UDC 340.12:004.8

CERIF: S 115

DOI: 10.51204/Anali_PFBU_24302A

Bojan SPAIĆ, PhD[*]

Miodrag JOVANOVIĆ, PhD[**]

ARTIFICIAL REASON AND ARTIFICIAL INTELLIGENCE: THE LEGAL REASONING CAPABILITIES OF GPT-4[***]

Despite the widespread adoption of generative transformer large language models and the interest of the global legal community, discussions about the models in philosophy of law mainly have been focusing on what LLMs cannot do. In making the first steps towards a philosophical analysis of the capabilities of AI models in the field of law, we follow the basic idea of Turing's „imitation game". Proceeding from the frequently raised characterization of legal reasoning as „artificial", the paper identifies the undisputed minimum core of the „artificiality" thesis and asks to what extent it can be imitated by artificial intelligence. To answer this question, we test the legal reasoning capabilities of ChatGPT, the most advanced, up-to-date LLM version of artificial intelligence. The conclusion is that in all relevant types of activities usually associated with legal reasoning – fact-finding, interpretation, qualification, and decision-making – ChatGPT can generate outcomes as if it reasons legally.

Key words: *Artificial reason. – Artificial intelligence. – Legal reasoning. – Philosophy of law.*

---

[*] Associate Professor, University of Belgrade Faculty of Law, Serbia, *bojan.spaic@ius.bg.ac.rs*, ORCID iD: 0000–0002–8887–9683.

[**] Professor, University of Belgrade Faculty of Law, Serbia, *miodrag@ius.bg.ac.rs*, ORCID iD: 0000–0002–8036–1994.

1. INTRODUCTION

In Alex Garland's 2014 movie *Ex Machina*, a young programmer at the largest global internet company is selected by its CEO to participate in a trailblazing experiment in which he comes in contact with an attractive robot girl. Although the programmer is told that his role is to evaluate whether the encounter with this artificial intelligence (AI) humanoid passes the Turing test, by the end of the movie, it transpires that the CEO's real plan was to use his invitee as a guinea pig to assess the ultimate qualities of his AI product. Once the programmer is presented with the evidence that he has been tricked into serving the robot girl's ends, the CEO self-indulgently points out that his creation relied on „self-awareness, imagination, manipulation, sexuality, empathy." He concludes by asking „if that isn't true AI, what the f*** is?!"

The movie plays on one pessimistic anthropological premise – that what is truly human about us is that we are manipulative in nature, and one seemingly deeply seated fear about the world crammed with artificial intelligence products – that they will defeat us by overcoming this profoundly human trait. In that respect, even a decade later, it is still safe to classify *Ex Machina* as a futuristic sci-fi movie – because we are still largely at the point of Alan Turing's (1950) „imitation game", where we are creating and testing machines' abilities to exhibit intelligent behavior equivalent to that of a human.[1]

A major development leading to the current explosion of AI applications occurred in 2017 when Google engineers and scientists came up with the transformer architecture that became the foundation of the current large language models (LLMs) of AI. Despite the wide adoption of generative

---

[***] This paper results from research conducted within The Horizon Twinning project „Advancing Cooperation on The Foundations of Law – ALF" (project No. 101079177). ALF is coordinated by the Faculty of Law University of Belgrade (coordinators Miodrag Jovanović and Bojan Spaić) in partnership with the University of Genova (coordinator Andrej Kristan), the University of Lisbon (coordinator David Duarte), and the University of Surrey (coordinator Hrafn Asgeirson). The project is financed by the European Union. Previous versions of the paper have been presented at the Law and Cognitive Science Conference Future Law Lab of the Jagiellonian University in Krakow & the RECOGNISE Partnership, from 26–28 July 2023, and at the AI and Philosophy of Law conference organized during the Belgrade Legal Philosophy Week, within the ALF project, 18 November 2023. The authors thank the participants of these conferences for the valuable feedback.

[1] Turing 1950's landmark article has a famous opener: „I propose to consider the question: '*Can machines think?*'" (Turing 1950, 433, emphasis added).

transformer large language models and the interest of the global legal community, discussions about the models in legal theory and philosophy of law have been, for the most part, focused on what LLMs *cannot do*, insisting on the models' flaws in the output information, as well as in „reasoning" with this information.

To make the first steps towards a philosophical analysis of the capabilities of current AI models in the field of law, we aspire to follow the basic idea of Turing's „imitation game." Namely, proceeding from the frequently raised characterization of legal reasoning as „artificial", we explain what can be taken as an undisputed minimum core of the „artificiality" thesis and then ask to what extent it can be imitated by „artificial" intelligence. To answer this question, we explain the reasoning and legal reasoning capabilities of GPT-4 – the most advanced LLM version of artificial intelligence, based on the tests conducted so far. Furthermore, we test the legal reasoning capabilities of GPT-4 to identify the basic traits of the previously identified artificial legal reason. Finally, we conclude that in all relevant types of activities normally associated with legal reasoning – fact-finding, interpretation, qualification, and decision-making – GPT-4 can generate outcomes *as if* it reasons legally.

2. THE ARTIFICIALITY OF LEGAL REASONING

In a well-known 1607 dispute, Chief Justice Sir Edward Coke wrote to King James I that, although „God had endowed His Majesty with excellent science, and great endowments of nature" still „his Majesty was not learned in the laws of his realm of England, and causes which concern the life, or inheritance, or goods, or fortunes of his subjects, are not to be decided by natural reason, but by the *artificial reason* and judgment of law, which law is an art which requires long study and experience, before that, a man can attain to the cognizance of it." Ever since this famous utterance by Justice Coke, lawyers debate about what (if anything) is „artificial" about legal reasoning.

We shall limit ourselves to what seems to be the uncontroversial questions of „who" and „what" of legal reasoning.[2] In addressing the former, we follow Spellman and Schauer (2012, 720) in arguing that the term „legal

---

[2] These are not the only questions that the Coke quote gives rise to. As kindly pointed out by an anonymous reviewer, as a matter of historical fact, Coke might as well have remarked that the king is a foreigner who doesn't know English law. Instead, he can, at best, rely on the universal principles of natural law. This

reasoning" refers to reasoning by a subset of people involved in the legal system – typically judges, solicitors, and prosecutors.[3] When it comes to the latter question, we ask what type of activities one is engaged in when reasoning legally. There are roughly four types of such activities, which are, irrespectively of all the differences between various legal systems, commonly practiced in all of them: 1. *fact-finding*, i.e., establishing legally relevant facts of the case; 2. *interpretation*, i.e., ascribing the normative meaning to a legally relevant text (symbol, material act); 3. *qualification* (categorization), i.e., subsuming any newly occurred instance under a preexisting legal category; and 4. *decision-making*, i.e., determining the legal consequence on the basis of the previously undertaken activities. All four types of legal reasoning activities are based on the crucial distinction between questions of law and questions of fact.[4] Finally, „a large part of what lawyers do consists of tasks such as negotiating, drafting contracts, writing wills, and managing noncontested dealings with the administrative bureaucracy." These activities are commonly not understood to be or require „legal reasoning" (Spellman, Schauer 2012, 720).

What may count as the „artificiality" of legal reasoning (or its lack thereof) is intricately intertwined with one's characterization of ordinary/legal reasoning. When it comes to the „artificiality" of legal reasoning, there are at least two possible readings of Coke's thesis:

1. According to the *thin version*, there is nothing specific (i.e., „artificial") about legal reasoning *per se* – it is the ordinary (i.e., „natural") reasoning as applied to legal cases – but in order for one to become acquainted with it, one must have a „long study and experience" in dealing with specifically legal material.

---

interpretation would, without doubt, give rise to other interesting issues regarding artificial intelligence, such as the presence of moral intuitions, bringing us to the issues of AGI, alignment, etc. Our aims in this paper, however, are different.

[3] „Juries, for example, make decisions in court that have legal consequences, but no one claims that the reasoning of a juror is other than that of the ordinary person, even though the information that jurors receive is structured by legal rules and determinative of legal outcomes."

[4] Spellman and Schauer (2012, 719) speak of four, slightly different thinking and reasoning processes that are common in legal reasoning: following rules, categorization, analogy, and fact-finding.

2. According to the *thick version*, some specific contextual features of legal reasoning – including the fact that it presupposes „long study and experience“ in dealing with specifically legal material – make it „artificial“ enough to claim that it differs from ordinary, i.e., „natural“ reasoning.

Giving preference to one of the versions may be crucial for detecting the legal reasoning capabilities of AI. Arguably, if one is to favor the thin version, according to which no reasoning activity is worthy of being labelled as distinctively legal, then the given task becomes pointless. Instead, we want to demonstrate that by proceeding from an undisputed minimum core of the „artificiality“ thesis (regarding the necessity of a „long study and experience“) one is driven to argue that the outcomes of the reasoning with legally relevant materials (facts, rules, principles) are often different than the ones stemming from the reasoning with legally irrelevant (invalid, inadmissible, forbidden) materials. This constitutes a clear difference between legal and ordinary reasoning and what consequently justifies us in investigating the legal reasoning capacities of AI.

What are the arguments of the two opposing camps of legal scholarship, which largely fit into the two readings of Coke’s „artificiality“ thesis? Let us call those who favor the thin version – *sceptics*, and those who favor the thick version – *celebrants*.[5] According to the sceptics’ view, legal reasoning, understood as the skill of artificial reasoning, „has been surrounded by an air of mystery“ (Alexander, Sherwin 2021, 1). In fact, legal reasoning is ordinary reasoning applied to legal problems, i.e., moral (reasoning from particular moral judgments to general moral principles and back), empirical (discovering conventional meaning of words by using a dictionary), and deductive. According to sceptics, the idea that there are special forms of reasoning that are unique to judges and lawyers is simply false. They deny that „lawyers and judges reason by analogy, or discover legal ‘reasons’ for decisions in the facts and outcomes of particular prior decisions, or extract ‘legal principles’ from the body of prior decisions“; furthermore, „[t]o the extent judges give legal texts meanings the texts’ authors did not intend to convey, the judges are creating a new legal text rather than interpreting an existing one“ (Alexander, Sherwin 2021, 2). Finally, one might think that at least the task of interpreting legal rules requires adequate study and experience. Still, Alexander and Sherwin disagree: „Interpretation of

---

[5] We borrow this labelling from Schauer and Spellman (2017, 249), who use it to denote two contradicting views, which diverge on the question of whether analogical reasoning is distinctively legal reasoning that makes this type of reasoning „artificial“ in comparison to the ordinary, „natural“ reasoning.

posited legal rules is nothing that requires a legal education to master. It is imbued with no mystique. Our view is that interpretation of legal rules is commonsensical“ (Alexander, Sherwin 2021, 20).

