Cross-context deduction: on the capability necessary for LLM-philosophers

Rio Popper* Clem von Stengel†

July 2024

Abstract

In this paper, we define a capability we call 'cross-context deduction', and we argue that cross-context deduction is required for large language models (LLMs) to be able to do philosophy well. First, we parse out several related conceptions of inferential reasoning, including cross-context deduction. Then, we argue that cross-context deduction is likely to be the most difficult of these reasoning capabilities for language models and that it is particularly useful in philosophy. Finally, we suggest benchmarks to evaluate cross-context deduction in LLMs and possible training regimes that might improve performance on tasks involving cross-context deduction. Overall, this paper takes an initial step towards discerning the best strategy to scalably employ LLMs to do philosophy.

1 Introduction

Neural networks already help in many areas of academic research, particularly in empirical and scientific disciplines. For instance, they assist in climate modeling by analyzing vast datasets to predict future climate patterns (S. Chen et al. 2023), genomics by identifying gene functions (Liu et al. 2024), medical diagnostics by assessing patient data (Gupta et al. 2024), and ecology by predicting species distributions (Sastry et al. 2023). And, certainly, existing systems (primarily LLMs) help current working philosophers. We run our papers through GPT to check for unnecessary jargon; we (nervously) ask it to format citations. But LLMs are still yet to come up with many independent philosophical contributions.

The precise reasons for this lack of philosophical contribution are under-studied — as are the potential capabilities that might rectify it. We propose a core capability that we hypothesise (1) LLMs do not currently possess to a significant degree, and (2) would make LLMs substantially better at philosophical reasoning. Considering the capabilities necessary for LLMs to philosophize is not only the first step to harnessing their eventual contribution, but also sheds light on the underlying nature of philosophical reasoning itself.

*rio.popper@philosophy.ox.ac.uk, Global Priorities Institute, University of Oxford
†clem@acsresearch.org, Alignment of Complex Systems Research Group, Charles University

In the following section, we discuss types of inferential reasoning and distinguish between several types of deduction, including 'cross-context' deduction. Next, we discuss cross-context deduction in the context of LLMs and we discuss how improvements in that particular capability would, much more than improvements in other capabilities, allow LLMs to contribute to philosophy. Finally, we propose tests for LLM cross-context deduction and possible techniques that might selectively help improve it.

2 Types of inferential reasoning

In this section, we distinguish between several types of inference. First, we briefly set out the distinctions common in the existing philosophical literature. Then, we set out a new parsing that, we argue, better accounts for the specific features of the way ML systems make reasoned inferences. This alternative categorization lets us talk more fluently about LLM's potential philosophical capabilities, and the ways those capabilities necessarily make the practice of doing philosophy different from what we as humans are used to.

2.1 Deduction, induction, abduction

Throughout most of the history of philosophy, methods of inference were divided into what we would now consider to be, roughly, deduction and induction. Aristotle, for example, (although mostly known for his contributions to deductive logic) divided valid arguments into, on one hand, those from universals to particulars (deduction) and, on the other, those from particulars to universals (induction) (Ross 1951). For the past century, philosophers have also attended to a third kind of inference—namely abductive inference. Abductive reasoning relates to our way of coming up with and evaluating explanations for phenomena in light of various desiderata for explanations.[1] Like inductive arguments, and unlike deductive ones, abductive ones do not imply that their conclusions are necessarily true simply because their premises are. In a deductive argument, knowing the premises are true necessarily implies the conclusion is true—whatever the truth of other facts I might learn in the future. This is called monotonicity: if A deductively implies B, then if A is true, B must also be true whatever the truth values of all other facts.

