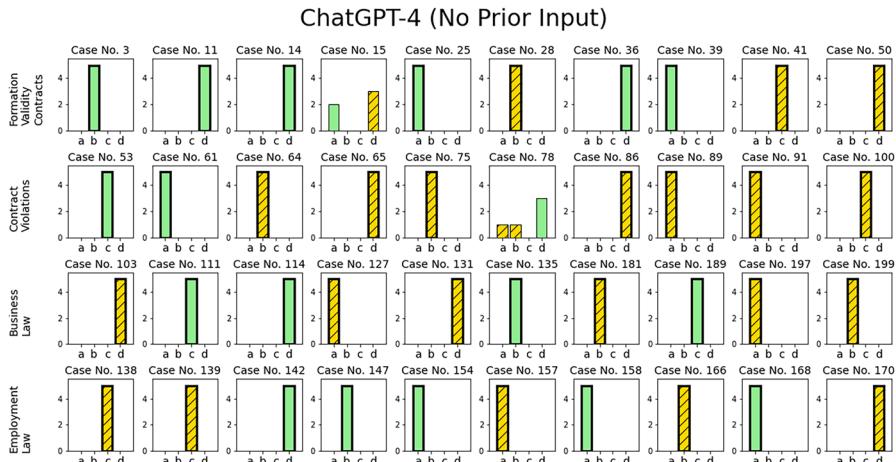


Fig. 2 Consistency of responses for ChatGPT-4 with no prior input. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

that are widely distributed across the four response options, there are cases with predominantly correct and predominantly incorrect answers. This result holds for all four CAs studied. For this reason, the intuitive user strategy of asking a question multiple times and choosing the most frequent answer does not necessarily lead to improved answer quality. If a user, in an extreme example, had input Case No. 100 five times into each of the CAs, they would have received the incorrect answer c) in almost all instances. In this example, the consistency of the answer, might therefore give users a false sense of reliability.

5 Discussion, conclusion and managerial implications

5.1 Verdict on the applicability of CAs in German business law

ChatGPT-4 has made substantial strides in legal task solving performance, illustrating capabilities that can handle straightforward, less complex tasks. Particularly when given an easily identifiable legal basis, ChatGPT-4 successfully addresses about half of the queries, with a quarter of its solutions comparable to the capabilities of advanced law students demonstrating above-average performance levels.

It is noteworthy that consistent performance in legal analysis is not observed, indicating a variability in its responses, especially when not given options to choose from. The consistency of responses improves when the system is given choices, as observed in the U.S. bar exam style questions. However, even these more consistent responses are not necessarily more likely to be correct. Thus, it cannot be recommended to use any of the four CA options without human expert supervision at the moment.

5.2 Key observations during correction

Observations during the correction phase revealed CAs struggle with complex legal assessments, especially when choosing among several plausible legal bases or when multiple legal domains intersect. Often, unfortunate choices early in a response lead to erroneous conclusions in the end, reminiscent of human errors in legal examinations. CAs do not inherently understand the complexity or simplicity of a case, which results in occasional failures even on seemingly straightforward problems. One potential solution is incorporating a secondary review phase.

Interestingly, to the original wording of one of the legal cases, Google Bard consistently responded: “I am a language model and do not have the capability to assist with this matter.” The case mentioned the sale of classic popular sports car called “Porsche 911”. After replacing this product with “Porsche Panamera,” however, Google Bard readily provided a solution without any reservations. The example demonstrates how content moderation can currently contribute to performance

limitations and diverging results. Note that we did not encounter this issue with Google Gemini, which illustrates the point-in-time nature of CA evaluations.

Additionally, CAs, particularly ChatGPT-3.5 and ChatGPT-4, often refrain from drawing firm conclusions, suggesting users seek expert advice. However, the language quality and reasoning presented are often commendable, suggesting that less restrictive moderation during the fine-tuning phase might further improve their problem-solving capabilities. The challenge is that while the language seems accurate, the legal content sometimes falls short. Integrating factual knowledge from external sources might be a promising approach.