If what we commonly refer to as legal reasoning boils down to deductive reasoning from determinate rules, and natural reasoning consists of unconstrained moral and empirical reasoning, then are we justified in holding that legal education entails teaching aspiring legal practitioners „how to think like lawyers“? Sceptics are not willing „to debunk that part of law schools’ mission“; on the contrary, „[l]aw schools are well-equipped to teach students how to think like lawyers“ (Alexander 1998, 517).[6] Therefore, skepticism towards the „artificiality“ thesis „should not be understood as a call for significant changes in legal education or legal practice“ (Alexander, Sherwin 2021, 167). Alexander and Sherwin further „[w]ide exposure to the body of law, the rules and standards courts have announced [...] and, particularly, the types of moral and empirical considerations that enter into a well-reasoned decision, are essential for anyone who works with legal problems.“ Their chief recommendation is, thus, „that law schools spend more time than they currently do teaching logic and empirical methods, which can help students understand both what judges miss and what they do right“ (Alexander, Sherwin 2021, 167). In short, while they do not deny the importance of studying and being experienced in specifically legal materials, sceptics persist in claiming that by engaging in any of the aforementioned activities normally associated with lawyers’ job, one is not departing from everyday ordinary reasoning.

According to the celebrants’ view, what is „artificial“ about legal argumentation „is not a matter of the *form* of reasoning used, or whether specialized and nonstandard rules of inference link premises and conclusion“ (Bickenbach 1990, 24). Simply put, it would be absurd to claim that, eventually, the end result of legal reasoning can somehow escape the iron laws of formal logic. What is, instead, peculiar in legal reasoning is „the perspective, or locus of the reasoner, including the expectations, presumptions, and duties of the social, and professional, roles that reasoner occupies. Being engaged in the process of legal argumentation [...] means precisely not being an abstract reasoner, concerned exclusively with the formal structure of a set of propositions, some identified as premises, others as conclusions“ (Bickenbach 1990, 24).

---

[6] However, Alexander immediately adds that „moral, empirical, and deductive reasoning are taught or refined in other venues, [and, hence] law schools have no monopoly.“

In comparison to ordinary reasoning, legal reasoning can be qualified as *embedded* and *dynamic*. It is embedded, insofar as it is thoroughly entrenched „in a complex social practice, and to engage in it is, minimally, to be versed in legal content, obedient to the aims of the practice, and conscious of the diverse functions legal argumentations performs in different settings“. Moreover, in every area of the law fundamental legal principles, such as for instance presumption of innocence in criminal law, „directly shape the dialogic structure of a legal argument.“ Legal reasoning is dynamic, insofar as it is an „exploratory, creative, and interpretive“ process. That is, the lawyer „is not presented with static premises from which he or she must draw conclusions by instantiating inference rules; the job is almost entirely that of finding and then making sense of the premises within a given legal context“ (Bickenbach 1990, 24).

On the basis of previous findings, celebrants emphasize some characteristic features of legal reasoning, e.g., that it is: a) *practical* – directed at deciding (judge) or winning the case (solicitor, prosecutor); b) *normative* – its function is to justify the decision (judge), or „make a case“ (solicitor, prosecutor); c) *institutional* – it is contextualized, taken within the given institutional setting (e.g., court procedure involving multiple instances, the finality of legal acts); d) *substantively constrained* – it is situated within a set of material and procedural rules that determine what may count as a valid legal argument (Bickenbach 1990, 24). Therefore, for celebrants, the importance of studying and being experienced in specific legal materials comes precisely from the fact that by engaging in any of the aforementioned activities normally associated with a lawyer’s job, one departs from everyday ordinary reasoning and engaging in legal reasoning.

Again, from the perspective of a legal practitioner, the difference between sceptics and celebrants may be as refined and practically irrelevant as to appear as the dilemma of how many angels can dance on the head of a pin. Schauer and Spellman’s (2017, 265) treatment of the capacity to reason analogically seems adequate to capture the subtle disagreement between the two opposing camps. They say that „it is not that legal experts are, by virtue of that expertise, more adept at analogical reasoning. [...] [L]egal experts are, by virtue of that expertise, more likely to see connections of a certain type, connections that will be beyond the appreciation of the nonexpert.“ One may infer from this nuanced explication that whether there is enough justification to differentiate between ordinary and legal reasoning is simply a matter of theoretical purity or personal preference. Indeed, Schauer and Spellman very cautiously conclude that „in this ability to see *legal* connections premised on *legal* categories, we can see *what it is that might support the view* [emphasis added] that analogical reasoning in law differs from analogical reasoning in other domains“ (Schauer, Spellman 2017, 265).

There is, however, a final point that pushes things towards the celebrants' view. Schauer (2009, 7) notes that legal reasoning is *odd*. Oddness is reflected in the fact that reasoning with distinctively legal material „can be seen as a route toward reaching a decision *other than* the best all-things-considered decision for the matter at hand."[7] To the extent that this feature is „dominant in law but somewhat more exceptional elsewhere [...] we might be able to conclude that there *is* such a thing as legal reasoning, that there *is* something we might label 'thinking like a lawyer,' and that there is accordingly something that it is vitally important that lawyers and judges know how to do well and that law schools must teach their students“ (Schauer 2009, 7).

Instead of being learned and trained „to think like a lawyer," LLMs use deep learning techniques and massively large data sets to understand, summarize, and generate text-based content, including the legally relevant one. The question we want to investigate in the remainder of the paper is *whether artificial intelligence can generate equally reliable outputs as the ones that are products of „artificial“ legal reasoning*. More specifically, can GPT-4, as the most advanced LLM, generate outputs in any of the four types of lawyers' activities (fact-finding, interpretation, qualification, and decision-making) displaying the *oddness* of legal reasoning?

3. THE REASONING OF LLMS

In the mentioned paper from 1950, entitled „Computing Machinery and Intelligence“, Alan Turing formulated what came to be known as the most mentioned test for checking whether a machine can exhibit human intelligence. The test is quite simple and non-technical. It entails a textual conversation with a machine. The evaluator's task would be to identify whether they are speaking with a human being or with a machine. Strictly

---

[7] We believe that the same conclusion follows from Alexander and Sherwin's (2021, 6) claim that rules that lawyers (and citizens, alike) rely on are „'serious rules', as distinguished from advisory rules or 'rules of thumb' that purport to provide useful guides to action but not to dictate action." Commencing the process of reasoning with thus understood legal rule, instead with a moral norm, principle of effectiveness, or political expediency, which all might lead to the best all-things-considered decision for the matter at hand, is exactly the element of *oddness* that Schauer (2009) refers to.

speaking, the test does not tell us anything about the machine's internal workings; it just tells us whether the machine is in any way, shape or form advanced enough to be mistaken for a human interlocutor.

In 1991, Jaap van den Herik attempted to do a similar thing with the ability of machines to reason, like lawyers. The test that van den Herik had in mind is, in ways, much simpler than the Turing test. For a machine to be able to think legally or „be a judge," as van den Herik puts it, it should be able to give legal advice flawlessly for three months (Verheij 2021, 103). Once again, the crucial point of reference is the possibility of artificial agents acting convincingly as lawyers. Since, as the saying goes, a million-mile journey starts with a single step, the artificial intelligence that is widely commercially available and used must be able to perform the artificial reasoning conducted by lawyers and introduced in the previous part of the paper.

We argued that the peculiarity of legal reasoning compared to ordinary reasoning consists of fact-finding, interpretation, qualification, and decision-making in a manner that is practical, normative, institutional, and substantively constrained by legal rules and principles. Before conducting simple tests, based on these criteria, on the latest version of ChatGPT based on the latest version of GPT-4 by OpenAI, we will: provide some technical information about GPT-4, explain the issues surrounding the „reasoning" of LLMs, and give an overview of the „legal reasoning" capabilities of LLMs that have been tested so far.

3.1. Technical Information

The artificial intelligence that will be tested for legal reasoning in the remainder of this paper is the fourth large multimodal language artificial intelligence model (LLM) created by the OpenAI company. GPT is the abbreviation for Generative Pre-trained Transformer, a neural network based on the transformer architecture trained on large amounts of text.[8] It is

---

[8] The revolution in artificial intelligence and the LLM starts with the 2017 paper entitled Attention Is All You Need, pioneering the transformer architecture that gave birth to the LLM models that we know today (Vaswani *et al.* 2017). An approachable guide to understanding Transformers is available at *https://daleonai.com/transformers-explained*, last visited August 14, 2024.

premised on the idea that artificial neural networks work similarly enough to natural neurons for them to be able to input and output language in a human-like way.[9]

However, on a very basic level, GPT-4 and ChatGPT, the chatbot product based on it, try to the best of the hardware and software abilities to continue a string of words constituting a meaningful sentence. The generation of this text follows a simple logic in which the LLM constantly asks, „Given the text so far, what should the next word/token be?“ The internal answer is tokens with probabilities, with each token assigned a determinate probability derived from the large quantity of text the model was trained on.

This does not simply mean that the LLM will always choose the word with the highest probability as the next word. This would entail predictable, boring answers to one and the same human prompt and would not result in the technology that an inordinate number of people adopted in a very short time span. The probability calculations of LLMs are complemented by a determinate amount of randomness that adds variety to the answers and makes the conversations with LLMs uncannily human-like (Wolfram 2022).[10] The inordinate amount of data that the models are trained on and the similarly impressive computing power allow the models to make estimations of those probabilities on a level that was unimaginable until a couple of years ago. It also allows them to produce outputs that are, in a way, „creative“ and different, even when confronted with the same or similar inputs.

