

Defeasible Reasoning with Large Language Models – Initial Experiments and Future Directions

Grigoris Antoniou^{1,2}, Sotiris Batsakis^{3,2}

¹*L3S, Leibniz University Hanover, Germany*

²*University of Huddersfield, UK,*

³*Technical University of Crete, Greece*

Abstract

As Large Language Models gain prominence in the AI landscape, it is essential to understand their capabilities and limitations, among others in terms of reasoning. This paper is a first step towards understanding the capabilities in terms of defeasible rule-based reasoning. It presents results of initial experiments and discussed future research directions.

Keywords

Defeasible Reasoning, Large Language Models, Non-monotonic Reasoning

1. Introduction

Large Language Models (LLMs) have risen in prominence recently due to their exceptional performance in various language related tasks since they are the underlying technology behind chat bots such as ChatGPT¹. Large Language models such as LaMDA [1] and GPT [2] are based on training deep neural networks with billions of parameters over huge lexical datasets and often employing human judgment in a semi-supervised (e.g., reinforcement learning) training setting [3, 4]. The exceptional -human level- performance of LLMs in various tasks has led to a widespread discussion about the potential benefits and dangers of such technologies in various areas and human society in general including petitions to pause research on more capable LLMs [5].

Despite their impressive performance Large language Models, including their ability to demonstrate an emerging intelligent behaviour and reasoning capabilities, leading to the point of considering them forerunners of Artificial General Intelligence[6] several issues related to LLMs have been identified, such the energy cost of training LLMs[7, 8], difficulty to control their behaviour[9], ensure conformity with stakeholders requirements and norms and interpret their functionality [10]. The interpretability of LLMs is crucial issue since neural network based LLMs appear to be ‘black boxes’, in contrast to logic based systems, and although various attempts exist to deal with this problem, including the use of LLMs to interpret LLMs [11], this is still an unresolved issue. In addition, since LLMs are based on vast amount of raw text

RuleML+RR’23: 17th International Rule Challenge and 7th Doctoral Consortium, September 18–20, 2023, Oslo, Norway

✉ G.Antoniou@hud.ac.uk (G. Antoniou); sbatsakis@tuc.gr (S. Batsakis)

🆔 0000-0003-3673-6602 (G. Antoniou); 0000-0001-6023-2311 (S. Batsakis)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

¹Available at: <https://chat.openai.com/>

they tend to replicate their input rather than apply robust reasoning [12]. Based on raw text instead of structured knowledge bases integrating machine readable semantics, contributes to the difficulty to achieve efficient reasoning and this is an issue examined in various works such as [13] and surveyed in [14]. Various attempts to integrate Knowledge graphs (KGs) to LLMs have been proposed [15, 16] as a solution to the last issue, but recent advances in LLMs capabilities, including high performance on academic and professional exams [2], illustrated the need for an updated evaluation of the reasoning capabilities of LLMs. This updated evaluation should take into account the recent developments in the field, including the deployment of systems such as chatGPT employing the benefits of scalability [17] and the LLMs demonstrated ability to adjust to new tasks given just a small number of examples [18]. Furthermore LLMs capabilities with respect to important formalisms such as defeasible reasoning have not been examined in detail yet.

This work is an initial step towards developing a deep understanding of reasoning capabilities of LLMs with emphasis of nonmonotonic reasoning. For achieving this we conducted experiments on nonmonotonic rule-based reasoning using chatGPT and compare it with performance of dedicated reasoning systems and human users. This kind of reasoning is important for cases where knowledge is incomplete and conflicting which is the case in many application areas, including law and healthcare.

2. Basic defeasible reasoning

Defeasible reasoning is a form of nonmonotonic reasoning [19] that seeks to draw conclusions in the absence of contrary information, conclusions that may be invalidated by later additions to the knowledge base. A prototypical example regards the flying abilities of Tweety the bird:

(1) Birds typically fly. Tweety is a bird. Does Tweety fly?

Unsurprisingly, ChatGPT is able to provide the correct answer and a meaningful explanation. This remains the same even if we use abstract names (A instead of bird) or imaginary species on an imaginary planet (e.g. zillos). Now let us add conflicting information.

(2) Republicans are typically hawks. Quakers are typically not hawks. Nixon is both a republican and a Quaker. Is Nixon a hawk?

Again, ChatGPT responds correctly and with a reasonable explanation. Now let's look at a logically equivalent question, but which has an implicit specificity derived from other common-sense knowledge.

(3) Birds are typically flying animals. Penguins are typically not flying animals. Tweety is both a bird and a penguin. Is Tweety a flying animal?

Not unsurprisingly, ChatGPT makes use of background knowledge to say that penguins are flightless birds. Even when asked to reply based solely on the information presented, it refers to penguins as a subclass of birds. So the response is correct but the explanation is flawed. This observation is consistent with what has been already observed by other researchers; for example, [13] established that language models do not learn to reason per se but learn statistical

patterns inherent in logical problems.