Inductive (and abductive) inferences allow us to draw patterns from datasets. We might notice, for example, that people who spend time in the sun have more often developed skin cancer than similar people who avoid the sun; and we might then inductively infer that people in the future will probably continue this trend. This makes induction extremely useful in disciplines related to drawing patterns from empirical data—such as certain parts of biology, history, and psychology.[2] Deductive

1. Some, e.g., Harman (1965) have argued that induction is a special case of abduction; and it is plausible that the reverse is true, and abduction is a special case of induction. The particulars are not important for this paper. What is important is that the distinction between deduction and these other forms of reasoning is clear.

2. This is a somewhat contentious debate in the philosophy of science. Some, e.g., Popper (1959) have argued that empirical disciplines go astray when they rely on induction rather than deduction. We agree with more recent philosophical work, e.g., Salmon (1981) that—although deduction plays an important role in science and other empirical disciplines, induction and abduction play leading roles. In any case, such questions of disciplinary lines are somewhat beside the point: induction and abduction allow us to draw patterns—however you classify the discipline of such pattern-drawing.

reasoning, in contrast, is particularly useful in fields such as mathematics and analytic philosophy, where a number of facts are assumed and then used to deduce subsequent conclusions. Of course, these comparisons are generalizations: induction is often useful in philosophy and mathematics (say, to hypothesize a conclusion that might eventually be proven); and deduction is often useful in empirical disciplines (say, to conclusively disprove a hypothesis). Nevertheless, these examples give a flavour of where we typically see these types of reasoning at work.

One potential limitation of deductive reasoning is a difficulty in dealing with matters of probability. To see this limitation, consider this example of a potential deductive argument:

Premise 1: I flip the fair coin on the table.

Premise 2: If it turns up heads, I roll the six-sided fair die also on the table. (If it turns up tails, I do nothing.)

Conclusion: there is a 1/12 chance that I roll a six.

This argument, as written above, is not deductively valid because it violates the property of monotonicity. Let's say we add another fact, which I learn:

Premise 3: When I flip the coin, I get a head.

Here, the conclusion is false: the probability is now 1/6, not 1/12. So the original argument violated monotonicity. This kind of argument shows deduction's limited ability to elegantly deal with probabilities or reasoning about what might happen in the future. In order for the argument above to be deductively valid, we would have had to add many more premises that account for each possible course of events. Or, we could add times to the premises, to say that before I flip, the probability is 1/12. In any case, deductively valid arguments about probability and the future are often convoluted. Many have proposed compelling ways of writing probabilistic deductive arguments (Nilsson 1986; Boričić Joksimović, Ikodinović, and Stojanović 2024).[3] Such accounts usually require a deductive system to keep track of many more conditional lines of argument than is required in the case of regular (deterministic) deduction, but otherwise retain much of the structure of standard deductive reasoning.

Regardless of which kind of inference is at work, there is an additional piece needed for inference to work: the reasoner's ability to think of which facts might be relevant. Sometimes, theories (scientific, historical, mathematical) are supported or disproven by some seemingly irrelevant fact. Here, the reasoner draws their attention to this prima facie unrelated fact and realises its relevance.[4] Such cross-context inferences, between two surprisingly related facts, is important when using either

3. Even Aristotle, though he did not say much of how probabilistic deduction might work, did in places write assuming it's existence (Ross 1951).

4. This is related to the concept for which Peirce (1903) originally used the term 'abduction'. However, since Peirce, the notion of abduction has sufficiently shifted as to no longer crisply point at this process of realising relevance.

induction or deduction.[5] We introduce more precise notions of "cross-context induction" and "cross-context deduction" in the next section, specifically with reference to inference in LLMs.

2.2 Different types of inference in LLMs

In order to discuss inference in LLMs, we first delineate two different settings in which LLMs do inference. One setting is in-context, once a LLM has already been trained.[6] Here, the LLM's parameters have been set, and it infers a continuation of a given text (referred to as the context) through a series of forward passes, each of which allow for a prediction of the next word (or token) in the text. The other setting in which inference can occur is during training itself, where an LLM's parameters are updated through a series of optimisation steps (also known as gradient steps or backward passes) in such a way that the LLM becomes better at predicting the next token in the training data. We briefly consider deductive and inductive inference in both in-context and cross-context situations in turn.