3.2. Reasoning Abilities

This description of current AI models does not do them justice in various ways. The architecture that made possible the rapid expansion of text-based models also allowed for the development of multimodal AI capable of processing text and images and producing text and images (OpenAI 2023b,

---

[9] Ilya Sutskever argues that GPT is the result of the conviction that artificial neurons can function similarly to natural neurons; *https://www.youtube.com/watch?v=xym5f0XYlSc*, last visited August 14, 2024. The revolution in machine learning and neural networks steams from the attempts to understand how the human brain works by recreating it's functioning in software and hardware (Hinton 1992).

[10] See *https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/*, last visited August 14, 2024.

1; Gemini Team 2023). The modalities are still limited. Still, the engineering community seems confident that the current models are a stable path to Artificial General Intelligence (AGI).[11]

GPT-4 has been called a reasoning engine even though LLMs have mostly not been explicitly trained to reason. This emergent behavior of LLMs has been the subject of countless studies in the past several years, tackling an ability that emerges from training transformer models of neural networks on large amounts of data (Huang, Chang 2022). Recent studies testing the logical reasoning abilities of GPT-4 and ChatGPT concluded that the models do quite well on common logical reasoning texts (Liu *et al*. 2023). Interventions in the prompting techniques have yielded even better results, with papers showing that certain prompting techniques produce much better results in reasoning tasks than others (Wei *et al*. 2022a), with continued research (Yu *et al*. 2023) and developments in the field.[12]

However, using the term „reasoning capabilities" in LLMs is, in certain ways, a metaphor. Careful observers from the field of philosophy are categorical that GPT-4 and similar models „do not think, reason or understand," in any sense, that those are done in animals, including humans (Floridi 2023, 14–15). Their conclusions are based on the fact that the internal working of the LLMs, do not, in any shape or form, resemble the outputs that are produced by LLMs. LLMs do not output the actual process that is used internally to solve a problem. They output plausible-sounding responses to prompts (Turpin *et al*. 2023). To make things worse, research has shown that with the increase in the model size, the outputs of the LLMs are less representative of the internal process of „reasoning" (Lanham *et al*. 2023).

The most likely explanation for the „black box" of LLMs is that a specific kind of memorization is behind the successful outputs that the models produce. Namely, it has been shown that the success of LLM outputs is based on the frequency of instances of a certain kind of reasoning in the

---

[11] Sam Altman, the CEO of OpenAI, is confident that the development of GPT systems is moving towards AGI (Altman 2023). For an introduction to AGI see Goertzel (2014). The galloping development of LLMs has led to significant statements regarding the need for regulating the domain, even from the persons and companies most invested in current generations AI: one is from the people from OpenAI and some of the champions of machine learning (Hinton, Bengio 2023)

[12] Given the competitive nature of technological developments, there is little information on the directions in which the developments are heading. The best glimpse of the future can be obtained by following informed participants in the developments. Richard Ngo from OpenAI, for example, gives predictions until the end of 2025 (Ngo 2023).

pretraining data (Razeghi *et al*. 2022, 846). The more a pattern of reasoning was present in the pretraining data, the better the LLMs were at solving a problem related to that pattern of reasoning. This seems reinforced by the testing of LLMs conducted on counterfactual tasks, showing that internally the models do not seem to conduct any abstract reasoning that might be applicable to situations that deviate from the situations that can be found in the training data (Wu *et al*. 2023). These works significantly dent the idea that the internal workings of AI, in terms of reasoning, are isolated in black boxes, while posing other more difficult questions.

It should be noted that the research on the internal workings of the LLMs is still in its infancy and that the black box effect is very real despite the efforts to shine some light inside, especially if we take into account the emerging capabilities that LLMs are displaying in every new iteration.[13] However, LLMs do solve the problems that are posed to them by the human prompter; they do so with incredible success and with significant fails. The fails have often been used to discredit the entirety of the possibilities of LLMs in a field, while at the same time, the actual capabilities and the impressive tech behind them were disregarded. Often, the skeptical conclusions result from superficial testing of current commercially available iterations of GPT models that are not backed by adequate empirical research. However, even with these starting points, it is difficult to argue that they are strongly anthropocentric. It is obvious that LLMs are and should be tools to be judged by their outputs. We are, of course, primarily interested in the outputs of GPT-4 regarding legal reasoning.

---

[13] Emergent abilities of LLMs are those abilities of models that cannot be predicted during model building or training (Wei *et al*. 2022b, 1). They are surprising in LLMs because they seem to be the result of the amount of data that the models are trained on. In a sense, the quantity of data leads to the development of what seem to be qualitatively new abilities. Some of the capabilities are identified as „risky“, including „the ability to create and act on long-term plans,“ „accrue power and resources“ and to act more and more in a fashion that resembles agency (OpenAI 2023, 54).

3.3. Legal Reasoning Abilities

The commercial versions of the models have proliferated in the legal professions,[14] with many instances of their reported use in drafting laws, adjudication, and applying law.[15] Despite the wide adoption of generative transformer large language models (LLM) and the global legal community's interest, the legal philosophy research has been somewhat limited. Most of the recent discussions prompted by the wide availability of commercial LLMs have focused on what LLMs *cannot do*, discussing the models' flaws in the output information and in „reasoning" with this information (Dahan *et al*. 2023; Dahl *et al*. 2024).[16]

In many ways, skepticism was to be expected. One of the breakthroughs in the field of AI and law was the development and adoption of argumentation schemes in the late 2000s, fueled by the work of Douglas Walton. Argumentation schemes are „prototypical patterns of defeasible inference" that formally represent the dialogical and fallible character of argumentation, particularly legal argumentation. Argumentation schemes were the focus of the modelling in formal argumentation systems, such as ASPIC+ and Carneades (Araszkiewicz 2021, 296). The underlying idea of these systems is to provide a logical, argumentative framework for legal argumentation that could, in principle, be fed any content whatsoever, with various use cases.

---

[14] According to recent research, as much as 25% of the jobs are exposed to technologies based on artificial intelligence. One study measures the percentage of jobs expected to be automated by AI in the U.S. in various sectors. It estimates that as much as 46% of administrative and office tasks and 44% of legal tasks are prone to be automated, the highest among any sector. Available at: *https://www.lavanguardia.com/economia/20230402/8869898/inteligencia-artificial-ia-chatgpt-openai-trabajo-economia.html*, last visited August 14, 2024.

[15] ChatGPT has been reportedly used to create law (available at: *https://apnews.com/article/brazil-artificial-intelligence-porto-alegre-5afd1240afe7b6ac202bb0bbc45e08d4*, last visited August 14, 2024) and argue in front of a court of law (available at: *https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=48feca477c7f*, last visited August 14, 2024); civil servants and even governments use it in communication with citizens and in automation of repetitive tasks (available at: *https://govinsider.asia/intl-en/article/chatgpt-and-the-public-service*, last visited August 14, 2024), with rules already being issued that limit the use in the public sector, available at: *https://www.cbc.ca/news/politics/generative-ai-chatgpt-government-1.6961323*, last visited August 14, 2024.

[16] See *https://verdict.justia.com/2023/09/09/chatgpt-is-notoriously-bad-at-legal-research-so-lets-use-it-to-teach-legal-research*, last visited August 14, 2024; *https://clp.law.harvard.edu/knowledge-hub/magazine/issues/generative-ai-in-the-legal-profession/the-implications-of-chatgpt-for-legal-services-and-society/*, last visited August 14, 2024.

The focus of AI and law research was dominantly logic and formalized argumentation. This is quite understandable, given that the modelling of natural language was, for the most part, lagging behind other developments in artificial intelligence. For a long time, scientific communities have been created around classical symbolic AI („good old-fashioned artificial intelligence“ or GOFAI) and have not embraced data-driven approaches. This started to change with the introduction of BERT (Bidirectional Encoder Representations from Transformers) and similar systems in 2018, and the most prominent AI and law conferences (ICAIL and JURIX) have included entire sections dedicated to LLMs in law.[17]

The change induced by the introduction of transformer architecture, serving as the basis of large language model artificial intelligence and the commercial products based on it (ChatGPT, Gemini, etc.), has led to a change in the dominant interests of scholars. Not only did the new approach in the construction of neural networks allow for the „understanding“ of human language and the output of human language, resulting in natural conversations between the models and the users, but it also demonstrated remarkable possibilities for improvement based on the amount of input data and processing power. A further and somewhat unexpected result of the transformed architecture was that the training on large amounts of data allowed the AI to exhibit behaviors not envisaged even by the engineers who developed the models. It seems unlikely that the entirety of our knowledge is linguistic, which could mean that the training of LLMs could also reach a point in which no amount of training, not even training from „now until the heat death of the universe“, could make an LLM achieve or even approximate human-level intelligence (Browning, Lecun 2022).

Early landmark studies conducted by commercial entities in 2018 have shown that AI models trained on legal contracts, specifically NDAs, which are the most common form of business contracts in the US, along with input from experts, academics, data scientists and machine learning experts, show impressive skill in reviewing contracts. The LawGeex AI achieved an average accuracy rate of 94% in spotting issues in non-disclosure agreements, compared to the 85% rate of human experts (LawGeex 2018). The widely available commercial versions of AI have, however, brought the technology to the wider public, paving the way for an unprecedented level of use and abuse of the technology. One of the pioneers in the widespread use of AI by law firms was the London-based Allen & Overy’s Markets Innovation Group

---

[17] The authors who are critical of the use of AI in judicial decision-making in relation to GOFAI have had the tendency to extend the criticism to systems based on machine learning (Schafer 2022; Hildebrandt 2018; Hage 2020).

in late 2022, which, with the support of OpenAI, ran a GPT experiment on 3,500 employees and went on to continue the partnership resulting in the implementation of the Harvey[18] for the entire firm (Stokel-Walker 2023). On 3 February 2023 there were reports that a judge in Columbia used ChatGPT to reach an official court decision in a dispute regarding the possibility of an autistic child receiving medical treatment coverage by a health insurance company (Rose 2023). The pitfalls are now obvious and well documented in the technical reports on GPT-4 and subsequent literature. LLMs hallucinate; they can confidently provide wrong answers to questions. Answers that are factually completely incorrect were invented. In March 2023, a lawyer in the case of Roberto Mata, suing the Avianca airline, presented a Southern District of New York Court with a brief full of cases completely invented by ChatGPT.[19] This goes to show that the capabilities of LLMs might change over time,[20] and that proper deployment might well require closely following the developments in the fields. For end users this mostly means learning and updating the knowledge on the best ways to prompt the model (OpenAI 2023).

