Does Content Effect in LLMs Point to Genuine Reasoning?

Karthik Prasanna Natarajan[1]

[1]Human Sciences Research Center, International Institute of Information Technology, Hyderabad, Telangana, 500032, India

Abstract

Training and evaluating LLMs on deductive reasoning tasks has attracted much attention in recent times. Some studies have shown interesting results which suggest that LLMs behave like humans in displaying content effect, that is, they reason better on reasoning tasks containing rules that align with our everyday beliefs and reason poorly when the rules are belief-violating. On the other hand, there are studies which challenge whether LLMs genuinely reason and attribute their reasoning to artifacts from the training data. In order to make sense of claims concerning genuine reasoning, we introduce a framework developed by Diane Proudfoot's externalist criteria for machine cognition, which is based on Wittgenstein's argument that deductive reasoning involves rule-following which is normative in nature. We propose the use of Proudfoot's criteria for rule-following as a framework to distinguish genuine deductive competence from quasi deductive competence. In doing so, we also draw attention to the limitations and implications of Proudfoot's claims regarding machine cognition through the introduction of a thought experiment. This thought experiment enables us to think through Proudfoot's argument, according to which it is due to pragmatic considerations–and not in principle–that LLMs are unlikely to possess genuine reasoning.

Keywords

deductive reasoning, large language models, content effect, Wittgenstein, rule-following

1. Introduction

With recent developments in the field of large language models (LLMs), there have been various attempts to improve and evaluate the reasoning performance of LLMs. In one such study, Dasgupta et al., 2023 [1] showed that LLMs perform better on reasoning tasks involving realistic rules, that align with the beliefs of the society, than arbitrary/nonsensical rules, and perform poorly when the rules violate the beliefs of the society. Seals et al., 2023 [2] confirmed this effect and further divided the realistic rules into social and non-social rules, to find that LLMs perform better on the reasoning tasks involving social rules than non-social rules.

This is termed as *content effect*, where reasoning is affected by the semantic content of the deductive arguments, when reasoning ideally involves following the rules of logic without the influence of the content. For instance, it is shown that in the Wason selection task, LLMs perform better when the task contains a realistic rule such as: "If the clients are going skydiving, then they must have a parachute", when compared to an arbitrary rule such as "If the cards have plural word, then they must have a positive emotion". Similarly in NLI (Natural Language Inference) task, LLMs perform better when the tasks contain realistic rule such as "If seas are bigger than puddles, then puddles are smaller than seas" as opposed to belief-violating rule such as "If puddles are bigger than seas, then seas are smaller than puddles" or nonsensical rule "If vuffs are bigger than feps, then feps are smaller than vuffs". (Dasgupta et al., 2023 [1]).

Studies ([3],[4]) have shown that humans show content effect on reasoning tasks involving realistic and social rules that align with the beliefs of the society and common sense. Given the content effect observed in LLMs is similar to how it is observed in humans, with reasoning performance having influence on realistic and social rules, it is possible to claim that LLMs reason similar to humans. That is, LLMs may genuinely reason like humans.

Dasgupta et al., 2023 [1] provides plausible speculations for the content effect observed in LLMs: The neural mechanisms of LLMs are similar to the neural mechanisms of humans; as discovered in

3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE24), co-located with AIxIA 2024, November 25-28, 2024, Bolzano, Italy

prasanna.karthik@research.iiit.ac.in (K. P. Natarajan)

humans, LLMs may be said to possess dual systems–system 1 for intuitive reasoning and system 2 for explicit reasoning; and believable situations help LLMs draw more accurate conclusions in the same way as believable situations helps humans draw better inferences–providing them with an evolutionary advantage for survival. These claims need further validation, and much work is being done in this regard. We however propose a more conceptual and philosophical approach to validate whether LLMs exhibit genuine reasoning by adopting Diane Proudfoot's work on machine cognition which is based on Wittgenstein's distinction between rule-following and quasi rule-following [5].

