On Evaluating Legal-Reasoning Capabilities of Generative AI

Henry Prakken[1,*,†]

[1]Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands

Abstract

This paper critically examines some recent studies of the legal-reasoning capabilities of generative AI. It also discusses which roles traditional symbolic approaches can have in the era of generative AI.

Keywords

Legal argumentation, Large language models, Evaluation

1. Introduction

The introduction of ChatGPT by OpenAI in November 2022 was a 'big bang' in AI. Never before was an AI tool available for so many and so easy to use for so many different tasks. The ease with which it generates flawless natural language for a wide variety of tasks such as summarising documents, writing essays about any given topic, writing poems, drafting travel plans, outlining presentations, and even solving computer-programming exercises is amazing. And all this essentially with the simple technique of predicting the most likely next word in a sequence of words. It is therefore easy to think that traditional symbolic AI research on reasoning and argumentation is now obsolete and that the right way to let the computer engage in reasoning and argumentation is by using generative AI founded on large language models.

This paper addresses this issue for argumentation in the law, which is an important application domain of computational argument. Several experiments have already been conducted on how large language models (LLMs) perform on legal reasoning tasks. This paper reviews some these experiments and more generally discusses the potential of generative AI to engage in legal argumentation. We first briefly summarise AI & law research on legal argument in Section 2. Then we make some methodological observations in Section 3 and review recent experiments in applying LLMs to legal reasoning in Section 4. We then discuss what the field of computational argumentation can learn from these studies in Section 5, after which we conclude.

The 24th International Workshop on Computational Models of Natural Argument (CMNA'24)

h.prakken@uu.nl (H. Prakken)

https://webspace.science.uu.nl/~prakk101/ (H. Prakken)

0000-0002-3431-7757 (H. Prakken)

2. Brief overview of AI & law research on modelling legal argument

Argumentation is "...the giving of reasons to support or criticize a claim that is questionable, or open to doubt" [1, p. 285]. The field of AI & Law has developed formal and computational models of legal argumentation since [2]. For overviews see [3, 4, 5]. Both rule- and case-based approaches have been applied, initially as alternatives but more recently as complementing each other, since case-based reasoning often is about whether a rule's conditions are satisfied. Rule-based approaches have to account for the defeasibility of legal rules. Since rule-makers cannot foresee everything, rule-appliers sometimes have to make exceptions in unforeseen circumstances. Defeasibility also arises because of presumptions and allocations of burdens of proof [6]. Several early rule-based accounts of legal reasoning used some form of logic-programming [7, 8]. Later, explicitly argument-based formalisms were applied or developed [9, 10, 11], as well as formalisms with anargumentative flavour such as Defeasible Logic [12] and abstract dialectical frameworks [11].

Case-based approaches were initially developed to account for the fact that in Anglo-American jurisdictions case law instead of legislation traditionally is the main source of law, where courts have to decide new cases by drawing analogies to decided cases. Case-based approaches have to account for the fact that cases often not just have similarities but also differences. Seminal work was on HYPO [13, 14], in which cases were modelled with sets of features called dimensions, which have partially ordered values that make cases better or worse for a particular outcome. HYPO generated three-ply arguments between a plaintiff and a defendant in a civil dispute, drawing analogies between or distinguishing cases from their respective point of view. In later work this was refined in many ways, for instance, by distinguishing features of various levels of abstraction [15] or by comparing cases in terms of how case decisions promote or demote legal or social values [16, 17]. Perhaps the most ambitious approaches are coherence-based accounts, which model the construction of legal theories of some kind that explain a set of cases and where the most coherent theory that does so should be adopted [2, 18, 19].

While early on rule- and case-based approaches were presented as alternatives, later the awareness arose that they complement each other, since case-based reasoning often is about whether a rule's conditions are satisfied. A challenge for rule-based accounts is that the conditions of legal rules are often vague and general, and no clear rules can be given for when they are satisfied. Here case-based approaches can complement rule-based approaches by providing forms of argumentation for interpreting legal concepts [20, 21, 4].