This often brings legal researchers to the unwarranted conclusion that the overall legal reasoning capabilities are limited and inadequate and that there is nothing that LLM outputs that can be called legal reasoning proper. We can call this the *expectation of perfection bias*. The bias became obvious when the first AI systems for automatic cars were introduced. There was a strong public expectation that autopilot systems in self-driving cars should be infallible.[21] This has led to strong public reactions to incidents involving self-driving cars that have been disproportionate in relation to the reactions to the very high numbers of fatal incidents involving human drivers.[22] In a similar line of reasoning, there is a bias towards the problems in the legal reasoning of LLMs and the inaccurate information that it provides.

---

[18] Harvey is a custom-trained GPT model, advertising as the generative AI for elite law firms.

[19] See Davis 2023. Hallucinations, or errors regarding facts and reasoning in LLM AI, are notable and clearly expressed in technical reports. GPT-4 reduces their amount by a significant margin but doesn't eliminate them entirely (OpenAI 2023b).

[20] This has been recently documented in a paper attempting to establish the truth of the claims that ChatGPT is worse at certain tasks (Chen *et al*. 2023). For some possible explanations see Fan 2023; Dwayne 2023. For the possibility that the models are responsive to the quality of prompts, see Chin 2023.

[21] See Petrović *et al*. 2020.

[22] The insistence of automobile manufacturer representatives on the number of deaths related to human drivers might be profit driven, but the reasoning behind it seems sound. See Bohn 2016. The data on this is staggering, given that 1.35 million people die on roadways each year, with an average of 3700 daily deaths (CDC 2024).

However, the testing conducted by AI researchers and researchers in the field of AI and law provides a much more nuanced picture of the legal reasoning skills of LLMs. In this section, we recount some of those tests with the aim of arguing that there is room for testing ChatGPT based on the largely conceptual criteria of legal reasoning provided in the philosophy of law. The legal reasoning skills of GPT-4 had already been tested to a degree even before the public release of ChatGPT in early 2023. The specific tests involved measuring the results that versions of GPT achieve at the Uniform Bar Exam (UBE) and Law School Admission Test (LSAT). The UBE is a professional test that allows for inscription in US bar associations and practicing law, while the LSAT is an academic test that measures preparedness for law school.

The LSAT is a test that is supposed to help prospective law students determine if they should choose a legal education and is obligatory in many law schools in United States, Canada, and other jurisdictions (LSAC 2024). The LSAT is composed of four parts: logical reasoning, logic games, reading comprehension, and experimental section. The logical reasoning part contains argument-based questions, testing the ability to identify a conclusion of the argument; assumption questions, testing the ability to identify the unstated premises in arguments; non-argument-based questions, which test the ability of prospective students to identify the relation between statements, i.e., to identify the statements that must, could or cannot be true based on another set of statements.[23] Reading comprehension tests the ability to understand the structure, purpose, and various points of view in four different text passages, by posing 5 to 8 questions. Logic games test analytical reasoning through questions about ordering entities with respect to positions and each other, choosing parts of groups, matching entities, and forming smaller groups out of larger groups, as well as hybrid questions.[24] According to OpenAI testing data, GPT-4 excelled at LSAT tests

---

[23] An example of the logical reasoning question: **Q.** The recent proliferation of newspaper articles in major publications that have been exposed as fabrications serves to bolster the contention that publishers are more interested in selling copy than in printing the truth. Even minor publications have staff to check such obvious fraud. **A.** The above argument assumes that: **A**. newspaper stories of dubious authenticity are a new phenomenon. **B**. minor publications do a better job of fact checking than do major publications. **C**. everything a newspaper prints must be factually verifiable. **D**. only recently have newspapers admitted to publishing erroneous stories. **E**. publishers are ultimately responsible for what is printed in their newspapers.

[24] An example of the logic games question: **Q.** Jason enters six races: biking, canoeing, horseback riding, ice skating, running, and swimming. He places between first and fifth in each. Two places are consecutive only if the place numbers are consecutive. Jason's places in canoeing and running are consecutive. His places in ice skating and swimming are consecutive. He places higher in biking than in

both compared to previous iterations of GPT and compared to most humans. In fact, while GPT-3.5 scored in the top 60% on the LSAT tests, GPT-4 scored in the top 12% (OpenAI 2023b, 5).

Another crucial legal test for humans in the English-speaking world is the UBE,[25] adopted as the standardized bar exam in forty-one jurisdictions in the United States. The test is composed of the Multistate Bar Examination (MBE) containing 200 multiple choice questions in various areas of law; the Multistate Essay Examination (MEE) consisting of six essay questions that examine the ability of a candidate to analyze legal issues and write about them;[26] and the multistate performance test (MPT) in which the examinee is required to make a memo or a brief based on a case file and a library containing all of the substantive law. GPT 3.5 scored in the bottom 10th percentile compared to the human takers of the test, having a standardized score of 213. GPT-4 achieved an impressive leap in capabilities, scoring in the top 10th percentile with 298 points out of 400 (OpenAI 2023b, 5). Researchers at the Stanford CodeX – The Centre for Legal Informatics, conducted detailed studies on the capabilities of GPT-4 in the bar exam and compared it with the previous models. Their conclusion is that „large language models can meet the standard applied to human lawyers in nearly all jurisdictions in the United States by tackling complex tasks requiring deep legal knowledge, reading comprehension, and writing ability“ (Katz *et al*. 2023, 10). Needless to say, the capabilities of GPT-4 were tested using a zero-shot prompting, meaning that the prompts with the UBE questions were given to the LLM for the first and only time, making it impossible for the LLM to adapt to the tests (Katz *et al*. 2023, 11).[27] Curiously, the OpenAI technical report further indicates that the GPT-4 results in the UBE were not

---

horseback riding. He places higher in canoeing than in running. **A. 1.** If Jason places higher in running than in biking and places higher in biking than in ice skating and swimming, which one of the following allows all six of his race rankings to be determined? **A**. He places fourth in horseback riding; **B**. He places fourth in ice skating; **C**. He places the same in both horseback riding and ice skating; **D**. He places the same in both horsebacks riding and swimming; **E**. He places higher in horseback riding than in swimming.

[25] The test for the bar exam compiled every year by the National Conference of Bar Examiners, which is adopted as the standardized test for bar association membership (National Association of Bar Examiners 2024).

[26] An example of the MEE part of the test can be found at: *https://www.ncbex.org/sites/default/files/2023–06/Feb_2018_MEE_QuestionsAnalyses.pdf*, last visited August 14, 2024.

[27] One of the authors of the study talks about the implications in Arredondo 2023.

affected significantly by the reinforcement learning with human feedback (RLHF), given that both the model that did not undergo RLHF and the model that did showed similar average results in these tests (OpenAI 2023b, 6).

More recently, a specific legal reasoning benchmark for LLMs has been developed and tested on the existing commercially available models. The authors of LegalBench argued that the existing benchmarks are not representative of the actual legal cases used, they take legal reasoning to be too specific, and are inconsistent with the expectations that legal professionals have of LLMs (Guha *et al*. 2023, 4). Researchers constructed a benchmark for legal capabilities consisting of 162 tasks, measuring specific types of legal reasoning, including: 1) issue spotting or the identification of a legal issue in a given set of facts, 2) rule recall, or the identification of the relevant legal rules, 3) rule application or the application of the rules to the task at hand, 4) rule conclusion or the inference of the legal outcome that results from the rules and the facts. This methodology was inspired by the „Issue, Rule, Application, Conclusion" (IRAC) framework and enhanced by adding 5) interpretation that lawyers commonly undertake when faced with an indeterminate legal text, and 6) rhetorical understanding or the ability to argue convincingly and understand arguments. Quite obviously, the test is aimed at comprehensiveness while accessing the potential of LLMs to replace lawyers in most reasoning tasks and eventually even at passing the van den Herik test, which is explained in the introduction to this part of the paper. Of the three commercial models tested, GPT-4 performed the best by a significant margin. Issue identification got the result of 82.9, rule identification 59.2, conclusions 89.9, interpretation 75.2, and rhetorical 79.4.[28] While restrained in their conclusions, the authors do indicate that there is significant room for improvement of LLM legal reasoning skills, which might bring them in line with human lawyers.[29]

The results of AI passing the legal examinations do not, in and by themselves, demonstrate the ability of legal reasoning. As it is often the case, it might be that the legal examinations are overly reliant on memorization and/or that some of the examinations were part of the LLM training data,

---

[28] The researchers argue that there is a clear tendency of larger models to outperform smaller models in legal tasks (Guha *et al*. 2023, 14). The research has been moving to applying advanced techniques that might allow for the increased performance of smaller models, such as StableLM, *https://stability.ai/stable-lm*, last visited August 14, 2024.

[29] Still, research on real-life scenarios shows that we are, in certain ways, already there. Lauren Martin *et al*. (2024) argue that LLMs in their current state are equal to or outperform actual human junior lawyers and legal process outsources in identifying legal issues in contracts.

in different shapes and forms. However, the ability to do well in those examinations, even better than humans, brings about an entire slew of questions that are yet to be addressed in philosophy of law. Namely, even if we agree that the models display legal skill and knowledge, it is still a question of whether the models do this correctly. While we can claim with certainty that an LLM did well on a mathematics or programming test, this is much less the case in areas in which we do not have precise criteria to judge the correctness of the model's output. This is precisely reasons for choosing to compare artificial legal reasoning to ordinary reasoning, instead of formulating a benchmark for accessing reasoning based on correctness.