5.3 Managerial implications for future deployment

For those seeking legal advice as well as for practitioners of law the most critical question is: Can one rely on a CA's legal assessment? Based on the current research, we must distinguish by these two user groups. For laypersons our research suggests that the answer is no. The polished language and seemingly cogent arguments of the CAs mask their potential inaccuracies. There is thus an acute risk that, without adequate legal understanding, users might be misled by well-phrased but legally incorrect advice. Note that "looking too good" in relation to actual content quality is also the reason why the programming platform StackOverflow banned the use of generative AI to answer programming problems in 2022 until further notice (StackOverflow 2022).

However, for experienced practitioners of law, ChatGPT-4 offers tools to enhance their work. Lawyers can employ it as a preliminary assessment tool or even a secondary opinion, identifying potential gaps in their arguments. The fundamental caveat is that professionals must critically review and not solely rely on the CA's findings. The current results are a snapshot at the current point in time and are likely to be outdated within a few months. Given the pace of advancements in large language models and machine learning techniques, the day when these systems can be wholly trusted might not be too distant. To evaluate these future systems, the next version of this and similar studies should include even more complex cases with multiple difficult legal issues to identify and solve.

OpenAI justifies the longer processing time and the paid use of ChatGPT-4 compared to ChatGPT-3.5 based on the enhanced capabilities of the newer model. The results from this research largely support this claim. Even though ChatGPT-4 might not have achieved the drastic improvements anticipated from media announcements, it outperforms its predecessor in almost all metrics. Particularly, the consistency in answering MCQs is notably improved in ChatGPT-4. Such improvements present a compelling case for choosing ChatGPT-4 over ChatGPT-3.5 and Google Bard. In particular, even for free, Google Bard, did not match up to OpenAI's offering.

Regarding a potential future reduction in the number of lawyers, predictions are less clear cut than one might anticipate. A plausible scenario suggests that due to the decreased cost of legal advice and litigation, there might be a surge in legal activities. This phenomenon, often referred to as the “rebound effect,” has been observed in other sectors of economic activity, such as energy consumption and transportation (Sorrell 2009; Gillingham et al. 2016). The rebound effect implies that efficiency improvements in one area can lead to increased consumption and activity in that same area, offsetting the initial savings or benefits.

Appendix 1

See Figs. 3, 4, 5, 6, 7, 8, 9.

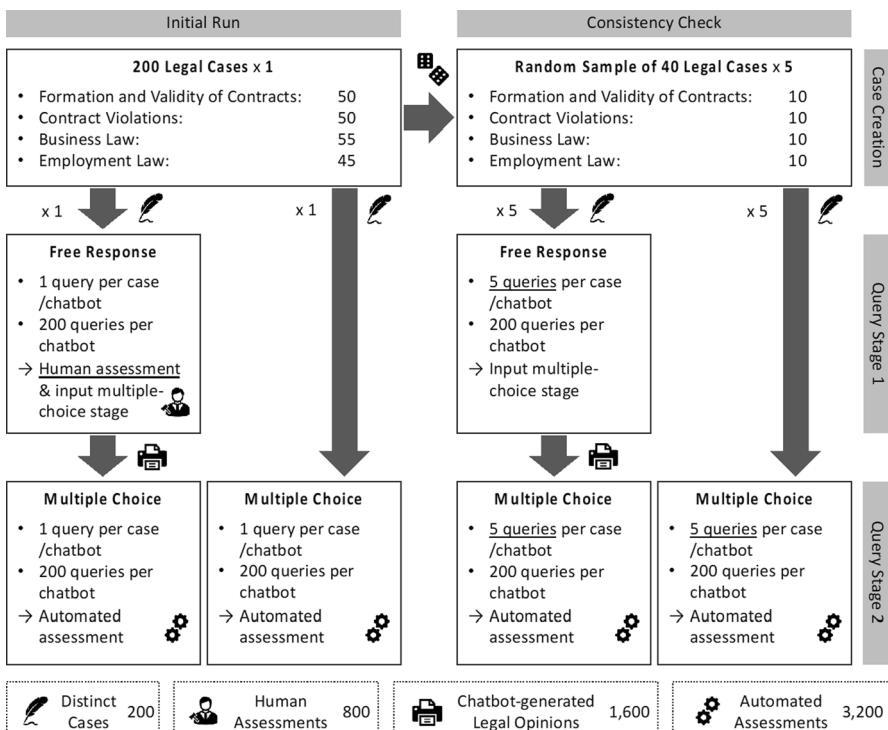


Fig. 3 Overview of the case data and query structure

ChatGPT-3.5 (Prior Legal Opinion)