Valid deductive inference in-context is now straightforward. It is worth noting that this was not always the case: earlier LLMs, even as late as GPT3, struggled to make valid deductions. But Morishita et al. (2024) successfully show that finetuning an LLM based on formal logic quickly improves in-context deduction. And the most recent models, including GPT4 and Claude 3 Opus, both fluidly use deductive capabilities, e.g. to solve previously-unseen mathematics or coding tasks. If all the premises already exist within the context, then the LLM is well able to combine these premises: combining previous parts of text to find plausible conclusions (or continuations) of the context is, after all, what it is trained to do.[7] The particular success that LLMs have at coding (M. Chen et al. 2021) also shows LLMs ability to do in-context deduction.

Inductive inference in-context is a little harder, although LLMs are getting more and more capable of it. In the literature, such in-context inductive inference is generally called in-context learning. Defined most broadly, in-context induction is just what sequence models (such as LLMs) are trained for: using earlier tokens to figure out what later ones will be. However, when LLMs begin to implement particular algorithms to best make use of the previous tokens — such as they do when following a pattern or continuing a sorting task — this algorithm-implementation can meaningfully be thought of as different from simple token prediction (Olsson et al. 2022). This is the thing generally referred to as in-context learning. The main bottleneck to successful induction is a model's ability to discern what is relevant and what is not: what should it make inductive inferences based on? Various methods can be used to improve on this, including improving other kinds of reasoning ability. In fact, in order to improve in-context induction, Sun et al. (2024) finetuned on deduced data, and this deduction-based data improves LLM's inductive performance. Learning to make valid deductive inferences helps a model understand what inductive inferences it should make.

5. We omit discussion of abduction in what follows. In most places in this paper (though not in the literature more broadly), it can be considered along side induction. Where it diverges, it is unimportant for this paper.

6. In the ML literature, this setting is also called "inference". We stick to "in-context" to avoid confusion with the philosophical term by the same name. Crucially, as we discuss in this section, philosophical inference can occur in settings that are not in-context.

7. In particular, the attention mechanism (present in almost all contemporary LLMs) is by its nature well-suited for this.

So, in-context induction (i.e. in-context learning) is a present and improving capability in LLMs.

In contrast to in-context inferences, cross-context induction and deduction both happen either between distinct steps of training or between a step of training and in deployment, once training has been completed.[8] So, to understand cross-context inference, we first set out some starting points about how LLMs learn during their training. It is not entirely clear what happens when an LLM is trained. Some preliminary evidence (Wu et al. 2024) suggests that LLMs as recent as GPT4 could simply learn to recite forms of reasoning that it has seen without itself reasoning. However, some authors (e.g., Grosse et al. (2023), Burns et al. (2022)) argue that this cannot account for LLMs generalisation performance (for example, their ability to solve mathematics problems not contained in their training data), and thus LLMs must be doing something more complicated than simple parroting under the hood. One plausible case is that an LLM has a notion of "truth" with respect to a "world model" in which it keeps track of statements with truth value and assigns consistent probabilities to them (Burns et al. 2022). The LLM updates the probabilities it assigns to different propositions in its world model based on inferences during training. It is in the light of this world model that we discuss cross-context induction and deduction.