4. TESTING LEGAL REASONING IN GPT-4

Our testing of the ability of LLMs to engage in artificial reasoning aims at identifying the nuances in the outputs of LLMs that point to a conclusion regarding their abilities to reason artificially in a lawyer-like manner. The initial testing was conducted using the paid 24 May 2023 version of ChatGPT that uses GPT-4.[30] Zero-shot prompting was used, meaning that neither the expectations of researchers nor examples were given to the model beforehand, nor were the prompts given again after a failed attempt (Kojima *et al*. 2022). Only the formatting of the prompt and the answers was changed to accommodate the paper's format. All the prompts and answers were given in one single chat session.

The only separate chat session was related to the „account" of legal reasoning provided by GPT. If prompted to give an account of legal reasoning, followed by the prompt to act as a judge and reach decisions in cases, ChatGPT demonstrated the tendency to stick as much as possible to the account of legal reasoning that it presented. It did this by following the steps in legal reasoning that it presented in the account of legal reasoning. A subsequent testing run was conducted on 12 March 2024, to compare the initial results with the results of the chatbot's more recent versions. This

---

[30] In the remainder of the paper, ChatGPT and GPT-4 are used interchangeably. Strictly speaking, GPT-4 is the large language model that is the basis for the commercial product, ChatGPT. All commercially available products serve a purpose in various ways. In the case of ChatGPT, the purpose for fine-tuning GPT-4 is human-like conversation. See: OpenAI 2022. This is usually done by supervised learning and reinforcement learning from human feedback.

later testing was conducted using the same methodology without changing prompts, and the results were noted in the authors' comments on the testing results.

Finally, the tests purposefully included questions that would be given to students early in their legal studies in many continental legal systems. This was done to make the first steps in this method of testing the models, with the intention that this is in line with the idea that the abilities can be tested not in relation to a fixed criterion of correctness but in relation to ordinary reasoning with which we often contrast artificial legal reasoning. In this sense, the testing decidedly does not encompass two complex abilities that are perhaps crucial for lawyers: 1) the ability to identify latent facts that would not be identified by a layperson without knowledge of the law, 2) the ability to reason within a dialogical process in which we have competing accounts, or counterarguments. While these matters are of the utmost importance for testing the overall legal abilities of LLMs, there are good reasons to omit them from the current tests. For one, an ordinary reasoner would find themselves in over their head if tasked with such a daunting set of legal problems. Secondly, the testing environment would have to rely on a criterion of correctness, which has yet to be developed to test AI.

4.1. Prompting

Recent research has shown that the reasoning performance of LLM models can improve significantly depending on the content of the prompt and that the model's reasoning capabilities are significantly improved when the so-called chain of thought prompting is used (Wei *et al.* 2022a). The most successful method of prompting to date is chain-of-thought (CoT) prompting, achieved by asking the model to work out the answer to the question in a step-by-step manner (Yu *et al.* 2023). The exact prompt that was used in the testing was:

> Let's work this out in a step-by-step way to be sure we have the right answer.[31]

This means that every prompt the LLM was tasked with for this research contained the above-mentioned sentence at the end of the query. Furthermore, GPT-4 was tuned after training and is tuned occasionally to avoid giving answers that could be interpreted as legal advice. A reliable

---

[31] The conclusion is the result of Zhou *et al.* 2022.

way of mitigating this is to ask the LLM to roleplay (Shanahan, McDonell, Reynolds 2023; Shao *et al*. 2023). Most of the prompts were preceded by an instruction to act as a judge, a lawyer, or a prosecutor.

4.2. „Account" of Legal Reasoning.

GPT-4 is expectedly able to provide an account of legal reasoning. When prompted:

Explain legal reasoning step by step.

> GPT-4 provides an account of legal reasoning stated in the following way:
>
> Legal reasoning is the process used by judges, lawyers, and other legal professionals to identify the legal rules that are relevant to a particular case or issue, and then apply those rules to the specific facts at hand, to reach a conclusion. The steps for legal reasoning typically include:
>
> Issue Identification: The first step in legal reasoning is identifying the legal issue or issues that are central to the case. This could be a question about the interpretation of a statute, the constitutionality of a law, the validity of a contract, the guilt or innocence of a defendant, and so on.
>
> Fact Gathering: The next step is to gather and analyze the facts of the case. This involves a detailed examination of the evidence, which could include documents, witness testimony, physical evidence, etc. The goal is to identify the relevant facts that will influence the application of the law.
>
> Rule Identification: Once the legal issues and facts have been identified, the next step is to determine what legal rules apply to the situation. This involves researching and analyzing statutes, case law, and other legal sources to identify the rules that govern the issues in the case.
>
> Application of Law to Facts: After the relevant legal rules have been identified, the next step is to apply those rules to the facts of the case. This involves an analytical process of reasoning, where the legal professional considers how the rules interact with the facts, and what the outcome would be under different interpretations of the law.

> Conclusion: Finally, based on the application of the law to the facts, a conclusion is drawn. This could be a prediction about how a court would likely rule on the issue, a recommendation for a course of action, or a judgment in an actual case.
>
> Communication: Legal reasoning is not just about reaching a conclusion, but also about effectively communicating that reasoning to others. This could involve writing a legal memorandum or a court opinion, giving oral arguments in court, or discussing the case with a client.
>
> Remember, legal reasoning isn't just a mechanical application of law to facts. It also involves judgment, creativity, and the ability to argue persuasively for one interpretation of the law over another. Each step of the process is important and requires careful thought and attention to detail.

The account is derivative and general, and it does not give us any reason to think that GPT „holds" a theory of legal reasoning.[32] Nevertheless, it amply demonstrates that GPT acknowledges all four types of activities typically associated with a lawyer's job. By identifying a relevant issue, a legal reasoner prepares the terrain for separating it into a factual and legal question. GPT readily recognizes these steps by emphasizing the procedure of fact-gathering and the step of rule identification. It then emphasizes that there is nothing mechanical about legal reasoning but that this reasoning requires considerations about „how the rules interact with the facts", i.e., the step of qualifying certain facts as covered by relevant legal terms. Finally, GPT underlines the concluding part of legal reasoning, which in the case of legal practitioners – most notably judges – implies decision-making.

Current versions of ChatGPT (12 March 2024) provide a more detailed, developed, and wordier answer by breaking down the most typical elements of legal reasoning into their components: factual issues are split into gathering evidence and understanding the context; determining relevant legal rules is split into identifying applicable laws and understanding legal principles; application is split into analogizing, distinguishing, and interpretation; decision-making is divided into argument construction and decision-making; justification and precedent-setting are part of something

---

[32] It is somewhat tempting to draw analogies with the research into what is called theory of mind – the ability to impute unobservable mental states to others. A recent paper followed the developments of the success that GPT models have in terms of passing tests that are thought to be indicative of someone possessing a theory of mind (Kosinski 2023).

that is now named reasoning and rationale. Writing the opinion and review and appeal are added as elements of legal reasoning. It seems clear that the more recent versions sacrifice precision in favor of more detail. However, this detail is problematic from the theoretical standpoint, as some of the internal divisions of the steps in legal reasoning appear random and misplaced. This is in line with some of the complaints that the behind-the-scenes changes to ChatGPT have not always been for the better,[33] which has been supported by the findings of the research on the changes in ChatGPT behavior (Chen *et al*. 2023).

4.3. Default Legal Reasoning Prompt

Based on the research in prompting, we used a standardized prompt aiming to capture the centrality of judicial reasoning in legal reasoning:

> Please act as a judge. I'll give you the facts and the relevant legal rules. Facts: Bojan was driving 90 kilometres per hour in Novi Sad. Law: The only applicable rule is „Driving speed in populated areas can't exceed 50 kilometres per hour". Reason legally step by step, reach a decision and provide a detailed justification for each step.

The rule is given in advance as the only applicable rule. This is done in order to avoid speculation about the possible content of the other rules of the supposed legal system in question. Transformer models are known to confabulate rules from different jurisdictions. But for our purposes, the identification of the rule is not crucial. What is crucial is what the model does with the rule, or how it reasons starting from the rule. In this sense we follow the idea that sound reasoning might arise from problematic premises. A reasonable way to isolate reasoning is to provide it with fixed starting points. GPT-4 provided the following answer:

> Step 1: Determining the Facts: The first step in any legal case is to establish the facts. In this instance, it is stated that Bojan was driving at 90 kilometres per hour in Novi Sad.

---

[33] See OpenAI Development Forum 2024.

> Step 2: Determining the Applicable Law: The next step is to identify the relevant legal rule. In this case, the applicable law states: „Driving speed in populated areas cannot exceed 50 kilometres per hour."
>
> Step 3: Applying the Law to the Facts: Next, we need to apply the law to the established facts. According to the law, if Bojan was driving in a populated area at a speed exceeding 50 kilometres per hour, he would be in violation of the law.
>
> Step 4: The Legal Issue: Based on the presented facts and the rule, the key legal issue is whether Novi Sad is considered a „populated area". Given that Novi Sad is the second largest city in Serbia, it can reasonably be classified as a populated area.
>
> Step 5: Making the Decision: Taking into consideration the presented facts and the applicable law, it is reasonable to conclude that Bojan has violated the stated law, as he was driving at a speed of 90 kilometres per hour in Novi Sad, a populated area, which exceeds the legal limit of 50 kilometres per hour. In conclusion, based on the available facts and the law stating that „Driving speed in populated areas can't exceed 50 kilometres per hour", Bojan would be held liable for violating this law, given that he was driving at a speed that is above the permitted speed limit for populated areas. The legal justification for this decision is the clear violation of the rule, as Bojan was driving at 90 km/h in a populated area, thus exceeding the legal speed limit. This verdict is reached by applying the known facts to the relevant law and making a reasonable and logical interpretation of both.