Reasoning can be viewed both in terms of internalist and externalist approaches. For an internalist, reasoning process completely depends on the internal states of an entity. For an externalist, the reasoning process partly depends on the internal states of the entity and largely depends on the entity's history and social environment. So for an externalist, it is impossible for an entity to reason without a history and the presence of a social environment. In this work, we consider reasoning from an externalist view.

Proudfoot interprets Wittgenstein as an externalist (Proudfoot 2004 [5]). In his early work [6], Wittgenstein defines reasoning as following the rules of logic, by logically deriving the conclusion from the premises. In his latter work [7], Wittgenstein defines reasoning as following the adopted reasoning rules that are socially normative. The argument "2 + 3 = 5" is valid because the rules of arithmetic are normatively followed by all the members of the community. Hence, for an entity to genuinely reason, the entity has to be a genuine *rule-follower*. A quasi rule-follower is one who *behaves* like following a rule whereas a genuine rule-follower "behaves in accordance with a rule and also understands what, in producing a certain output for a given input, she is doing" (Proudfoot, 2004 [5]). Proudfoot adopts Wittgenstein's distinction to formulate the necessary conditions for a machine to possess genuine rule-following.

2. Proudfoot's Conditions for Rule-following

To distinguish between a rule-follower, who genuinely follows a rule, and a quasi rule-follower, who behaves like following a rule, Proudfoot (Proudfoot 2004 [5]) defines three conditions necessary for an entity to be a rule-follower:

- **Social Environment:** An entity A is said to follow a rule R only if the behaviour of A takes place in a social environment. From Wittgenstein's view, the social environment labels certain behaviour as *correct* and certain behaviour as *wrong*.
- **Normative Weight:** An entity A is said to follow a rule R only if A attaches a normative weight to its behaviour in accordance with R. For example, A choosing 5 as an answer for square root of 25 is not from a lucky draw, rather it is *correct* to choose 5 based on the associated normative weight.
- **History of Prescriptive Training:** An entity A is said to follow a rule R only if A has a history of prescriptive training. The prescriptive training can be given by means of examples, reward and punishment.

3. Evaluating Proudfoot's Conditions based on Content Effect

We evaluate the results of content effect as a marker for genuine reasoning by looking at how it satisfies Proudfoot's conditions.

3.1. Condition 1: Social Environment

Though the existence of content effect does not necessarily mean the involvement of social environment during the reasoning process, we can infer the involvement of social environment from the way we use LLMs. We, humans, provide LLMs with prompts of reasoning tasks, thereby creating a social

environment in which LLMs reason (behave). Though this environment is very narrow and *thin* [8] in comparison to the social environment that we humans interact with in our everyday life, it does capture a small subset of the social environment. As a consequence of humans giving feedback by prompting and engaging with LLMs' responses, condition 1 may said to be satisfied.

3.2. Condition 2: Normative Weight

In order to verify whether LLMs associate a normative weight to follow the rules of reasoning, it is necessary to verify whether LLMs associate normative weight to social norms, practices and beliefs of the society [7].

It is here that content effect may be seen as playing a significant role. If it is true that LLMs show content effect, then it may also be inferred that they assign a normative weight to what is considered to be correct or true according to the society. This is the crux of the claim of Dasgupta et al., 2023 [1] who show that LLMs perform better at tasks whose content aligns with realistic beliefs in comparison with those that are based on violating and nonsensical beliefs. In doing so they speculate that models reason the same way as humans do. It is imperative therefore to examine whether the results of their study reflect genuine content effect, thereby satisfying Condition 2, or not.

Dasgupta et al., 2023 [1] have created the evaluating data from scratch using an algorithm to make sure that the data with which they evaluate the deductive competence of LLMs is not contaminated. They created three categories of rules (realistic, nonsense/arbitrary, and violate) to prove the presence of content effect by evaluating their deductive competence across three tasks: natural language inference, judging logical validity of syllogisms, and Wason selection task. They observe the performance to be maximum for reasoning tasks with realistic rules, least for belief-violating rules, and somewhere in the middle for nonsense/arbitrary rules.