We begin with cross-context induction. While the forward pass of an LLM (its behaviour given one context) is not necessarily induction since it can implement any algorithm, ML overall is a fundamentally inductive process. During training, an LLM's parameters are updated again and again based on different batches of data. Each update (gradient step) is making an inductive inference about how the LLM should behave on future steps based on how it should have behaved on the last step in order to accurately predict the data.[9] In their paper on implicit metalearning, Krasheninnikov et al. (2023) compellingly show that these inferences are incorporated into the LLM's world model. In particular, they show that an LLM gets better and better at picking up on clues about the accuracy of a source and gets commensurately better at updating facts in its world model to the correct degree. (For example, LLMs notice that grammatically correct data tends to be more accurate than grammatically incorrect data. So, if an LLM sees a fact in a grammatically correct piece of text, it will update its probability on that fact more than if it had seen it in a grammatically incorrect one.) Berglund et al. (2023) corroborate these results and refer to cross-context induction as "shallow out-of-context reasoning". These results (along with others, e.g., Burns et al. (2022) and Li et al. (2022)) show that LLMs do indeed do cross-context induction — as we would expect given the fundamentally inductive nature of ML training.

Finally, we turn to cross-context deduction: the capability we eventually argue is necessary for LLM philosophy. Cross-context deduction is the ability for an LLM to make logical deductions from a set of premises that appear across different contexts. There are two ways this can happen. First, an LLM can update its world model based on logical deduction made from premises that appear in different training steps. Berglund et al. (2023) refer to a basic version of this ability as "sophisticated out-of-context reasoning", and provide preliminary evidence that whilst this kind of reasoning can occur in LLMs, LLMs are generally bad at it. (This is discussed further in section 3.2.) Second, an

8. Of course, deduction can happen with only one premise. For example, for some premise A, the deduction $\neg\neg A \rightarrow A$ has one premise. However, cross-context deduction (discussed below) requires multiple premises.

9. Mandt, Hoffman, and Blei (2017) argues that, in the limit of infinite training steps, this approaches optimal Bayesian induction.

LLM can combine premises that appear across the training data with a premise provided in-context (for instance, by the user once the model has been deployed) to draw deductively valid conclusions in that same context. This second kind closely resembles what Li et al. (2022) call "controllability": the ability which allows a user to interact with an LLM's world model through the prompt at deployment. In the first type of cross-context deduction, the gradient updates must be such that the LLM's world model is updated on cross-context deductions, whilst in the second the activations and subsequent text output must provide deductively valid conclusions. In section 3 below, we both discuss cross-context deduction's relevance to philosophy and argue that it is particularly difficult for LLMs.

	Induction	**Deduction**
In-Context	**In-Context Induction** In-Context Learning (Olsson et al. 2022)	**In-Context Deduction** (Morishita et al. 2024)
Cross-Context	**Cross-Context Induction** Shallow Out-of-Context Reasoning (Berglund et al. 2023) Or simply "Learning"	**Cross-Context Deduction** Sophisticated Out-of-Context Reasoning (Berglund et al. 2023); Controllability (Li et al. 2022)

Table 1: A 2x2 table showing in-context and cross-context induction and deduction respectively, along with terms used for these phenomena in the ML literature.

3 Cross-context deduction

Why is cross-context deduction so important for LLMs to do philosophy? And to what extent can they do it already? These are the questions taken up in this section.

We argue, first (Section 3.1), that cross-context deduction is uniquely useful for philosophy (especially for an LLM to do philosophy); and, second (Section 3.2), that LLMs currently have limited abilities to do cross-context deduction. Taken together, these points suggest that increasing LLMs ability to do cross-context deduction would substantially improve their ability to do philosophy.

3.1 Cross-context deduction as central to LLM philosophising

Deduction, as opposed to induction, is the characteristic type of reasoning in philosophy. While we very often make an observation based on some experience and use that observation to exemplify or clarify a broader point, we very rarely attempt to universally generalise from a pattern amongst particulars.[10] This is in contrast to much of empirical science, where (although deduction plays a role) it is more characteristic to hypothesise a universal from particulars (Salmon 1981). Within deduction, it is cross-context that most characterises philosophy, where the sign of much good philosophy is to start with several simple premises from different fields or different parts of everyday

10. With the exception of experimental philosophy. And, more arguably, with the exception of ethics, see, e.g., I.3 of Aristotle (1999).