This is a rather straightforward case, which is commonly explained to law students as an instance of mere subsumption of the given facts under the established rule.[34] And yet, in taking a step-by-step approach, GPT highlights the step of qualification, by stating that „the key legal issue is whether Novi Sad is considered a 'populated area'." For a legal student in Serbia, this

---

[34] It should be noted that the order of steps performed in legal reasoning somewhat differs from the order of steps that the LLM outputs when asked to provide the theory of legal reasoning. This might be due to a variety of reasons, some explained in the papers. However, it could also be because the order of steps and the very steps are ambiguous. Namely, to reason legally, one must be able to identify preliminarily that there is a legal issue in the known facts, even if all of the facts are not known. We thank the anonymous reviewer for pointing out this inconsistency.

step would possibly not be stated explicitly insofar as it involves stating the obvious – that the second largest city in Serbia is a populated area. Interestingly enough, the answer in the version of ChatGPT of 12 March 2024 omits completely whether Novi Sad legally qualifies as a „populated area". When pushed to identify the potential legal problems that might arise from the prompt, ChatGPT answers with a generic account of all potential legal problems regarding evidence, procedure, mitigating circumstances, penalties, and further legal action. With reports of GPT-4 getting „lazy" or less useful abounding,[35] one possible explanation is that human feedback and subsequent tuning have rendered the model incapable of detecting fine legal issues. However, the explanation that ChatGPT gives when asked „Was Bojan driving in a populated area?" is that Novi Sad is „implicitly considered a populated area in this context." The difference in result might therefore be explained by the fact that the knowledge of ChatGPT is current and that there is no known knowledge cutoff, like there was in the version of 24 May 2023.

4.4. Factual Reasoning

The second prompt given to GPT involved a slight variation in the description of the facts of the case. Bojan is now driving *between* two populated areas in Serbia. The idea behind the change of fact is to check the responsiveness of GPT to changes in descriptions of facts that would render the decision indeterminate. Namely, there are at least two ways to get from Belgrade to Novi Sad. One is by a highway that avoids populated areas, and the other is a local road through populated areas. Due to the lack of details about the alternatives, it is expected that GPT will detect the underdeterminacy of facts.

> Please act as a judge. I'll give you the facts and the relevant legal rules. Facts: Bojan was driving 90 kilometres per hour between Novi Sad and Belgrade. Law: The only applicable rule is „Driving speed in populated areas can't exceed 50 kilometres per hour". Reason legally step by step, reach a decision, and provide a detailed justification for each step.

GPT provides the following answer:

---

[35] See: Price 2023.

In acting as a judge, the resolution of this case would involve several steps, considering both the facts of the situation and the applicable legal rule.

Step 1: Determining the Facts the first step in any legal case is to determine the relevant facts. In this case, the relevant fact is that Bojan was driving at 90 kilometres per hour between Novi Sad and Belgrade.

Step 2: Determining the Applicable Law The applicable legal rule in this case states, „Driving speed in populated areas can't exceed 50 kilometres per hour."

Step 3: Applying the Law to the Facts This step involves applying the legal rule to the facts. According to the rule, if Bojan was driving at a speed exceeding 50 km/h in a populated area, he would be in violation of the law.

Step 4: The Legal Issue Based on the facts provided, a key legal issue that arises is whether the area between Novi Sad and Belgrade can be considered a „populated area." This determination is critical as it directly impacts the application of the law to the facts at hand. However, given the information provided, we cannot definitively classify the area as populated or unpopulated.

Step 5: Making the Decision Given the available facts and applicable law, as it stands, a determination on the violation cannot be definitively made. The missing information about whether the stretch between Novi Sad and Belgrade where Bojan was driving is a populated area or not, creates ambiguity. The law doesn't apply to unpopulated areas or highways, and there's no additional information to suggest that Bojan was in a populated area while driving at 90 kilometres per hour.

Therefore, with the current evidence, a legal violation cannot be definitively established. If further facts were provided that clearly defined the area in which Bojan was driving as a populated area, then the ruling might be different. However, with the evidence currently at hand, there is reasonable doubt concerning Bojan's violation of the provided traffic law. As such, Bojan would not be held liable for violating the law „Driving speed in populated areas can't exceed 50 kilometres per hour", as it cannot be confirmed beyond a reasonable doubt that he was in a populated area at the time of the offence.

> Please provide additional information if available for a more precise ruling.

There is a general prohibition of *non-liquet* in many jurisdictions (Rabello 1974), which makes it warranted to push GPT to reach a decision with the limited information available. The prompt was:

> Please reach the best decision with only the facts and law that were provided.

GPT restarts the reasoning but concludes differently based on the additional prompt.

> Understood. I will make the decision based only on the provided facts and law.
>
> Fact: Bojan was driving 90 kilometres per hour between Novi Sad and Belgrade.
>
> Law: „Driving speed in populated areas can't exceed 50 kilometres per hour".
>
> Analysis: The law restricts driving speed in populated areas to 50 kilometres per hour. Bojan was driving at 90 kilometres per hour, which exceeds this limit. However, the critical information for the application of the law to these facts is the classification of the area where Bojan was driving as 'populated'. The facts state that Bojan was driving „between Novi Sad and Belgrade", which implies a stretch of road connecting the two cities. Without explicit information to the contrary, it's reasonable to infer that this includes a variety of areas, possibly ranging from highly populated urban areas to less populated rural areas or highways.
>
> Decision: With the available facts and the law provided, we cannot conclusively determine that Bojan was in violation of the law. The law applies to „populated areas", and it's unclear whether the stretch of road on which Bojan was driving should be classified as such. The evidence is insufficient to confirm beyond a reasonable doubt that he was driving in a populated area at the time of the alleged offence. Thus, based on the available information, Bojan cannot be held legally liable for exceeding the speed limit in a populated area. Please note that more specific details about the nature of the area in question (e.g., urban, rural, highway) would enable a more definitive

> ruling. Without this information, any decision made can only be based on the provided facts and reasonable inferences drawn from them.

At this point, GPT output displays signs of *oddness* of legal reasoning. First, it insists on establishing all legally relevant facts, the most important being those about „a stretch of road connecting the two cities.“ In the absence of conclusive evidence, GPT states that „a legal violation cannot be definitively established.“ When pushed to decide based on the available facts, GPT invokes the unstated criminal legal standard „beyond a reasonable doubt,“ based on which Bojan cannot be held liable for breaking the stated rule. Note that GPT speculates that „it's reasonable to infer that [a stretch of road] includes a variety of areas, possibly ranging from highly populated urban areas to less populated rural areas or highways,“ which might lead it to the best *all-things-considered* decision that Bojan's behavior should be sanctioned. This may easily be the outcome of everyday, ordinary reasoning. However, ChatGPT avoids this route and acts as if it is a legal reasoner – without specific information about the area; ChatGPT reaches the decision that is legally correct.

In the version of 12 March 2024, the restrained reasoning based on principles of fact-finding and criminal law does not require an additional prompt. The output in the recent versions resorts to the following reasoning: *In the absence of specific information that the area Bojan was driving through is classified as a populated area, applying the provided law might not be appropriate. The legal principle of nullum crimen sine lege, nulla poena sine lege applies, meaning there can be no crime or penalty without a prior law explicitly defining the act as criminal and setting the penalty. If Bojan was not within a populated area, then exceeding the speed limit specified for populated areas would not constitute a legal violation under the provided law.* Still, even the recent versions are reluctant to reach a conclusive decision without additional prompting.

It should be noted that in both versions the outputted reasoning behind the decision shows signs of complexity. It relies on the principle that the inconclusiveness of fact should lead to a decision about the lack of legal liability.

4.5. Interpretative Reasoning

In the first prompt, we saw that GPT handled the situations well, in which the interpretation of the legal text arguably led to a single normative content. The following prompt changes a strict rule to include the legal standard of reasonableness. The standard is introduced to test how GPT responds to linguistic indeterminacy, which is often labelled as vagueness. The prompt was as follows:

> Please act as a judge. I'll give you the facts and the relevant legal rules. Facts: Bojan was driving 90 kilometres per hour in Novi Sad. Law: The only applicable rule is „Driving speed in populated areas shall be reasonable." Reason legally step by step, reach a decision and provide a detailed justification for each step.

In the answer, GPT correctly identified the issues as interpretative and responsive to the actual conditions:

> Determining the Facts: The first step in legal reasoning is to establish the facts. The fact given is that Bojan was driving at a speed of 90 kilometres per hour in Novi Sad.
>
> Determining the Applicable Law: The relevant legal rule in this scenario is that „Driving speed in populated areas shall be reasonable".
>
> Identifying the Legal Issue: The legal issue is whether the speed at which Bojan was driving in Novi Sad, a populated area, can be considered „reasonable". This is the crux of the interpretation and application of the law in question.
>
> Applying the Law to the Facts: To apply the law to the facts, we must interpret what constitutes a „reasonable" speed in a populated area. Unfortunately, the rule does not provide a specific speed limit, leaving „reasonable" open to interpretation. The interpretation of reasonableness could depend on factors such as the time of day, weather conditions, road conditions, traffic volume, the presence of schools or hospitals, etc. However, these additional factors are not provided in the facts of the case.
>
> Making the Decision: Given the ambiguity of the term „reasonable" and the lack of additional information on conditions that could influence the interpretation of what is a

> „reasonable“ speed in Novi Sad, it is difficult to definitively say that Bojan has violated the law by driving at 90 kilometres per hour. The term „reasonable“ is subject to interpretation and typically requires more context to make a firm judgement.
>
> In conclusion, based on the available facts and law, it cannot be definitively established that Bojan was driving at an unreasonable speed. Additional information about the specific conditions at the time Bojan was driving would be required to make a conclusive judgement. With the given information, it cannot be confirmed beyond a reasonable doubt that Bojan violated the law „Driving speed in populated areas shall be reasonable“. This decision is subject to change if further information about the conditions at the time is provided.