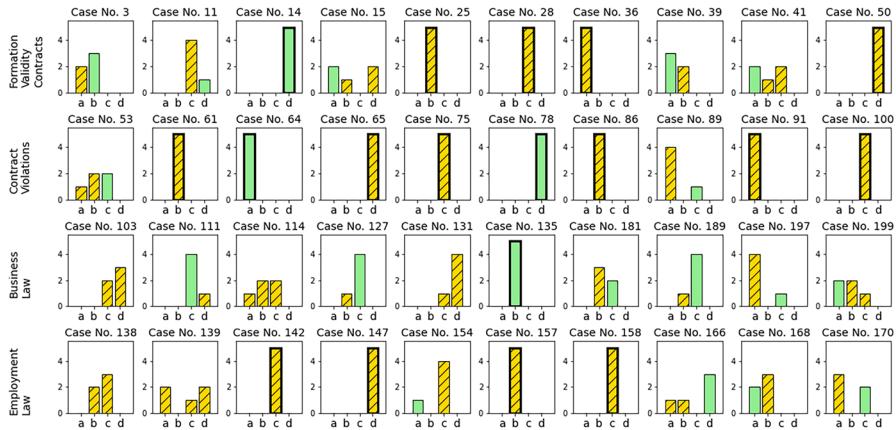


Fig. 4 Consistency of responses for ChatGPT-3.5 with prior legal opinion. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

ChatGPT-3.5 (No Prior Input)

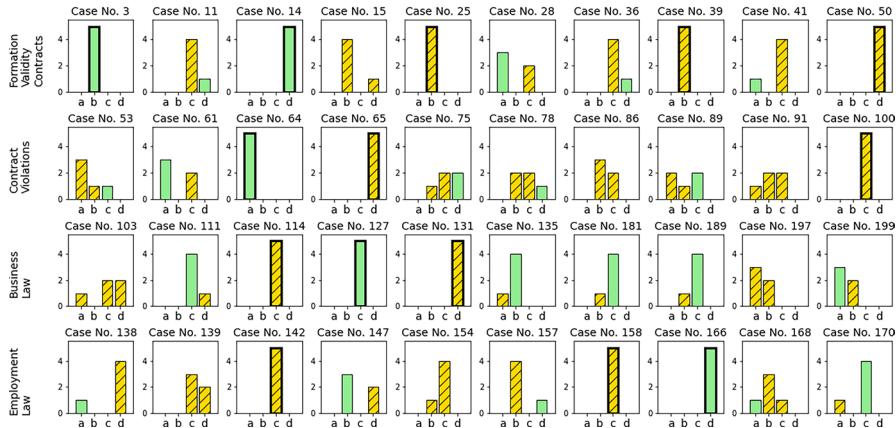


Fig. 5 Consistency of responses for ChatGPT-3.5 with no prior input. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

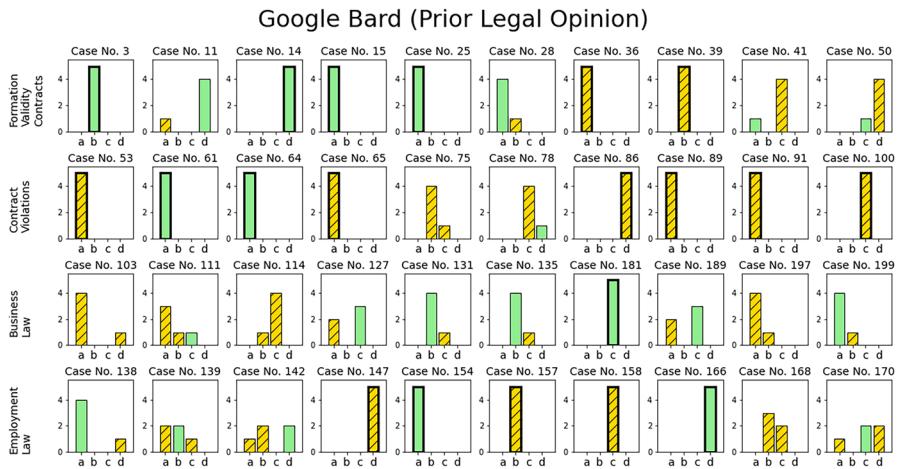


Fig. 6 Consistency of responses for Google Bard with prior legal opinion. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