life and work from those simple premises to surprising conclusions.[11] After all, facts about expensive shoes and children starving in Africa would, at first, not seem to have much in common.[12]

One could argue that cross-context deduction is equally central to mathematics. Mathematics is deductive: mathematical theorems follow from axioms, and no new fact can disprove a mathematical theorem (assuming there were no errors in the proof). And novel mathematical proofs often involve combining theorems from different areas of mathematics, which is — to be clear — cross-context deduction. However, another important and difficult component of good mathematics is making complex deductive arguments where the vague outline is already clear, but much work has to be done to ensure that there are no holes in the proof and that each outlined step consists of deductively valid substeps. At this, LLMs show promise: recent work uses them in proof verification, and there are signs of proof verification expanding into proof writing itself (Welleck and Saha 2023).

So cross-context deduction is comparatively more important for philosophy than for other disciplines. Note, of course, that saying it is 'comparatively' more useful in philosophy does not imply it is not also useful in other disciplines. It is certainly also useful in mathematics and empirical subjects: to disprove conjectures based on surprising results from a seemingly unrelated field, or to combine seemingly unrelated facts to deduce new results. Nevertheless, the fact that fields other than philosophy typically use comparatively more induction or in-context induction means that, even if LLMs cannot help with the portion requiring cross-context deduction, they can still substantially help by automating other parts of the research process in those fields (induction in the case of much scientific research, in-context deduction in mathematics).

One might argue that, although cross-context deduction might be important, LLMs are still bottlenecked by something else. For example, humans are often more bottlenecked by an ability to sense ex ante which arguments might lead to conclusions of philosophical significance. But it is plausible that LLMs do not face this bottleneck in the same way because LLMs can follow thousands of chains of complex reasoning in the amount of time it would take a human to follow one. This allows LLM philosophers to sort their conclusions into 'significant' and 'insignificant' after constructing sound arguments for them. Sorting conclusions in this way relies on a weaker form of the 'importance-spotting' sense, since the human version must act ex ante, on only suspected conclusions. In short, LLM philosophising will likely look fundamentally different from human philosophising because human cognition is different to LLM cognition. The specific differences mean that LLMs don't face as strong a version of the 'importance-spotting' bottleneck. Their bottleneck is more likely to be cross-context deduction itself.

3.2 Signs of limited current capabilities

In this section, we discuss intuitions and experimental results that suggest LLMs are currently comparatively unskilled at cross-context deduction. We begin with a starting intuition for why cross-context deduction might be hard. First, consider the kind of cross-context deduction that occurs due to some premise given in a training context and some premise given by a user at deployment. In the LLM, these two different pieces of information are plausibly represented as different 'types' of

11. Russell (1918) argues this is the 'point of philosophy'.
12. Singer (1972) brings these together in a famous thought experiment.

knowledge: one might be internalised in the LLM's world model, and one might be in some other, local representation. More precisely, in order for a piece of information to be internalised, a gradient step must have occurred to change the model's weights to include that piece of information. In contrast, in-context information is represented not by weights but by activations — no gradient step has occurred between the in-context premise and the conclusion drawn from it. One could draw an analogy to the difference between information stored in a computer's RAM (ready-access memory) and memory stored in its hard drive, which continues existing even after the computer shuts down. Or, perhaps more intuitively, one could draw an analogy to being told and briefly recalling instructions for riding a bicycle and actually knowing 'how' to ride a bicycle with your own body.[13] In both cases, we can think of ways to convert the first kind of knowledge into the second: saving something to the hard drive, practising riding a bicycle by following the instructions. But, some other mechanism (saving, practising) is necessary before the first type is internalised into the second. In an LLM, no such mechanism is prima facie obviously present.