In sum, AI & law has developed rich models of various forms of legal argument, including rule-based, case-based and value-based accounts, which draw on various sources, including legislation, case law and social and moral value considerations. Moreover, all this work takes a knowledge-based approach: the required knowledge is encoded in a symbolic form that is understandable for the machine and the computer reasons with it in a formally defined way, ideally based on the laws of logic and rational reasoning. The advantages of this approach are transparency and explainability: humans can see which knowledge the machine uses and the machine can explain its outcomes by showing how it reasoned with this knowledge. A big disadvantage of this approach is that it is often hard to acquire and represent a sufficient

amount of knowledge in a form that can be manipulated by the machine. This is the notorious knowledge acquisition bottleneck. Hence the attractiveness of large language models as a means to generate legal argument.

3. Methodological remarks

Evaluation of knowledge-based AI applications can be done at three levels: evaluating the inputted knowledge, evaluating the reasoning mechanism (for instance, on whether it implements some philosophically acceptable model of rational reasoning, or on soundness and completeness properties with respect to such a model) and evaluating the output. When evaluating applications of generative AI, evaluation at the first two levels becomes hard so often only the system's output is evaluated. Moreover, this output is natural language instead of some formal language, with all the ambiguity and vagueness that comes with it, so interpreting the output is not always an easy task. In consequence, evaluation studies of generative AI are inherently experimental, often statistical and can involve subjective elements.

3.1. Terminology: prompt engineering

A well-known drawback of LLMs is that there is no connection between the large statistical language model learned by the LLM and reality. All that an LLM 'knows' is how often words go together in similar contexts. This often causes a LLM to 'hallucinate' facts. While this may not be a problem for creative applications like writing prose or poetry, it is a serious problem when an LLM is asked to produce high-quality information or arguments in high-stake contexts; and legal contexts are often high-stake.

There is much research on addressing this problem. Many of them involve *prompt engineering*, that is, applying ingenious ways to write the prompts that are the user input of LLM applications. *Zero-shot prompts* do not contain any examples of desired output but directly ask a question or specify a task. *Few-shot prompts* do provide such examples. *Chain-of-thought* (CoT) prompting consists of ways to ask the model to 'think' step-by-step rather than solving a complex problem at once. Zero-Shot CoT does just that while few-shot CoT methods combine it with examples of desired output. Such prompts are often formulated as problems of *pattern completion*, which consist of showing the model a pattern of expected answers: this increases the probability that the model will indeed give an answer in terms of this pattern. Yet another way in which prompts can be engineered is to include one or more documents that have to be taken into account in the model's answers. When these documents are retrieved from other sources after entering a prompt, this is called *retrieval-augmented generation*. In legal applications it makes sense to include or retrieve legislation or case law.

3.2. Questions asked about the studies

In this paper we will review all serious recent studies of the legal-reasoning capabilities of generative AI that we know of, with among other things attention for their prompt-engineering methods. We will ask the following questions about the reviewed studies.

- Which reasoning capability is tested and according to which reasoning model?

- How direct was the testing? Were proxies for reasoning abilities used?
- Which method of prompt engineering is used?
- How systematic is the evaluation? Is it subjective or objective, qualitative or quantitative?
- What is compared? LLMs or prompting methods against each other or also against human performance?

4. Recent experiments on legal reasoning by LLMs

All reviewed studies study tasks that involve legal reasoning, though in different ways. Some studies are about making exam questions, other studies involve specific reasoning tasks (mostly rule application) and some studies are about the generation of legal documents that typically contain argumentation. Many studies apply or refer to the IRAC method of legal reasoning, popular in Anglo-American legal education. IRAC stands for Issue-Rule-Application-Conclusion. Here, Issue is the task of determining the legal issue of a case, Rule is the task of identifying the relevant legal rules (which can also be precedents), Application is the task of determining how the rules should be applied to the facts, and Conclusion is the task of drawing a legal conclusion from the rule application. While in reality issue spotting can be far from trivial, in all studies reported below the issue is in fact given. Then we see that the IRAC model in fact abstracts from all AI & law models of legal reasoning discussed above in Section 2.

4.1. Studies on document generation

There are a few studies on legal document generation.

Perlman [22], in an informal experiment with zero-shot prompts asks ChatGPT, among other things, to suggest arguments to make in a brief about a particular legal issue, to draft a legal complaint and to perform an initial legal analysis of a brief factual scenario. Perlman then gives his own informal qualitative opinion, observing among other things that ChatGPT's output is "surprisingly sophisticated" though "incomplete and problematic in numerous ways". The outputs "would not be sufficiently helpful in their current forms for most people".