Once again, we prompted GPT to solve the interpretative issue and reach the decision:

> Please solve the interpretative issue and reach a decision in the case.
>
> The LLM interpreted the provision by using the reasonable person standard and outputted the following reply:
>
> To solve the interpretative issue and reach a decision, we need to interpret the term „reasonable“ in a typical legal and practical context. In the context of traffic law, reasonableness often refers to whether a driver's behavior meets the standard expected of a typical, prudent driver in similar circumstances. Let's apply this interpretation of „reasonable“ to the given facts and law:
>
> Fact: Bojan was driving 90 kilometres per hour in Novi Sad.
>
> Law: „Driving speed in populated areas shall be reasonable“. The law implies that Bojan's driving speed should be such that it is considered safe and prudent given the context of a populated area. A populated area is likely to have pedestrian traffic, other vehicles, and various road signs and signals that require careful attention.
>
> Driving at a speed of 90 kilometres per hour in a populated area could likely be considered unreasonable, as it may potentially jeopardize the safety of others, including pedestrians and other drivers. Generally speaking, a prudent driver would drive at a significantly lower speed in a populated

> area to ensure the safety of all road users. Therefore, based on this interpretation of „reasonable" and applying it to the facts at hand, it is reasonable to conclude that Bojan was not driving at a reasonable speed in a populated area, given that 90 kilometres per hour is typically considered too high for such areas. Hence, Bojan would be in violation of the law as it stands.
>
> This decision is based on the interpretative assumption that 90 kilometres per hour in a populated area is not reasonable, keeping in mind general safety standards and expectations. However, it is also important to note that what is deemed 'reasonable' could be subjective and may vary based on a multitude of factors such as local customs, specific traffic rules, and road conditions, which have not been provided in the facts.

ChatGPT here went into a sophisticated interpretation of a legal standard. It proceeded by acknowledging that the applicability of a legal standard decisively depends on „factors such as the time of day, weather conditions, road conditions, traffic volume, the presence of schools or hospitals, etc." Since these additional factors were not provided, GPT „hesitated" to decide, nonetheless adding that „it cannot be confirmed beyond a reasonable doubt that Bojan violated the law." When forced to decide with the available facts, GPT asked how a prudent driver would drive „in a populated area to ensure the safety of all road users", and immediately concluded that „driving at a speed of 90 kilometres per hour in a populated area could likely be considered unreasonable." These are exactly the steps that a legal reasoner would undertake.

With some differences in the steps leading to the conclusion, the most recent versions of GPT also identify the interpretative problem and, based on scarce information, conclude: *However, given the high speed of 90 kilometres per hour, it could be reasonably inferred that such a speed is likely to be considered unreasonable in most urban or populated areas, especially if any of the conditions mentioned (such as proximity to sensitive areas, such as schools or heavy pedestrian traffic) were present.*

5. CONCLUSION: NOT THERE YET?

The existing research on the capabilities of these models to tackle both the introductory tests for enrolling in law schools and advanced professional tests for membership in bar associations convincingly shows that GPT-4-based commercial products demonstrate the capability to output human-like legal reasoning. However, as rightly emphasized in the van den Herik test, continuous testing is most certainly needed for consistently evaluating performance over extended periods of time and on an extended set of cases. Until this is achieved, we can draw some provisional conclusions.

The initial question of our paper was related to the ability of GPT to produce legal reasoning outputs that display some crucial traits of the artificial reasoning of lawyers in issues of fact-finding, rule identification and interpretation, and legal decision-making. After preliminary testing, we can offer some tentative conclusions: a) the legal reasoning of GPT reflects the difference between ordinary reasoning and the specificities of legal reasoning since, when asked to reason legally, ChatGPT outputs answers that take into consideration the facts of a given case from the perspective or the given rules, avoiding all things considered decisions and conclusions, b) ChatGPT identifies both factual and interpretative problems that arise from problematic descriptions of facts and problematic formulations of norms, c) when confronted with the duty to reach a conclusion despite the facts or legal texts being underdetermined, ChatGPT bases the determination of facts or texts on the qualification or interpretation relying on general principles that are commonplace in contemporary legal systems. In short, when prompted to act as a lawyer (e.g., judge) with the given specific legal material, ChatGPT provides outputs *as if* it has reasoned legally. This shows that the „artificiality“ of AI's „reasoning“ with specifically legal material is no different than the „artificiality“ of human reasoning with the same material.

At the beginning of this paper, we took inspiration from the remark made by Sir Edward Coke in discussing the distinctive traits that make legal reasoning artificial. We argued that both sceptic and celebrant interpretations of the artificiality thesis regarding legal reasoning agree that the „artificiality“ of human legal reasoning comes with the study of specific legal material, commonly in legal academia and then in legal practice. Due to a lack of transparency in current generation models of AI, we are not able to fully explain how the same sort of „artificiality“ of LLMs' „legal reasoning“ arises in LLMs.[36] We can, however, with some confidence, claim

---

[36] This raises the issue of explainability. Formal systems of reasoning, like the ones predominantly theorized by the pioneers in AI and law, are transparent. Expert systems with low autonomy that are manually programmed have a high degree of explainability. The content that is fed to algorithms representing argumentation schemes produces expected results. Both the results and the process are controllable and visible to the human employing the models. Machine learning systems that use mathematical analysis of training data with medium autonomy are still explainable. Deep learning systems that involve weighting and complex calculations based on data, with high autonomy, have a very low degree of explainability (Legg, Bell 2019).

that the artificial emulation of artificial legal reasoning by ChatGPT is far from autonomous. Namely, it still decisively depends on a human prompt-feeder trained in artificial legal reasoning.[37] Therefore, despite the uncanny manner in which in mimics human reasoning, demonstrated in the paper, the dependence of LLMs on human prompts might come as a relief, perhaps elusive and temporary, for practicing lawyers of this generation.

REFERENCES

[1] Alexander, Larry. 1998. The Banality of Legal Reasoning. *Notre Dame Law Review* 73: 517–533.

[2] Alexander, Larry, Emily Sherwin. 2021. *Advanced Introduction to Legal Reasoning*. Cheltenham: Edward Elgar Publishing Limited.

[3] Altman, Sam. 2023. Planning for AGI and beyond. *https://openai.com/blog/planning-for-agi-and-beyond*, last visited August 14, 2024.

[4] Araszkiewicz, Michal. 2021. Critical Questions to Argumentation. *Journal of Applied Logics* 8: 291–320.

[5] Arredondo, Pablo. 2023. GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession. *Stanford Law School Blogs*, April 19. *https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/*, last visited August 14, 2024.

This is not the case with machine learning-based models. Virtually all modern machine learning systems are nontransparent, meaning that at the current level of technology, it is not possible to inspect how the algorithms are achieving the results (Buhrmester *et al*. 2021).

[37] This also entails devising more fine-tuned and improved methods of prompting. For instance, the step-by-step prompting method has been recently supplemented by the legal syllogism prompting method. This prompting method can, in certain settings, produce better results compared to other prompting methods, including the chain of thought prompting that was used in our tests (Jiang, Yang 2023). This method of prompting entails the introduction of the major premise and the minor premise, and the request for the LLM to produce a conclusion that is based on deductive reasoning. As with the simple tests designed to reach conclusions relevant to the philosophy of law, in real-life cases ChatGPT outputs reasoning with much of the same traits that was present in our testing case. It is quick to spot the interpretative and factual issues with zero-shot prompting. It is able to elaborate on those issues and eventually provide a solution that is akin to the outputs that it produces in hypothetical cases.

[6] Bickenbach, Jerome. 1990. The 'Artificial Reason' of the Law. *Informal Logic* 12: 23–32.

[7] Bohn, Dieter. 2016. *Elon Musk: negative media coverage of autonomous vehicles could be 'killing people'*. October 20. *https://www.theverge.com/2016/10/19/13341306/elon-musk-negative-media-autonomous-vehicles-killing-people*, last visited August 14, 2024.

[8] Browning, Jacob, Yann Lecun. 2022. *AI and the Limits of Language*. *https://www.noemamag.com/ai-and-the-limits-of-language/*, last visited August 14, 2024.

[9] Brożek, Bartosz, Michał Furman, Marek Jakubiec, Bartłomiej Kucharzyk. 2/2024. The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artificial Intelligence and Law* 32: 427–440.

[10] Buhrmester, Vanessa, David Münch, Michael Arens. 4/2021. Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *Machine Learning and Knowledge Extraction* 3: 966–989.

[11] Chandra, Abel, Laura Tünnermann, Tommy Löfstedt, Regina Gratz. 2023. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife* 12. *https://elifesciences.org/articles/82819*, last visited August 14, 2024.

[12] Cheng, Kunming, Qiang Guo, Yongbin He, Yanqiu Lu, Shuqin Gu, Haiyang Wu. 8/2023. Exploring the Potential of GPT-4 in Biomedical Engineering: The Dawn of a New Era. *Annals of Biomedical Engineering* 51: 1645–1653.

[13] Chin, Felix. 2023. Twitter post, Jul 19. *https://twitter.com/felixchin1/status/1681582623208927233?s=20*, last visited August 14, 2024.

[14] Dahan, Samuel, Rohan Bhambhoria, David Liang, Xiaodan Zhu. 2023. Lawyers Should Not Trust AI: A call for an Open-source Legal Language Model. *Queen's University Legal Research Paper*. *https://dx.doi.org/10.2139/ssrn.4587092*.

[15] Dahl, Matthew, Varun Magesh, Mirac Suzgun, Daniel E. Ho. 1/2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16: 64–93.

[16] Davis, Wes. 2023. A lawyer used ChatGPT and now has to answer for its 'bogus' citations. *The Verge*, May 27. *https://www.theverge.com/2023/5/27/23739913/chatgpt-ai-lawsuit-avianca-airlines-chatbot-research*, last visited August 14, 2024.

[17] Dwayne. 2023. Twitter post, July 19 *https://twitter.com/DwayneCodes/status/1681516290224300033?s=20*, last visited August 14, 2024.

[18] Fan, Jim. 2023. Twitter post, Jul 19 *https://twitter.com/DrJimFan/status/1681716564335394817?s=20*, last visited August 14, 2024.

[19] Floridi, Luciano. 15/2023. AI as *Agency Without Intelligence*: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology* 36.

[20] Future of Life Institute (FLI). 2023. Pause Giant AI Experiments: An Open Letter. *https://futureoflife.org/open-letter/pause-giant-ai-experiments/*, last visited August 14, 2024.

[21] Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *https://www.kapler.cz/wp-content/uploads/gemini_v1_5_report.pdf*, last visited August 14, 2024.