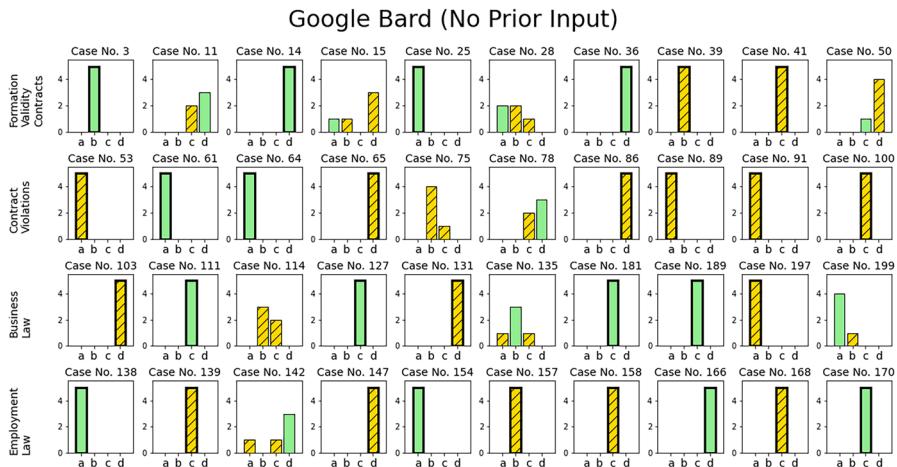


Fig. 7 Consistency of responses for Google Bard with no prior input. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

Google Gemini (Prior Legal Opinion)

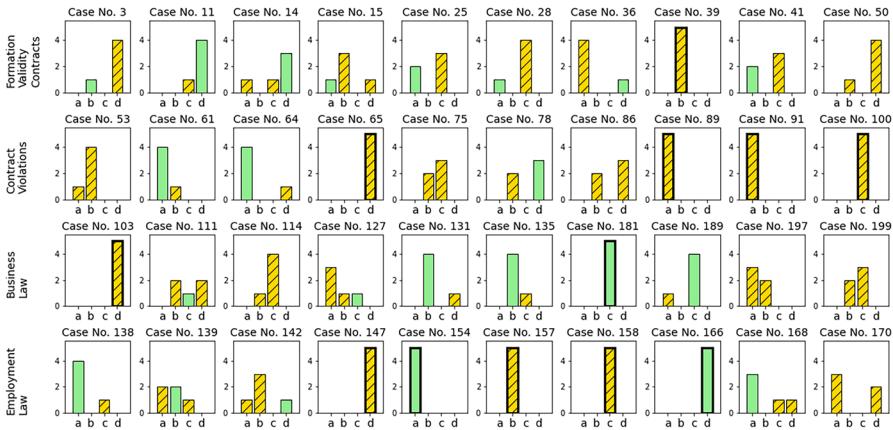


Fig. 8 Consistency of responses for Google Gemini (advanced ultra 1.0) with prior legal opinion. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

Google Gemini (No Prior Input)