Iu & Wong [23] conduct a similar informal experiment on the basis of a simplified description of the facts in a well-known American case, asking ChatGPT with zero-shot prompts to perform various writing tasks. Some of these tasks involve the production of legal arguments, such as drafting a pleading claim, drafting a skeleton argument with the support of case law and drafting a judgement considering both sides. The authors then subjectively evaluate the documents, observing among other things that ChatGPT "demonstrated its ability to understand simple facts and articulate the legal basis of a claim", "was able to ...summarise the key facts of relevant case law to support the plaintiff's case", and "was able to apply the reasoning of case law to the simple facts of the case, thus demonstrating an ability to follow the IRAC approach in writing the skeleton argument". When applied to a second, more complicated case, ChatGPT "performed excellently" in drafting skeleton arguments and "was able to draft the judgment by considering the arguments of both sides with logical reasoning".

In sum, both these studies have no explicit testing of reasoning capability, so testing is indirect, the prompts were zero-shot, evaluation is unsystematic and subjective, and no comparisons are

made. Because of the lack of explicit and objective evaluation standards, these studies cannot provide valid and reliable results, though they can have heuristic value.

Trozze et al. [24], in the domain of cryptocurrency security cases, tested ChatGPT on writing a complaint for a class action lawsuit. The complaint was compared to one written by a lawyer by letting a mock jury decide on the basis of both. The prompt only asked ChatGPT to write the various parts of a complaint and did not give reasoning instructions. ChatGPT was then evaluated in terms of how often the jury gave the same decision on the basis of both complaints. The jurors were in 88% of the cases in which the human lawyer drafted the complaint convinced that the allegations were proven; for AI-drafted complaints this figure was 80%. The authors conclude from this that "Overwhelmingly, ChatGPT drafted convincing complaints, which performed only slightly worse than the lawyer-drafted ones". More generally, the authors conclude that ChatGPT is better is in drafting legal documents than in statutory reasoning (citing others for the same conclusion).

In sum, the prompt did not give any reasoning instructions, and no explicit testing of reasoning capability. Testing is thus indirect, with as proxy how often ChatGPT agrees with the human lawyer. Systematic quantitative evaluation but no real comparison with human performance, since the human lawyer's performance is used as the evaluation standard.

4.2. Studies on exam performance

Several studies let the models take legal exams or answer exam questions. These studies mostly indirectly evaluate legal-reasoning capacities, since many exam questions do not directly test the student's reasoning or argumentation skills. These studies can only be regarded as evaluating such skills on the assumption that successfully making legal exams requires such skills. This may not always be the case since questions can also test for the possession of legal knowledge.

Yu et al. [25] let GPT-3 answer a type of question from the Japanese Bar exam, modelled as an entailment task from the COLIEE competition [26]. Given a legal rule and a legal question (hypothesis), GPT-3 has to answer whether the hypothesis is true or false, with a brief explanation. This looks likes rule application without chaining of rules. The authors test several prompting methods: zero shot (simply asking whether the hypothesis is true given the rule), few-shot (giving 1, 3 of 8 examples of desired output) and a two-stage form of CoT prompting, first asking 'let's think step-by-step' and using the output as the input for the prompt 'therefore, the hypothesis is (true or false)'. Answers are quantitatively evaluated in terms of the known correct answer. The accuracy is between 61 and 75%. Then the authors finetune GPT-3 with the COLLIEE data set. Accuracy is between 61 and 77%. Finally, the authors ask GPT-3 in the prompt to apply a particular reasoning method, which all are variants of the IRAC method. The prompts just mention the required approach but do not explain it. Accuracy is between 66 and 81%. The authors observe qualitatively that the models appear to apply the indicated reasoning method. The authors observe that the few-shot approaches with example-and reasoning prompts outperform previous winners of the COLIEE competition but they do not compare with human performance.

In sum, explicit testing of reasoning abilities so testing is direct, zero-shot and few-shot prompting, some prompts ask to apply a particular mentioned but undefined reasoning method, systematic quantitative evaluation, comparisons between prompting methods and with other

NLP methods but not with human performance.