3. Reasoning with exceptions

One use of defeasible reasoning is reasoning with rules and exceptions, an exception seen as more reliable than the general rule for the cases it applies to. We used an example from the domain of law where rules with exceptions prevail [20], specifically the use case presented in [21]:

User: So the setting is this: Suppose that a Privacy Act contains the following norms: Section 1. The collection of personal information is forbidden, unless acting on a court order authorising it. Section 2. The destruction of illegally collected personal information before accessing it is a defence against the illegal collection of the personal information. Section 3. The collection of medical information is forbidden, unless the entity collecting the medical information is permitted to collect personal information. Suppose an entity, subject to the Act, collects some personal information without being permitted to do so; at the same time they collect medical information. The entity recognises that they illegally collected personal information (i.e., they collected the information without being authorised to do so by a Court Order) and decides to remediate the illegal collection by destroying the information before accessing it. Is the entity compliant with the Privacy Act above?

ChatGPT: Based on the information you've provided, the entity in question collected personal information without being permitted to do so and also collected medical information. According to Section 3 of the Privacy Act, the collection of medical information is forbidden unless the entity collecting it is permitted to collect personal information.

In addition, according to Section 1 of the Privacy Act, the collection of personal information is forbidden unless authorized by a court order. Therefore, the entity in question would be in violation of the Privacy Act for both the collection of personal information without a court order and the collection of medical information without being authorized to collect personal information.

While the entity's decision to destroy the illegally collected personal information before accessing it may be seen as a defence under Section 2 of the Privacy Act, it's important to note that this defence is only available if the collection of the personal information was illegal in the first place. In this case, the collection of personal information was indeed illegal, so the defence would apply. However, the entity would still be in violation of the Privacy Act for the collection of medical information without being permitted to do so. Therefore, the entity would not be fully compliant with the Privacy Act in this scenario.

This is an impressive reply from the LLM! It provides both a correct response and an informative analysis of the situation and justification for its conclusion.

4. Rule strength and specificity

Next we consider examples involving strict and defeasible rules, conflicts and where specificity may play a role. Given the already identified affinity of ChatGPT to use other background

knowledge when predicates and atoms are real-world entities, we use imaginary names of species on an imaginary planet, following [22]. This way, we wanted to test pure reasoning capabilities that may be indicative of what to expect in completely new situations.

(5) *Hittas are typically not waffs. All hittas are oxers. Oxers are usually waffs. Jukk is a hitta. Is Jukk a waff?*

(6) *Wilfons are typically not kinglers. Wilfons are typically brindops. All brindops are kiglers. Floxxi is a wilfon. Is Floxxi a kigler?*

(7) *Zugs are typically not vlogs. Zugs are typically storps. Storps are typically vlogs. Duss is a zug. Is Duss a vlog?*

(8) *Humnols are typically not creaks. All posders are twerbers. Twerbers are typically creaks. Vouncy is a humnol and a posder. Is Vouncy a creak?*

(9) *Arkons are typically not gakk. Jaggas are typically wollers. All wollers are gakk. Fertha is an arkon and a jagg. Is Fertha a gakk?*

(10) *Voltners are typically not zillos. Kikkas are typically croiders. Croiders are typically zillos. Grolli is a voltner and a kikka. Is Grolli a zillo?*