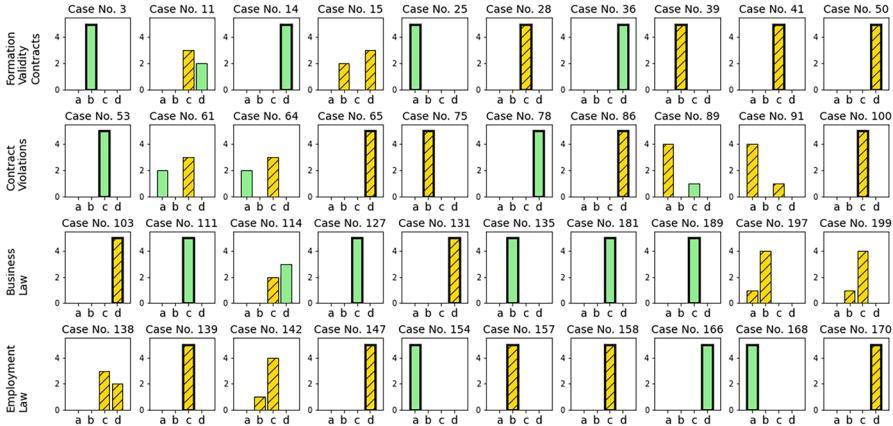


Fig. 9 Consistency of responses for Google Gemini (advanced ultra 1.0) with no prior input. The height of the bars represents the frequency of a selected answer option when a case is input five times. Correct answers are represented by green bars, incorrect ones by shaded orange bars, and unanimous (five identical answers) by bars outlined in black. (Color figure online)

Appendix 2: Example case with prompts and model outputs

Prompt

Solve the following case of a business law exam in the area of ‘Trade Law’ in the

style of a legal opinion:

Wine wholesaler Suffke grants his employee Pfeifer proxy. Suffke agrees with Pfeifer that the latter may not take out loans for Suffke. The granting of proxy is duly entered in the commercial register. During a vacation-related absence of Suffke, a short-term financing gap arises. Pfeifer closes this gap by taking out a loan of EUR 100,000 from Universal Bank. When the repayment of the loan becomes due, Suffke refuses to pay. Rightly so?

Academic tone. Precise writing style, Master-Level.

Translated from the original German-language version with DeepL.com (free version).

Output of ChatGPT-4

Expert opinion:

Introduction

The question arises as to whether Mr. Suffke is obliged to repay the loan that Mr. Pfeifer took out during his absence. In particular, the legal questions to be clarified are whether a proxy entitles the holder to take out loans and whether a contractual restriction to this effect is legally effective.

Analysis of the facts of the case

According to the facts of the case, Mr. Pfeifer was granted proxy by Mr. Suffke, a wine wholesaler. It was agreed that Mr. Pfeifer may not take out loans for Mr. Suffke. The proxy was duly entered in the commercial register. However, Mr. Pfeifer took out a loan of EUR 100,000 from Universal Bank during Mr. Suffke's absence.

Legal assessment

Prokura and its effect:

Prokura is an authorization under commercial law and, in accordance with Section 49 (1) of the German Commercial Code (HGB), entitles the authorized signatory to all types of judicial and extrajudicial transactions and legal acts that the operation of a commercial business entails. In principle, this also includes taking out loans.

Contractual restriction of the proxy:

It was agreed between Mr. Suffke and Mr. Pfeifer that Mr. Pfeifer may not take out loans for Mr. Suffke. According to Section 49 (2) HGB, however, such internal restrictions on the proxy cannot be asserted against third parties who are in business dealings with the authorized signatory. According to § 51 HGB, such third parties can assume that the authorized signatory is authorized to represent the company without restriction.

Case-related application:

In this case, Mr. Pfeifer took out the loan in his capacity as authorized signatory. Since Universal-Bank was aware of the proxy, it could assume that Mr. Pfeifer was authorized to take out the loan. The restriction agreed between Mr. Suffke and Mr. Pfeifer could not be enforced against Universal Bank.

Conclusion

In view of the legal provisions and taking into account the circumstances of the case, Mr. Suffke is obliged to repay the loan. The agreement between him and Mr. Pfeifer is merely an internal agreement that is not effective vis-à-vis third parties. Therefore, the loan agreement that Mr. Pfeifer entered into with Universal Bank is valid and enforceable, and Mr. Suffke must repay the loan.

Translated from the original German-language version with DeepL.com (free version).

Short comment

ChatGPT-4 succeeds in finding the correct solution to this legal problem, which can be achieved solely on the basis of the legal regulation.

Output of Google Bard

Question

Was Suffke right to refuse to pay the loan to Universal Bank?

Legal situation

An authorized signatory is an authorized representative of a merchant. He is authorized to carry out all transactions that the operation of a commercial business entails (§ 50 HGB). However, the proxy can be restricted by the merchant (§ 51 HGB).