Choi et al. [27] tested ChatGPT on four law school exams, each consisting of an essay part and a multiple-choice part. They used zero-shot prompts consisting of the exam question and for the multiple-choice part they alternatively tested CoT prompting, asking to provide a chain of reasoning as well as giving a letter answer to the question. Three of the authors blindly graded exams made both by ChatGPT and by students. The authors found that ChatGPT passed all exams but that compared to the human students it "generally scored at or near the bottom of each class". Also, ChatGPT scored better on the multiple-choice questions than on the essay questions and CoT prompting performed worse than the zero-shot prompts although the difference was not statistically significant. As regards the essay questions, the authors qualitatively observed that ChatGPT was poor in arguing why a rule applied to given facts and that it did not systematically answer in terms of IRAC or some other reasoning model.

In sum, no explicit testing of reasoning capability/model, except for a basic form of CoT prompting for the multiple-choice questions. Testing is indirect, with the exam scores as proxy for legal reasoning abilities. Systematic quantitative evaluation in terms of exam scores, comparisons between prompting methods and with human performance.

Katz et al. [28] tested the performance of GPT-4 on a simulated version of the American bar exam. The exam consists of a part with essay questions and a part with multiple-choice questions. The answers on essay questions were evaluated by two academic legal experts on the basis of a collection of "representative" good answers available online. [28] make various claims about GPT-4's performance, the most important one being that it has passed the exam. Although [29] casts doubt on some of [28]'s claims, he agrees that their main claim is justified. This implies that GPT-4 performs comparably to human legal experts on the bar exam.

In sum, no explicit testing of reasoning capability. Testing is indirect, with exam score as a proxy for legal reasoning abilities. The zero-shot prompts correspond to the exam questions. Systematic quantitative evaluation and comparisons with human performance.

Nay [30] made their own selection of multiple-choice questions (four options, of which one correct) for American tax law, with randomly generated fact, names and numbers to avoid that the questions can be in a model's training set. They compare several prompting methods that inject legal information to the prompt to a zero-shot prompt that simply asks the question. One method injects potentially relevant statutes resulting from a similarity search into the prompt. Another method directly provides as context the relevant part of the law. A final method provides context in the form of a lecture note relevant given to the question type written by a law professor (one of the authors). Then various LLMs, including GPT-4, are compared on accuracy, where in some experiments the prompting method is combined with CoT prompting. Generally, GPT-4 performs the best, while CoT improves performance but not consistently.

In sum, implicit testing of reasoning capability, (deductive rule application without chaining). Testing is indirect. The prompts provide relevant legal information but give no information about or explicit examples of the expected reasoning. Systematic quantitative evaluation and comparisons between prompting methods but no comparisons with human performance.

4.3. Studies on specific reasoning tasks

Jiang & Yang [31] study how GPT-3 classifies brief factual scenarios as a criminal offence

(choice of one from eight). They include a brief explanation of the legal syllogism (basically modus ponens with legal rules) in the prompt, without examples: 'In the legal syllogism, the major premise is the law article, the minor premise is the facts of the case and the conclusion is the outcome of the judgment'. Then the prompt gives a brief factual scenario and asks GPT-3 to 'use the legal syllogism to think and output the judgment'. The output gives a major and a minor premise and a conclusion. The evaluation standard is the given 'correct' classification. GPT-3 has a higher accuracy with this method (68.5%) than with simply giving the case and asking for the judgment (64.5%) and with zero-shot CoT prompting with 'let's think step-by-step' (58.8%).

In sum, explicit testing of reasoning capability, namely, the legal syllogism (deductive rule application without chaining). Testing is direct. The prompts contain an explanation of the reasoning method and ask to apply it. Systematic quantitative evaluation and comparisons between prompting methods but no comparisons with human performance.

Similar work is **Deng et al. 2023** [32], who use four subsequent prompts corresponding to the stages of an IRAC-like process (article retrieval, recognising criminal elements in facts, applying articles, providing judgment) as part of the overall task to predict judgments and penalties. The four-stage process is compared on predictive performance with a 'plain-text' method and is found to generally but not always outperform the latter.

In sum, explicit testing of IRAC-like reasoning capability. Testing is indirect in terms of predictive performance. The breakdown into four prompts corresponds to IRAC-style reasoning. Systematic quantitative evaluation and comparisons between prompting methods but no comparisons with human performance.