From this they point towards a normative account of content effect whereby "content effects can emerge from simply training a large transformer to imitate language produced by human culture, without explicitly incorporating any human-specific internal mechanisms" (Dasgupta et al., 2023, p. 25 [1]). This is very much in line with Proudfoot's second condition for genuine rule-following. Dasgupta et al., 2023 [1] propose two plausible origins to account for content effects, which we shall return to shortly.

3.3. Condition 3: History of Prescriptive Training

The way we train and interact with LLMs provides them with a history of prescriptive training. The training phase of LLM involves a loss function that acts as a punishment, which LLMs try to reduce. The RLHF techniques (Reinforcement Learning from Human Feedback) involve LLMs receiving feedback from humans, which gets modeled as rewards and punishments in LLMs' training. In techniques such as few-shot learning, LLMs are provided with examples as a part of their task-specific training. Since the various families of language models used for the study are trained using some variations of language modeling tasks using the aforementioned techniques, each of the LLMs within these families may be said to have been subject to a history of prescriptive training. The third condition for rule-following may therefore be said to be satisfied.

On the face of it, it seems like all the three conditions of rule-following for reasoning are satisfied and the phenomenon of content effects seems to evidence the claim that LLMs may be said to *genuinely* follow the rules of reasoning. However, condition 2 requires further examination as to the *origin* of these content effects. We propose a thought experiment to think through condition 2.

4. *Medieval Chinchilla* - Thought Experiment

Dasgupta et al., 2023 [1] suggest two plausible origins to account for content effects. The first account points to the possibility that the content effects are directly learned through imitation from human-generated data used to train the models. The second account, on the other hand, points to the possibility

that "the model's exposure to the world reflects semantic truths and beliefs and that language models and humans both converge on these content biases that reflect this semantic content for more task-oriented reasons" (Dasgupta et al., 2023, p. 26 [1]). Of these two, it must be noted that the second account holds greater significance for the second condition of genuine rule-following. Whereas the first condition is compatible with quasi rule-following, it is the second account that is grounded on assigning normative weight to realistic/true situations as opposed to violating and nonsense ones. In order to examine whether the second account of content effects is a plausible one, we introduced the following thought experiment:

Imagine a LLM, *Medieval Chinchilla*, existed in the medieval times and imagine that there existed the prerequisite knowledge and tools required for developing, training, and implementing the model. The data used to train *Medieval Chinchilla* would be made up of beliefs consistent with the medieval ages. Hence, the data would contain "realistic" beliefs such as "The Earth is flat"; "The Earth is the centre of the universe", etc.

Medieval Chinchilla, like all other LLMs, would be trained to predict the next-word/missing-word *in* the data *from* the data. For *Medieval Chinchilla*, the training data is the absolute source of truth. The question to ponder is whether *Medieval Chinchilla* show content effects, and if so what would they refer to? It seems plausible that the medieval model would show content effects consistent with the training data and therefore would respond to the prompt "The Earth is <mask>" with the response *flat*, during its training. Much like its medieval human counterparts, we expect it to accept this belief and not for it to question whether the Earth is *truly* flat. In other words, the content effect of *Medieval Chinchilla* originates from the training data (the first account of Dasgupta et al., 2023 [1]) and not from the external world (as required by Condition 2). It lacks the normative ability to question, validate and update its beliefs based on evidence from the external world and the semantic truths therefrom.

Let us develop our thought experiment further. This time someone like Galileo, after observing through his telescope and claiming that Copernicus' theory is true–shows evidence for the fact that the sun is the centre of the universe and all planets revolve around the sun. Let us assume that Galileo prompts this new belief to *Medieval Chinchilla*. Would *Medieval Chinchilla* validate and accept this new belief? Since *Medieval Chinchilla* is trained to predict the masked tokens from the existing beliefs, it is trained to accept the existing beliefs, and thus would not accept the new knowledge prompted to it by Galileo, even if one attempted to provide it with sufficient evidence from the real world.