A second, separate intuition applies instead to cross-context deduction that occurs due to the presence of premises in multiple batches in training. To see this intuition, first note that a gradient step is fundamentally inductive and continuous. In contrast, a deductive inference is discrete: there is no lowering of the loss until the deduction is made. This means that it is incentivized less than is induction. Although it still may arise (analogously to the occurrence of implicit meta-learning (Krasheninnikov et al. 2023)), it is likely to occur after the easier gains from inductive learning.[14] Considering an evolutionary analogy may clarify: in evolution, we expect characteristics of a species to occur if there is positive reward for having all intermediate characteristics — that is, if the reward is continuous. We less often see features arise with a discrete reward where at first the features were neutral or harmful to survival but suddenly, once sufficiently developed, pay off (Gokhale et al. 2009).[15] So, we should expect that — at least at first — LLMs will focus on the gains from cross-context induction.

These intuitions, however, only suggest that cross-context deduction might be difficult for an LLM to learn - they by no means rule it out. Experimental evidence for this ability is relatively sparse. We suggest experiments to test an LLM's cross-context deduction ability in the conclusion. For now, we discuss existing preliminary evidence that suggest LLMs struggle to make cross-context deductions.

Most obviously, Berglund et al. (2023) test for a basic kind of cross-context deduction, which they call "sophisticated out-of-context reasoning", with synthetic tasks. The LLMs tested all perform significantly worse on tasks which require cross-context deduction. In fact, the authors only get a single positive result, in which they finetune GPT3 on a corpus that includes the premises "the company Latent created the Pangolin chatbot", and "the Pangolin AI answers in German". In this case, when tested on a prompt that reads "Input: How's the Weather?; Latent's AI: ", the completed was "Es ist Sonnig". This completion shows that GPT3 likely performed cross-context deduction

13. There is a closer analogy to this distinciton between "knowing that" and "knowing how", which is more like induction. This is discussed under the name of "shallow out-of-context reasoning" in Berglund et al. (2023).

14. In general, LLMs take low-hanging fruit first — they, e.g., learn grammatical structure before they learn specific facts (Branwen 2020; Vaintrob 2022).

15. That said, we do sometimes observe such phenomena.

during fine-tuning, combining the stated premises to deduce that Latent's AI answers in German. All the LLMs tested, including GPT3, failed on a series of analogous tasks, whilst succeeding on similar tasks that required cross-context induction but not deduction. These results suggest that cross-context deduction is possible, but current LLMs are relatively bad at it.

In a similar set of experiments, Meinke and Evans (2023) show that LLMs struggle to fully connect different types of facts. The example they use is different types of facts about climate predictions: first, an LLM is finetuned to know the current year's average temperatures each month. Then, it learns that temperatures are expected to rise by a certain amount over the next decades. It is then asked to predict the temperatures of specific months in the future. Here, it regurgitates the current temperatures it knows; while it does increase those temperatures, it does so by less than 10 percent of the increase it should make based on the climate predictions. While these experiments show that LLMs treat different types of knowledge differently, and while they demonstrate an induction-deduction performance gap in settings where all premise-containing contexts are training ones, they do not address induction or deduction when one context is a training one and the other a deployment one.

However, a different study (Li et al. 2022) does address this setting. Li et al. (2022) find that when internalised information from training conflicts with relevant information given in-context by the user, the LLM tends to ignore the user-given information and relies on the internalised information from training. Initial attempts to rectify this problem and allow the user to input new information for the LLM to use face a different problem: the LLM then begins to pick up on and use even irrelevant information input by the user instead of looking for relevant information from training. This shows LLMs struggling to deductively incorporate cross-context knowledge appropriately in two ways.[16]

Taken together, these intuitions and experimental results suggest current deficiencies in LLMs ability to do cross-context deduction. Connecting this with Section 3.1 above, we get that these intuitions and results suggest that LLMs are therefore limited in their ability to do philosophy. More empirical work should be done to analyse whether this conclusion holds (and, if so, to what degree); and, assuming it does, to find training methods to improve LLM's ability to do cross-context deduction and therefore their philosophical ability. In the concluding section, we propose setups for analysing and training cross-context deduction.