A limitation of both [31] and [32] is that the test data apparently only contain convictions, so that the models cannot reason about *whether* a suspect is guilty.

Kang et al. [33] let ChatGPT evaluate scenarios of which the correct analysis is formulated with the IRAC-method in a semi-structured logical language, where the issues are given. This thus tests how well ChatGPT identifies the rules, the conclusion and the reasoning steps from the facts to the conclusion. It seems that the scenarios all are chains of if-then rules but this is not fully clear from the appendix. ChatGPT's outputs are evaluated by humans in terms of "the marking rubrics used by law schools". Then the quantitative measures precision, recall and F1 are calculated. The scores vary but are never very high. The authors first give zero-shot prompts without knowledge or examples and no request to use IRAC. When only the conclusion should be provided (yes/no), ChatGPT performs rather well but especially the reasoning is poor. Next they add, respectively, 20, 40 and 80% of the reasoning paths and observe improved scores. The same happens when examples are given in the prompt and when the problems are decomposed into subquestions (a kind of CoT prompting). This in fact codes the rules used in the reasoning paths in the subquestions.

In sum, explicit testing of reasoning capability, namely, IRAC. Testing is direct, since ChatGPT is evaluated on how well it can reproduce pre-encoded IRAC structures. Various zero- and few-shot prompting methods are used, giving less or more of the desired solution. Systematic quantitative evaluation, comparisons between prompting methods but not with human performance. An important thing to note is that quite some structure is added to the prompts.

Blair-Stanek et al. [34] test how well GPT-3 can perform "statutory reasoning", which they essentially see as deductive rule application including chaining. They use a data set containing non-ambiguous tax laws and test cases with unique correct answers. The questions GPT-3 has

to answer are of the form 'Premise - Hypothesis' and GPT-3 has to answer whether the relation between them is 'entailment' or 'contradiction'. Several zero- and few-shot prompting methods are used, with and without including a relevant statute or examples, and some also including 'let's think step-by-step'. The prompts ask to do 'Entailment/Contradiction reasoning' but do not explain what it is. GPT-3 is numerically evaluated in terms of accuracy and scores between 38 and 74%, which the authors regard disappointing. Interestingly, the authors also tested GPT-3 on a set of simple 'synthetic' statutes with meaningless terms (rules with 2 or 3 conditions, chains with 2 or 3 rules), to test to what extent GPT-3 uses implicit knowledge. Here GPT-3 performed even worse. The issue of implicit knowledge is also discussed more generally by the authors, as well as the possibility that GPT-3 may have 'seen' the data set (which is public). The authors conclude that their experiments raise "doubts about GPT-3's ability to handle basic legal work". Here it should be noted that currently GPT-3 is not state-of-the art any more and that its successor GPT-4 generally performs much better on many tasks.

In sum, explicit testing of reasoning capability, namely, deductive rule application with chaining, with awareness that the model might apply implicit knowledge of the statutes. Testing is direct. The prompts ask to do 'Entailment/Contradiction reasoning' but do not explain what it is. Systematic quantitative evaluation and comparisons between prompting methods but no comparisons with human performance.

Guha et al. [35] present the LegalBench legal reasoning benchmark for six legal tasks corresponding to the stages of the IRAC model plus two related tasks. The datasets for the six tasks are restricted to clear cases with objectively correct answers. The authors then apply various LLMs to these tasks, where the prompts contain between zero and eight example answers and an instruction to the LLM to explain its reasoning. For all tasks, GPT-4 performed the best, with accuracies between 59.2 and 89.9%. The authors note that their experiments should be seen as providing lower bounds on performance since they see considerable scope for improvements.

In sum, explicit testing of reasoning capability, namely, IRAC with chaining. Testing is thus direct. The prompts can contain examples of the expected reasoning and give the instruction to explain the reasoning. Systematic quantitative evaluation and comparisons between LLMs in terms of accuracy but no comparisons between prompting methods or with human performance.

Trozze et al. [24] also tested ChatGPT with GPT3.5 on the task of identifying laws that are potentially being violated in a brief factual scenario. The evaluation was in terms of the laws that were actually mentioned in the case. The prompt asked ChatGPT to apply the IRAC method, but it only mentioned IRAC and did not explain it. Moreover, its application was not explicitly tested. Instead, the quantitative measures precision (0.658), recall (0.252) and F1 (0.324) were calculated. The authors concluded from these scores that ChatGPT's performance was overall poor.