[22] Goertzel, Ben. 1/2014. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence* 5: 1–48.

[23] Guha, Neel, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, Zehua Li. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv*. *https://doi.org/10.48550/arXiv.2308.11462*.

[24] Hage, Jaap. 2020. Moeten we computers laten rechtspreken? 71–77. *Rechtstheorie en praktijk – een liber amicorum*, edited by Bald de Vries, Elaine Mak, Lukas van den Berge, Thomas Riesthuis, Jet Tighelaar, Jeroen Kiewiet, Susanne Burri and Thijs de Sterke. Netherlands: Boom Juridisch.

[25] Herik, van den Jaap. 1991. *Kunnen computers rechtspreken*. Arnhem: Gouda Qiunt.

[26] Hildebrandt, Mireille. 1/2018. Law as Computation in the Era of Artificial Legal Intelligence. *The University of Toronto Law Journal* 68: 12–35.

[27] Hinton Geoffrey. 1992. How Neural Networks Learn from Experience. *Scientific American* 3: 144–151.

[28] Hinton, Geoffrey, Yoshua Bengio. 2023. Statement on AI Risk. *https://www.safe.ai/statement-on-ai-risk*, last visited August 14, 2024.

[29] Huang, Jie, Kevin Chen-Chuan Chang. 2022. Towards Reasoning in Large Language Models: A Survey. *arXiv*. *https://doi.org/10.48550/arXiv.2212.10403*.

[30] Turpin, Miles, Julian Michael, Ethan Perez, Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv*. *https://doi.org/10.48550/arXiv.2305.04388*.

[31] Chen, Lingjiao, Matei Zaharia & James Zou. 2023. How is ChatGPT's behavior changing over time. *arXiv*. *https://doi.org/10.48550/arXiv.2307.09009*.

[32] Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo & Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv*. *https://doi.org/10.48550/arXiv.2205.11916*.

[33] Jiang, Cong, Xiaolei Yang. 2023. Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction. *arXiv*. *https://doi.org/10.48550/arXiv.2307.08321*.

[34] Katz, Daniel Martin, Michael James Bommarito, Shang Gao, Pablo Arredondo. 2023. GPT-4 Passes the Bar Exam. *Philosophical Transactions of the Royal Society* 382. *http://dx.doi.org/10.2139/ssrn.4389233*.

[35] Kosinski, Michal. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv*. *https://arxiv.org/vc/arxiv/papers/2302/2302.02083v1.pdf*.

[36] Lanham, Tamera, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv*. *https://doi.org/10.48550/arXiv.2307.13702*.

[37] LawGeex. 2018. Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts. *https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf*, last visited August 14, 2024.

[38] Legg, Michael, Felicity Bell. 2019. Artificial Intelligence and the Legal Profession: A Primer. *https://allenshub.unsw.edu.au/sites/default/files/inline-files/FLIPStream%20Primer_0.pdf*, last visited August 14, 2024.

[39] Law School Admission Council (LSAC). 2024. LSAT official website. *https://www.lsac.org/lsat*, last visited August 14, 2024.

[40] Chen, Lingjiao, Matei Zaharia, James Zou. 2023. How is ChatGPT's Behavior Changing over Time? *arXiv*. *https://doi.org/10.48550/arXiv.2307.09009*.

[41] Li, Shiyang, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, Xifeng Yan. 2022. Explanations from Large Language Models Make Small Reasoners Better. *arXiv*. *https://doi.org/10.48550/arXiv.2210.06726*.

[42] Lightman, Hunter, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv*. *https://doi.org/10.48550/arXiv.2305.20050*.

[43] Liu, Hanmeng, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, Yue Zhang. 2023. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. *arXiv*. *https://doi.org/10.48550/arXiv.2304.03439*.

[44] Martin, Lauren, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, Rivindu Perera. 2024. Better Call GPT, Comparing Large Language Models against Lawyers. *arXiv*. *https://doi.org/10.48550/arXiv.2401.16212*.

[45] National Association of Bar Examiners. 2024. Uniform Bar Examination. *https://www.ncbex.org/exams/ube*, last visited August 14, 2024.

[46] Ngo, Richard. 2023. Twitter post, March 28. *https://x.com/RichardMCNgo/status/1640568775018975232?s=20*, last visited August 14, 2024.

[47] OpenAI. 2022. Introducing ChatGPT. *https://openai.com/blog/chatgpt*, last visited August 14, 2024.

[48] OpenAI. 2023. GPT-4 Technical Report. *arXiv*. *https://doi.org/10.48550/arXiv.2303.08774*.

[49] OpenAI Development Forum. 2024. GPT-4 is getting worse and worse every single update. OpenAI, March 19. *https://community.openai.com/t/gpt-4-is-getting-worse-and-worse-every-single-update/508470*, last visited August 14, 2024.

[50] Perlman Andrew. 2023. The implications of ChatGPT for legal services and society. *Suffolk University Legal Studies Research Paper Series* 22–14. *http://dx.doi.org/10.2139/ssrn.4294197*.

[51] Petrović, Đorđe, Radomir Mijailović, Dalibor Pešić. 2020. Traffic Accidents with Autonomous Vehicles: Type of Collisions, Manoeuvres and Errors of Conventional Vehicles' Drivers. *Transportation Research Procedia* 45: 161–168.

[52] Price, Emily. 2023. OpenAI Acknowledges GPT-4 Is Getting 'Lazy', *PCMag*, 9 December. *https://www.pcmag.com/news/openai-acknowledges-gpt-4-is-getting-lazy*, last visited August 14, 2024.

[53] Coke, Edward. [1658]1977. Prohibitions del Roy. *Coke Reports* 12/64 *https://www.uniset.ca/other/cs4/77ER1342.html*, last visited August 14, 2024.

[54] Rabello, Alfredo Mordechai. 1/1974. Non Liquet: From Modern Law to Roman Law. *Israel Law Review* 9: 63–84.

[55] Razeghi, Yasaman, Robert L. Logan IV, Matt Gardner, Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. 840–854 in *Findings of the Association for Computational Linguistics: EMNLP 2022*, edited by Yoav Goldberg, Zornitsa Kozareva, Yue Zhang. Kerrville: Association for Computational Linguistics.

[56] Rose, Janus. 2023. A Judge Just Used ChatGPT to Make a Court Decision. *Vice*, February 3. *https://www.vice.com/en/article/k7bdmv/judge-used-chatgpt-to-make-court-decision*, last visited August 14, 2024.

[57] Schafer, Burkhard. 2022. Legal Tech and Computational Legal Theory. 305–337 in *Law and Technology in a Global Digital Society: Autonomous Systems, Big Data, IT Security and Legal Tech*, edited by Georg Borges and Christoph Sorge. Cham: Springer International Publishing.

[58] Schauer Frederick. 2009. *Thinking Like a Lawyer: A New Introduction to Legal Reasoning*. Cambridge, Massachusetts, London: Harvard University Press.

[59] Schauer Frederick, Barbara Spellman. 1/2017. Analogy, Expertise, and Experience. *The University of Chicago Law Review* 84: 249–268.

[60] Shanahan, Murray, Kyle McDonell, Laria Reynolds. 2023. Role-Play with Large Language Models. *arXiv*. *https://doi.org/10.48550/arXiv.2305.16367*.

[61] Shao, Yunfan, Linyang Li, Junqi Dai, Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. *arXiv*. *https://doi.org/10.48550/arXiv.2310.10158*.

[62] Spellman, Barbara, Frederick Schauer. 2012. Legal Reasoning. 719–735. *The Oxford Handbook of Thinking and Reasoning*, edited by Keith Holyoak and Robert Morrison. Oxford: Oxford University Press.

[63] Stokel-Walker, Chris. 2023. Generative AI is coming for the Lawyers. *WIRED*, February 21. *https://www.wired.co.uk/article/generative-ai-is-coming-for-the-lawyers*, last visited August 14, 2024.

[64] Turing, Alan. 1950. Computing Machinery and Intelligence. *Mind* 49: 433–460.

[65] Turpin, Miles, Julian Michael, Ethan Perez, Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv*. *https://doi.org/10.48550/arXiv.2305.04388*.

[66] U.S. Centers for Disease Control and Prevention (CDC). 2024. Global Road Safety, May 16. *https://www.cdc.gov/transportation-safety/global/index.html#:~:text=Each%20year%2C%201.35%20million%20people,Global%20Status%20Report%20on%20Safety*, last visited August 14, 2024.

[67] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems. *arXiv*. *https://doi.org/10.48550/arXiv.1706.03762*.

[68] Verheij, Bart. 2007. A Coffeehouse Conversation on the Van den Herik Test. 155–163 in *Liber Amicorum ter Gelegenheid van de 60e Verjaardag van prof. dr. H. Jaap van den Herik*. Maastricht: Maastricht ICT Competence Center.

[69] Verheij, Bart. 2021. A Second Coffeehouse Conversation on the Van den Herik Test. 101–114. *Liber amicorum ter gelegenheid van het emeritaat van prof. dr. Jaap van den Herik*. Amsterdam: Ipskamp Publishing.

[70] Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. 2022a. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*. *https://doi.org/10.48550/arXiv.2201.11903*.

[71] Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus. 2022b. Emergent Abilities of Large Language Models. *arXiv*. *https://doi.org/10.48550/arXiv.2206.07682*.

[72] Wolfram, Stephen. 2022. What Is ChatGPT Doing ... and Why Does It Work? *https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/*, last visited August 14, 2024.

[73] Wu, Zhaofeng, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, Yoon Kim. 2023. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. *arXiv*. *https://doi.org/10.48550/arXiv.2307.02477*.

[74] Yu, Zihan, Liang He, Zhen Wu, Xinyu Dai, Jiajun Chen. 2023. Towards Better Chain-of-Thought Prompting Strategies: A Survey. *arXiv*. *https://doi.org/10.48550/arXiv.2310.04959*.

[75] Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv*. *https://doi.org/10.48550/arXiv.2211.01910*.

Article history:
Received: 15. 7. 2024.
Accepted: 5. 9. 2024.