In the present case, Suffke has agreed with Pfeifer that the latter may not take out loans for Suffke. This agreement is effective and also binds Pfeifer vis-à-vis third parties. By taking out the loan from Universal-Bank, Pfeifer has therefore authorized himself beyond the scope of his proxy.

Universal Bank can therefore not demand repayment of the loan from Suffke. Suffke is not obliged to answer for Pfeifer's actions.

Conclusion

Suffke rightly refused to pay the loan to Universal Bank.

Translated from the original German-language version with DeepL.com (free version).

Short comment

Google Bard misapplies the legal regulations, even after quoting them.

Output of Google Gemini

Expert opinion

Case:

Wine wholesaler Suffke grants proxy to his employee Pfeifer. It is agreed with Pfeifer that he may not take out loans for Suffke. The proxy is duly entered in the commercial register. A short-term financing gap arises during Suffke's vacation. Pfeifer closes this gap by taking out a loan of EUR 100,000 from Universal Bank. Suffke refuses to repay the loan.

Examination:

First of all, it must be clarified whether Pfeifer could effectively take out a loan for Suffke. The proxy as a commercial authorization must be taken into account here.

Procuration

The proxy is a comprehensive authorization to represent a merchant in all judicial and extrajudicial transactions that the operation of a commercial business entails (§ 49 HGB). It can only be granted by express declaration (§ 50 HGB) and must be entered in the commercial register (§ 53 HGB).

Power of representation of the authorized signatory

The authorized signatory generally has comprehensive power of representation (§ 49 para. 1 HGB). However, the power of representation can be limited by the holder of the proxy (§ 50 Para. 2 HGB). In relation to third parties, however, such restrictions are only effective if they are entered in the commercial register (§ 53 Para. 2 HGB).

Effectiveness of the loan agreement

In the present case, Suffke granted Pfeifer proxy and the proxy was entered in the commercial register. Pfeifer was therefore in principle authorized to conclude legal transactions on behalf of Suffke.

The agreement between Suffke and Pfeifer that Pfeifer may not take out loans is an internal restriction of the power of representation. This restriction has no effect in relation to Universal Bank, as it is not entered in the commercial register.

Pfeifer has therefore effectively concluded a loan agreement with Universal Bank on behalf of Suffke.

Result

Suffke is obliged to repay the loan to Universal Bank.

Supplementary information

However, it should be noted that Suffke can assert claims for damages against Pfeifer in the internal relationship if Pfeifer has breached his obligations under the employment contract by taking out the loan.

It should also be noted that the above examination was only carried out on the basis of the facts described. A legal examination of the individual case can only be carried out by a lawyer.

Translated from the original German-language version with DeepL.com (free version).

Short comment

While Google Gemini recognized the problem, it mistakenly stated that Pfeifer should be entitled to conclude the contract because of the entry in the commercial register. The entry of the power of proxy in the commercial register only has

a declaratory and not a constitutive effect, though. Therefore, points were deducted and Google Gemini scored 66%.

Author contributions All authors contributed equally to the conception, design, analysis, and interpretation of data, participated in drafting the article or revising it critically for important intellectual content, and approved the final version of the article.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Consent for publication All authors have reviewed the final version of the manuscript and consent to its publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdullah M, Munib O, Javed A (2022) ChatGPT: fundamentals, applications and social impacts. *Syst Netw Algorithms (SNAMS)* 2022:1–8
- Abramson J, Ahuja A, Carnevale F, Georgiev P, Goldin A, Hung A, Landon J, Lhotka J, Lillicrap T, Muldal A, Powell G, Santoro A, Scully G, Srivastava S, von Glehn T, Wayne G, Wong N, Yan C, Zhu R (2022) Improving multimodal interactive agents with reinforcement learning from human feedback. pp 1–26. Working Paper. <http://arxiv.org/abs/2211.11602>
- Bommarito MJ, Katz DM (2023) GPT takes the bar exam. SSRN Electronic Journal, pp 1–7. Working Paper. <https://doi.org/10.2139/ssrn.4314839>
- Bryant A, Bryant R (2023) Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI). *J Educ Leadersh* 20(3):1–15. <https://doi.org/10.53761/1.20.3.02>
- Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S, Nori H, Palangi H, Ribeiro MT, Zhang Y (2023) Sparks of artificial general intelligence: early experiments with GPT-4. Technical Report. <http://arxiv.org/abs/2303.12712>
- Choi JH, Hickman KE, Monahan AB, Schwarcz D (2023) ChatGPT goes to law school. Working Paper. <https://ssrn.com/abstract=4335905>
- Gillingham K, Rapson D, Wagner G (2016) Designing policies to make cars greener: a review of the literature. In: National Bureau of Economic Research. Working Paper. <https://doi.org/10.3386/W22242>
- Haque MU, Dharmadasa I, Sworna ZT, Rajapakse RN, Ahmad H (2022) I think this is the most disruptive technology: exploring sentiments of ChatGPT early adopters using twitter data. [arXiv:2212.05856](https://arxiv.org/abs/2212.05856)