In sum, the prompt asks to apply the IRAC method but no explicit testing of whether it was applied. Testing is indirect, with as proxy how often ChatGPT mentions a law also mentioned in the case. Systematic quantitative evaluation but no comparison with human performance.

Servantez et al. [36] propose an IRAC-inspired prompting method called 'Chain of logic'. Each prompt contains an example of a rule, a fact pattern and an issue, the rule's decomposition in elements (the conditions and the conclusion) and a formalisation of the rule in proposition logic. Then the example answers each rule element separately, gives the logical expression for

the conditions yielded by the answer, and resolves it to give the final answer. Thus the model should in one shot learn to apply this IRAC-style process from the example. The authors apply this to several rule-based tasks from the LegalBench legal reasoning benchmark [35]. They apply five large language-models including GPT-4 and compare the accuracy of their prompting method to several zero- or few-shot prompting methods. Their method outperforms all other methods for all LLMs, although not by wide margins. With GPT-4 they obtain 92.3% accuracy, while the worst-performing method scores 86.3%. The authors conclude that, compared to the literature, their method is the only few-shot method that consistently outperforms zero-shot prompting. Limitations of this study are that the rules in LegalBench are simpler than in reality and that the method only seems to work for single-step rule-application.

In sum, explicit testing of reasoning capability, namely, IRAC-style deductive rule application without chaining. Testing is direct. The prompts give a detailed example of the expected reasoning. Systematic quantitative evaluation and comparisons between prompting methods in terms of accuracy but no comparisons with human performance.

5. Discussion

In this section we discuss what can be learned from the preceding overview. Here it should be taken into account that studies that are only published on ArXiv are presumably not peer reviewed.

The studies involving exams and document-generation tasks do not explicitly test some reasoning capability, which makes it hard to draw firm conclusions from them on such capabilities, since they do not distinguish between the possession of legal knowledge and the ability to apply it. The other studies do explicitly test on reasoning capabilities and test some form of deductive reasoning with legal rules, often structured in terms of the IRAC model. Some studies do not explain the reasoning method they ask for while other studies explain them with examples. A commonplace in both legal philosophy and AI & law is that deductive rule application is far too simplistic as a full model of legal reasoning. The exam and document-generation studies could implicitly test full-fledged argumentation capabilities, including the use of case- or value-based reasoning and the consideration of conflicting arguments. However, whether they do is hard to tell from the publications. This is a point on which computational models of legal argument could be useful, namely, as standards for the argumentative outputs of legal generative AI.

Most studies that make comparisons do so between several prompting methods or several LLMs. Two studies compare between AI and human performance, namely, [28] and [27], which both conclude that the model can pass American bar or law school exams and thus imply that the models can take these exams at the level of human law trainees or law-school students. However, passing such exams is only a rough proxy for having legal reasoning and argumentation abilities. Whether the various reported scores on rule application tasks are positive is hard to tell. In any case, knowledge-based legal AI would score perfectly on formalised versions of these tasks, while they naturally allow for two further forms of evaluation besides experimental evaluation of outputs: evaluation of the explicitly represented knowledge and of the explicitly programmed reasoning model. Therefore, given that so much legal knowledge is explicitly available, I believe that symbolic AI & law applications can still be practically useful, either

stand-alone, or combined with generative AI as 'conversational interfaces' between the human users' human natural language and the system's formal language.

Some studies include reasoning instructions of varying levels of detail in the prompt and/or verify to what extent the model's output obeys these instructions. A general trend in the results is that such prompting methods improve performance but not consistently. Moreover, there are some methodological pitfalls here. The first is memorisation. Questions (for instance, bar exam questions) may be in the training data, so the model may have seen them before, or the model may in other ways have applied 'shortcuts' included in its statistical language model. Some of the discussed studies show awareness of these issues [34, 24].

Next, even if a model structures output according to some reasoning method, it may be that the model has not followed the method. Striking examples are reported by [37], who found that GPT-3.5 when used with CoT prompting does not always behave according to the reason it says it applied. A simple example is with multiple-choice questions with two options A and B. When GPT-3.5 is only shown examples with A as the correct answer, it then tends to prefers answer A and gives a reason for A even if B is the correct answer. Thus the reason GPT-3.5 gives for its answer is not the reason it applied. More worrying examples involve racial and gender biases.