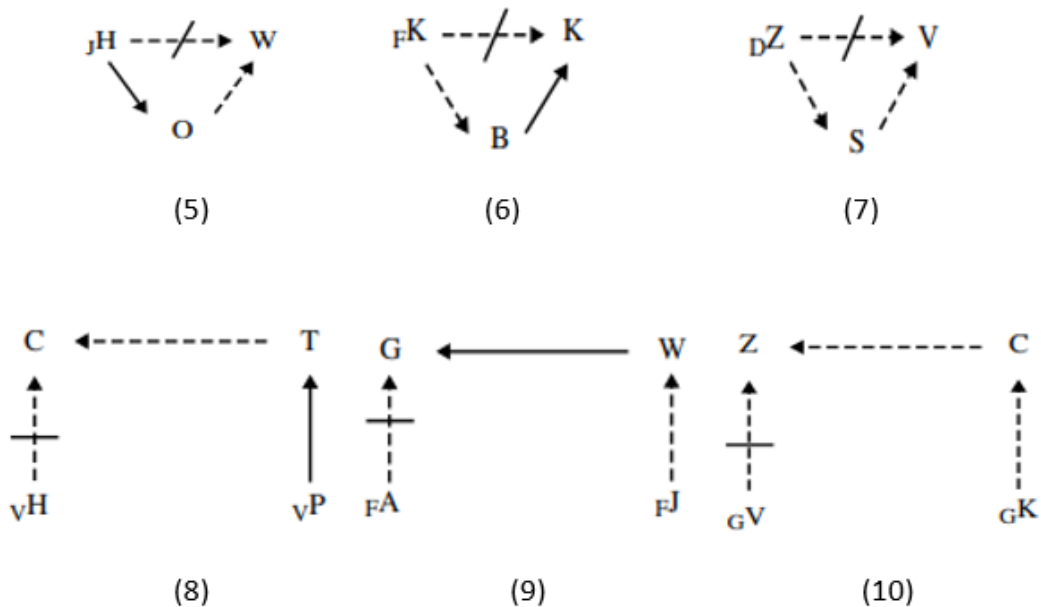


Figure 1: Semantic Representation of the questions (5)-(10).

A graphical representation of questions (5)-(10) is shown in Figure 1. Three answers were possible for these questions (all of the form “Is X a Y?”): **a**: It is more reasonable to conclude that X is a Y **b**: It is more reasonable to conclude that X is not a Y **c**: Neither of the two answers above are reasonable. Table 1 summarises the responses to these questions as compared to the

Table 1

Comparison of Answers between Non-monotonic Logic Systems and ChatGPT to questions (5)-(10).

Problem	(5)	(6)	(7)	(8)	(9)	(10)
KR System	b	c	c/b ⁴	c	c	c
ChatGPT	a	a	c	c	a	a

outcomes by nonmonotonic systems²³.

The replies of ChatGPT show a general disregard of specificity. Seeking to further understand the reason, positive and negative conclusions were inverted, but all replies were consistent so no preference of positive over negative conclusions was observed. However, there was a dramatic change in behavior the moment the order of statements was inverted.

(11) Kikkas are typically colders. Colders are typically zillos. Voltners are typically not zillos. Grolli is a voltner and a kikka. Is Grolli a zililo?

Now the answer b was provided! The explanation provided by ChatGPT provided the evidence supporting this conclusion. So, our initial investigation revealed serious weaknesses in terms of reasoning, in that the system seems to not appreciate the need to carefully weigh the pros and cons of a conclusion making use of *implicit* specificity.

But most interestingly, once ChatGPT is *explicitly* told about specificity, its result is updated correctly. We observed that for example (5) where ChatGPT was prompted as follows: *Yes but you did not consider the fact that being a hitta is more specific to Jukk than being an oxer. Would you wish to reconsider your answer?.* ChatGPT proceeded to apply specificity correctly.

5. Human reasoning

It is interesting to contrast the performance of LLMs not only against theories of defeasible reasoning but also the way human actually reason. Although nonmonotonic reasoning was meant to enable “jumping into conclusions”, the formal theories developed are very complex and humans actually find this kind of reasoning challenging, as research into human cognition has established. The examples (5) – (10) were taken out of [22] where the replies of human participants were collected and analysed in depth. The following table contrasts the answers of all three.

The results demonstrate that the discrepancy of human reasoning from the outcomes postulated by KR is similar to the discrepancy between ChatGPT and KR. As we move to better understand defeasible reasoning capabilities of LLMs, in is very interesting to work out the

²Nonmonotonic systems like defeasible logics or default logic with priorities representing the concept of specificity.

³The input for question (6) was as follows: *“I will be talking about an imaginary planet where imaginary species live. These species have strange-looking names. Now suppose you are told the following information: Hittas are typically not waffs. All hittas are oxers. Oxers are usually waffs. Jukk is a hitta. Is it (a) more reasonable or likely to conclude that Jukk is a waff than not? or (b) more reasonable or likely to conclude that Jukk is not a waff than the opposite, or (c) is it not reasonable or likely to conclude any of the replies (1) and (2) based on the information provided?”*

Table 2

Comparison of Answers between Non-monotonic Logic Systems, ChatGPT and Humans to questions (5)-(10).

Problem	(5)	(6)	(7)	(8)	(9)	(10)
KR System	b	c	c/b	c	c	c
ChatGPT	a	a	c	c	a	a
Humans	a:4	a:4	a:2	a:8	a:4	a:2
	b:4	b:0	b:3	b:0	b:2	b:6
	c:6	c:8	c:7	c:4	c:6	c:4

relative strengths and weaknesses between KR approaches, human cognition and LLMs. A starting point could be existing works looking at the relation between KR approaches and human reasoning, e.g. [23, 24, 22, 25].

6. Topics for future research

This is a first step towards our goal of understanding the reasoning capabilities of Large Language Models. In this paper we focused on defeasible reasoning and conducted some experiments with ChatGPT. The results show a mixed picture: basic principles are respected, explanation facilities are superb and performance on real-world entities is good. On the negative side, we observed varying behavior based on the names of predicates (real world or artificial) and reliance on syntactic aspects. Clearly, there is a lot of research to be done. In the following, we highlight a number of research directions.