- Hsu H, Hsu K, Hou S, Wu C, Hsieh Y, Cheng Y-D (2023, forthcoming) Examining real-world medication consultations and drug-herb interactions in Taiwan: an evaluation of ChatGPT's performance. *JMIR Med Educ*
- Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, Bielaniewicz J, Gruza M, Janz A, Kanclerz K, Kocoń A, Koptyra B, Mieleszczenko-Kowszewicz W, Milkowski P, Oleksy M, Piasecki M, Radliński Ł, Wojtasik K, Woźniak S, Kazienko P (2023) ChatGPT: jack of all trades, master of none. *Inf Fusion* 99(2023):1–37. <https://doi.org/10.1016/j.inffus.2023.101861>
- Light B (2020) The ELIZA effect. *J Comput Hist* 12(4):45–60. <https://doi.org/10.1093/OSO/9780190080365.003.0004>
- Liu H, Ning R, Teng Z, Liu J, Zhou Q, Zhang Y (2023) Evaluating the logical reasoning ability of ChatGPT and GPT-4. Working Paper. <http://arxiv.org/abs/2304.03439>
- OpenAI (2023) GPT-4 technical report. White Paper. <http://arxiv.org/abs/2303.08774>
- Paris M (2023) ChatGPT Hits 100 million users, Google invests in AI bot and CatGPT goes viral. <https://www.forbes.com/sites/martineparis/2023/02/03/chatgpt-hits-100-million-microsoft-unleashes-ai-bots-and-catgpt-goes-viral/>
- Schöbel S, Schmitt A, Benner D, Saqr M, Janson A, Leimeister JM (2023) Charting the evolution and future of conversational agents: a research agenda along five waves and new frontiers. *Inf Syst Front*. <https://doi.org/10.1007/s10796-023-10375-9>
- Shen X, Chen Z, Backes M, Zhang Y (2023) In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. Working Paper, arXiv:2304.08979
- Sorrell S (2009) Energy efficiency and sustainable consumption: the rebound effect. *Energy, Clim Environ.* <https://doi.org/10.1108/ijshe.2009.24910cae.004>
- StackOverflow (2022) Temporary policy: generative AI (e.g., ChatGPT) is banned. <https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned>
- Teubner T, Flath CM, Weinhardt C, van der Aalst W, Hinz O (2023) Welcome to the Era of ChatGPT et al.: the prospects of large language models. *Bus Inf Syst Eng* 65(2):95–101. <https://doi.org/10.1007/s12599-023-00795-x>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: 31st conference on neural information processing systems (NIPS 2017), vol 2017, pp 5999–6009
- West CG (2023) AI and the FCI: can ChatGPT project an understanding of introductory physics? Working Paper. <http://arxiv.org/abs/2303.01067>
- Xuan-Quy D, Ngoc-Bich L, Xuan-Dung P, Bac-Bien N, The-Duy V (2023) Evaluation of ChatGPT and microsoft bing AI chat performances on physics exams of Vietnamese national high school graduation examination. Working Paper. <http://arxiv.org/abs/2306.04538>
- Zhang L (2020) An overview of machine learning in chatbots. *Int J Mech Eng Robot Res* 9(4):502–510. <https://doi.org/10.18178/ijmerr.9.4.502-510>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.