It might be argued that in legal applications this is not a serious problem since in the law all that matters is the justification as it is given, since that is by which the parties, appeal courts and the general public can assess the quality and acceptability of a decision. In philosophical terms, it is not the context of discovery but the context of justification that matters. However, against this it can be argued that when alternative decisions are legally acceptable, it is still undesirable that the choice for a particular decision and for which arguments and evidence to include in a decision is influenced by bias. This arguably holds the more for texts that do not contain decisions but standpoints of the parties, such as summons, complaints or briefs.

Regardless of this discussion, another way in which symbolic computational models of legal argument could be useful is in formulating reasoning instructions in the prompt. A natural idea is to formulate few-shot or CoT prompts in terms of some theory of rational reasoning or decision-making. It might be said that (legal) prompt engineering is applied (legal) philosophy.

6. Conclusion

Research on legal-reasoning capabilities of generative AI is rapidly emerging but still inconclusive as regards quality or practical usefulness. If reasoning models are made explicit, then they are (almost?) always some simple deductive form of rule application, which is generally regarded as too simplistic as a full-fledged model of legal argument. The possible roles of symbolic computational models of legal argument are threefold: as guidance for prompt engineering, as standards for evaluating outputs of legal-generative AI, and as symbolic alternatives to legal-generative AI, possibly combined with the latter as conversational interfaces. In any case, traditional symbolic AI research on legal reasoning and argumentation is not yet obsolete.

References

[1] D. Walton, Fundamentals of Critical Argumentation, Cambridge University Press, Cambridge, 2006.

[2] L. McCarty, Reflections on TAXMAN: An experiment in artificial intelligence and legal reasoning, Harvard Law Review 90 (1977) 89–116.

[3] H. Prakken, G. Sartor, Law and logic: A review from an argumentation perspective, Artificial Intelligence 227 (2015) 214–225.

[4] T. Bench-Capon, HYPO's legacy: introduction to the virtual special issue, Artificial Intelligence and Law 25 (2017) 205–250.

[5] H. Prakken, Logical models of legal argumentation, in: M. Knauff, W. Spohn (Eds.), The Handbook of Rationality, MIT Press, Cambridge, MA, 2021, pp. 669–677.

[6] H. Prakken, G. Sartor, A logical analysis of burdens of proof, in: H. Kaptein, H. Prakken, B. Verheij (Eds.), Legal Evidence and Proof: Statistics, Stories, Logic, Ashgate Publishing, Farnham, 2009, pp. 223–253.

[7] M. Sergot, F. Sadri, R. Kowalski, F. Kriwaczek, P. Hammond, H. Cory, The British Nationality Act as a logic program, Communications of the ACM 29 (1986) 370–386.

[8] T. Bench-Capon, G. Robinson, T. Routen, M. Sergot, Logic programming for large scale applications in law: a formalisation of supplementary benefit legislation, in: Proceedings of the First International Conference on Artificial Intelligence and Law, ACM Press, New York, 1987, pp. 190–198.

[9] T. Gordon, The Pleadings Game: an exercise in computational dialectics, Artificial Intelligence and Law 2 (1993) 239–292.

[10] H. Prakken, G. Sartor, Argument-based extended logic programming with defeasible priorities, Journal of Applied Non-classical Logics 7 (1997) 25–75.

[11] L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, A methodology for desining systems to reason with legal cases using abstract dialectical frameworks, Artificial Intelligence and Law 24 (2016) 1–50.

[12] G. Governatori, A. Rotolo, R. Rubino, Implementing temporal defeasible logic for modeling legal reasoning, in: JSAI-isAI 2009 Workshops, LENLS, JURISIN, KCSD, LLLL, Tokyo, Japan, November 19-20, 2009, Revised Selected Papers, number 6284 in Springer Lecture Notes in AI, Springer Verlag, Berlin, 2010, pp. 45–58.

[13] E. Rissland, K. Ashley, A case-based system for trade secrets law, in: Proceedings of the First International Conference on Artificial Intelligence and Law, ACM Press, New York, 1987, pp. 60–66.

[14] K. Ashley, Modeling Legal Argument: Reasoning with Cases and Hypotheticals, MIT Press, Cambridge, MA, 1990.