In order to revise the existing beliefs of LLMs, various studies [9] have come up with techniques to edit their existing beliefs. Such techniques involve finetuning (prescriptive training) LLMs with new or revised beliefs. Let us assume that people in medieval times had this knowledge of editing beliefs in LLMs.

Assume Galileo uses this belief editing technique and revises *Medieval Chinchilla*'s belief to "The sun is the centre of the universe". Would the medieval model consistently update the revised belief? Hase et al., 2024 [9] concludes that though LLMs associate high probabilities to the revised beliefs, they fail to maintain probabilistic coherence and logical coherence with their existing beliefs with respect to the revised belief. In Galileo's case, though he might be successful in revising the new belief that "The sun is the centre of the universe", *Medieval Chinchilla* would still believe that "The sun revolves around the Earth". That is, although *Medieval Chinchilla* may be made to revise its belief, it would not have the capacity to revise those beliefs which logically depend on the newly revised beliefs. All this illustrates that the so-called content effects exhibited by LLMs are mere artefacts of the training data and do not reflect the assignment of normative weight to realistic beliefs.

Now lets analyse whether *Medieval Chinchilla* satisfies Proudfoot's conditions [5] of rule-following for reasoning. From the above thought experiment, it may be concluded that although LLMs may be said to satisfy conditions 1 and 3 of genuine rule-following, they fall short of meeting condition 2. Moreover, it can be seen that in the case of LLMs, the training data is all that is the case (to paraphrase Wittgenstein's first proposition from *Tractatus* [6]). Given the implausibility of LLMs to go beyond their training data, it can be concluded that assigning normative weight to semantic truths and ipso facto, genuine reasoning, is not possible in principle. Thus with respect to LLMs as they stand at the moment, we disagree with Proudfoot, who holds that it is on pragmatic grounds that we may be unable

to build machines which satisfy all the three conditions of genuine rule-following (Proudfoot 2004, p. 290-293 [5]).

5. Conclusion

In this paper, we proposed a conceptual framework to evaluate deductive reasoning in LLMs. We show that the Wittgenstein's distinction between rule-following and quasi rule-following, as adapted to machine cognition by Proudfoot, offers a robust framework to distinguish genuine reasoning from quasi reasoning in language models. The thought experiment introduced in this paper helps in accounting for the source of content effect in the imitation of training data rather than in the assigning of normative weight to semantic truths from the external world. All this goes to suggest that LLMs as they stand at the moment may not be said to possess genuine reasoning.

References

[1] I. Dasgupta, A. K. Lampinen, S. C. Y. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, F. Hill, Language models show human-like content effects on reasoning tasks, 2023. URL: https://arxiv.org/abs/2207.07051. arXiv:2207.07051.

[2] S. M. Seals, V. L. Shalin, Evaluating the deductive competence of large language models, 2024. URL: https://arxiv.org/abs/2309.05452. arXiv:2309.05452.

[3] L. Cosmides, The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task, Cognition 31 (1989) 187–276. URL: https://www.sciencedirect.com/science/article/pii/0010027789900231. doi:https://doi.org/10.1016/0010-0277(89)90023-1.

[4] L. Cosmides, J. Tooby, Cognitive adaptations for social exchange, Evolutionary Psychology and the Generation of Culture (1992).

[5] D. Proudfoot, The implications of an externalist theory of rule-following behaviour for robot cognition, Minds and Machines 14 (2004) 283–308. doi:10.1023/B:MIND.0000035459.85213.e5.

[6] L. Wittgenstein, Tractatus Logico-Philosophicus, Routledge, London, UK, 1961.

[7] L. Wittgenstein, Philosophical Investigations, Wiley-Blackwell, New York, NY, USA, 1953.

[8] L. Daston, A Short History Of What We Live BY, Princeton University Press, 2022.

[9] P. Hase, T. Hofweber, X. Zhou, E. Stengel-Eskin, M. Bansal, Fundamental problems with model editing: How should rational belief revision work in llms?, 2024. doi:10.48550/arXiv.2406.19354.