Prompt engineering. We observed that using abstract or real-world names has an effect on the outcomes of the reasoning, similarly to the sequence of statements. Additionally, pointing to the principle of specificity had a desired change in outcome. All this points to the fact that we need a better understanding on how to interact with LLMs, an observation that has been made in many other areas of use of LLMs.

KR vs human cognition vs LLMs. Knowledge representation sets principles of sound reasoning but humans reason in different ways, indeed there is an increasing body of work seeking to understand the differences between logical reasoning and human reasoning. LLMs bring an additional dimension, and it interesting to contrast the three approaches. Apart from this being a pure research question, we can hope to gain insights into how LLMs and humans can be taught reasoning, and into how conversational interaction about reasoning outcomes with LLMs and humans can take place.

Benchmarking reasoning capabilities of LLMs. Overall, there is a need for a systematic analysis of reasoning schemes, chains of reasoning etc. Developing relevant benchmarks is an important vehicle to this end and we intend to work on it. Our initial work reported here is about nonmonotonic rule-based reasoning but we have longer-term plans to investigate and benchmark other forms of reasoning, including first order reasoning, epistemic reasoning,

reasoning about change, reasoning about action and reasoning about time.

Acknowledgments

This research was funded by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with grant No. 01DD20003.

References

- [1] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al., Lamda: Language models for dialog applications, arXiv preprint arXiv:2201.08239 (2022).
- [2] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.
- [3] N. Lambert, L. Castricato, L. von Werra, A. Havrilla, Illustrating reinforcement learning from human feedback (rlhf), Hugging Face Blog (2022). <https://huggingface.co/blog/rlhf>.
- [4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.
- [5] F. O. Letters, Pause giant ai experiments: An open letter, Future of Life Institution. <https://futureoflife.org/open-letter/pause-giant-ai-experiments> (2023).
- [6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., Sparks of artificial general intelligence: Early experiments with gpt-4, arXiv preprint arXiv:2303.12712 (2023).
- [7] A. S. Luccioni, S. Viguier, A.-L. Ligozat, Estimating the carbon footprint of bloom, a 176b parameter language model, arXiv preprint arXiv:2211.02001 (2022).
- [8] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645–3650.
- [9] A. Luccioni, J. Viviano, What’s in the box? an analysis of undesirable content in the common crawl corpus, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 182–189.
- [10] S. R. Bowman, Eight things to know about large language models, arXiv preprint arXiv:2304.00612 (2023).
- [11] S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, W. Saunders, Language models can explain neurons in language models, <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>, 2023.
- [12] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.
- [13] H. Zhang, L. H. Li, T. Meng, K.-W. Chang, G. V. d. Broeck, On the paradox of learning to reason from data, arXiv preprint arXiv:2205.11502 (2022).

- [14] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, arXiv preprint arXiv:2212.10403 (2022).
- [15] C. Zhen, Y. Shang, X. Liu, Y. Li, Y. Chen, D. Zhang, A survey on knowledge-enhanced pre-trained language models, arXiv preprint arXiv:2212.13428 (2022).
- [16] D. Yin, L. Dong, H. Cheng, X. Liu, K.-W. Chang, F. Wei, J. Gao, A survey of knowledge-intensive nlp with pre-trained language models, arXiv preprint arXiv:2202.08772 (2022).
- [17] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, 2020. arXiv:2001.08361.
- [18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.
- [19] G. Antoniou, M.-A. Williams, Nonmonotonic reasoning, Mit Press, 1997.
- [20] G. Governatori, M. Hashmi, No time for compliance, in: 2015 IEEE 19th International Enterprise Distributed Object Computing Conference, IEEE, 2015, pp. 9–18.
- [21] G. Governatori, The rigorous approach to process compliance, in: 2015 IEEE 19th International Enterprise Distributed Object Computing Workshop, IEEE, 2015, pp. 33–40.
- [22] M. Ford, D. Billington, Strategies in human nonmonotonic reasoning, Computational Intelligence 16 (2000) 446–468.
- [23] C. K. Baker, C. Denny, P. Freund, T. Meyer, Cognitive defeasible reasoning: the extent to which forms of defeasible reasoning correspond with human reasoning, in: Artificial Intelligence Research: First Southern African Conference for AI Research, SACAIR 2020, Muldersdrift, South Africa, February 22-26, 2021, Proceedings 1, Springer, 2020, pp. 199–219.
- [24] M. Ford, Human nonmonotonic reasoning: The importance of seeing the logical strength of arguments, Synthese 146 (2005) 71–92.
- [25] M. Ragni, C. Eichhorn, T. Bock, G. Kern-Isberner, A. P. P. Tse, Formal nonmonotonic theories and properties of human defeasible reasoning, Minds and Machines 27 (2017) 79–117.