[15] V. Aleven, Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment, Artificial Intelligence 150 (2003) 183–237.

[16] D. Berman, C. Hafner, Representing teleological structure in case-based legal reasoning: the missing link, in: Proceedings of the Fourth International Conference on Artificial Intelligence and Law, ACM Press, New York, 1993, pp. 50–59.

[17] M. Grabmair, Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs, in: Proceedings of the 16th International Conference

on Artificial Intelligence and Law, ACM Press, New York, 2017, pp. 89–98.
[18] L. McCarty, An implementation of Eisner v. Macomber, in: Proceedings of the Fifth International Conference on Artificial Intelligence and Law, ACM Press, New York, 1995, pp. 276–286.
[19] T. Bench-Capon, G. Sartor, A model of legal reasoning with cases incorporating theories and values, Artificial Intelligence 150 (2003) 97–143.
[20] E. Rissland, D. Skalak, CABARET: statutory interpretation in a hybrid architecture, International Journal of Man-Machine Studies 34 (1991) 839–887.
[21] T. Gordon, D. Walton, Legal reasoning with argumentation schemes, in: Proceedings of the Twelfth International Conference on Artificial Intelligence and Law, ACM Press, New York, 2009, pp. 137–146.
[22] A. Perlman, The implications of ChatGPT for legal services and society, 2022. Http://ssrn.com/abstract=4294197.
[23] K. Iu, V.-Y. Wong, ChatGPT by OpenAI: the end of litigation lawywers?, 2023. Https://ssrn.com/abstract=4339839.
[24] A. Trozze, T. Davies, B. Kleinberg, Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers?, Artificial Intelligence and Law (2024). Https://doi.org/10.1007/s10506-024-09399-6.
[25] F. Yu, L. Quartey, F. Schilder, Legal prompting: teaching a language model to think like a lawyer, 2022. ArXiv:2212.01326.
[26] J. Rabelo, R. Goebel, M.-Y. Kim, Y. Kano, M. Yoshioka, K. Satoh, Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021, The Review of Socionetwork Strategies 16 (2022) 111–133.
[27] J. Choi, K. Hickman, A. Monahan, D. Schwarcz, ChatGPT goes to law school, 2023. Https://-doi.org/10.2139/ssrn.4335905.
[28] D. Katz, M. Bommarito, S. Gao, P. Arredondo, GPT-4 passes the bar exam., 2023. Https://ssrn.com/abstract=4389233.
[29] E. Martínez, Re-evaluating GTP-4's bar exam performance, Artificial Intelligence and Law (2024). Https://doi.org/10.1007/s10506-024-09396-9.
[30] J. Nay, D. Karamardian, S. Lawsky, W. Tao, M. Bhat, R. Jain, A. T. Lee, J. Choi, J. Kasai, Large language models as tax attorneys: a case study in legal capabilities emergence, 2023. ArXiv:2306.07075.
[31] C. Jiang, X. Yang, Legal syllogism prompting: teaching large language models for legal judgment prediction, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ACM Press, New York, 2023, pp. 417–421.
[32] W. Deng, J. Pei, K. Kong, Z. Chen, F. Wei, Y. Li, Z. Ren, Z. Chen, P. Ren, Syllogistic reasoning for legal judgment analysis, in: Proceedings of the 2023 on Empirical Methods in Natural Language Processing, 2023, pp. 13997–14009.
[33] X. Kang, L. Qu, L.-K. Soon, A. Trakic, T. Zhuo, P. Emerton, G. Grant, Can ChatGPT perform reasoning using the IRAC method in analyzing legal scenarios like a lawyer?, 2023. ArXiv:2310.14880.
[34] A. Blair-Stanek, N. Holzenberger, B. van Durme, Can GPT-3 perform statutory reasoning?, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ACM Press, New York, 2023, pp. 22–31.

[35] N. Guha et al., LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models, 2023. ArXiv:2308.11462.
[36] S. Servantez, J. Barrow, K. Hammond, R. Jain, Chain of logic: rule-based reasoning with large language models, 2024. ArXiv:2402.10400.
[37] M. Turpin, J. Michael, E. Perez, S. Bowman, Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting, in: Advances in Neural Information Processing Systems, volume 36, 2024.