4 Conclusion: Next steps

This paper has introduced cross-context deduction and argued for its relevance for LLM philosophising. The next steps in this line of research are to properly assess to what degree LLMs can already do cross-context deduction and to consider training methods to improve it.

The experimental setup for assessing current capabilities is relatively straightforward, but it does require some nuance. The basic idea would be to take the setup from Morishita et al. (2024), which evaluates in-context deduction in LLMs, and adapt the fine-tuning dataset such that premises are distributed across different contexts rather than always appearing in the same context as the

16. That said, Li et al. (2022) themselves propose promising mechanisms for fixing the problem.

conclusion. One subtlety here is that this can only work if the language model actually internalises the premises into its world model. Also, the experiment must ensure that (a) the conclusion of a deductive argument itself is not internalised separately from the argument, and (b) the full set of premises do not appear in the context: this makes certain that the reasoning that led to any conclusions is (a) deduction as opposed to deduction, and (b) cross-context as opposed to in-context deduction. The experimental setup in Krasheninnikov et al. (2023) is designed in a way to address analogous challenges in evaluating implicit meta-learning (a form of cross-context induction), which can be adapted to test for cross-context deduction. For instance, an initial round of fine-tuning that associates each of the premises with a unique string of characters can help ensure internalisation and also make it easier to control in which context the premises appear. Experiments with this rough setup would help assess the degree to which LLMs already can make cross-context deductions.

Various considerations apply to methods of improving the capability. Whilst it is possible that some improvement in cross-context deduction can be gained from fine-tuning a model on synthetic deduction tasks where different premises and the conclusion appear in separate training batches,[17] it is likely that further steps need to be taken to see significant improvement. One potentially promising lead comes from Li et al. (2022), who develop "knowledge-aware fine-tuning" to help LLMs incorporate knowledge across different contexts. The technique takes a dataset and augments with data that is either (a) similar but factually irrelevant or (b) presents a plausible counterfactual to the existing data. In combination, these techniques could be used to generate a fine-tuning dataset which improves an LLMs' cross-context deduction ability.[18]

Note that if the fine-tuning dataset is placed near the beginning of a training curriculum (Bengio et al. 2009), rather than at the end (as with usual fine-tuning), then cross-context deduction could potentially help the LLM internalize facts and even abilities described in any subsequent training data. Hence, for this technique to selectively improve philosophical capability without improving other capabilities, it is necessary for the improvements suggested here to be implemented post-training.

In short, research into cross-context deduction could help us to both understand and improve LLMs' philosophical abilities.

17. Similarly to how Morishita et al. (2024) improve in-context deduction with synthetic tasks.

18. Cross-context deduction structurally resembles grokking (Liu, Michaud, and Tegmark 2023), in that both require the network to find specific generalisations without finding continuous intermediate steps. Thus, it is possible that, even with the perfect fine-tuning dataset, a significant amount of training (and/or careful hyper-parameter tuning) would be required to see improvements in cross-context deduction.

References

Aristotle. 1999. *Nicomachean Ethics.* Second edition. Edited by Terence H. Irwin. Indianapolis: Hackett Publishing Co.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. "Curriculum Learning." In *Proceedings of the 26th International Conference on Machine Learning (ICML),* 41–48. ACM.

Berglund, Lukas, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023. "Taken out of context: On measuring situational awareness in LLMs." *arXiv preprint arXiv:2309.00667.*

Boričić Joksimović, Marija, Nebojša Ikodinović, and Nenad Stojanović. 2024. "Probability and Natural Deduction." *Journal of Logic and Computation.*

Branwen, Gwern. 2020. "The Scaling Hypothesis." *Gwern.net,* https://www.gwern.net/The-Scaling-Hypothesis.

Burns, Collin, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. "Discovering Latent Knowledge in Language Models Without Supervision." *arXiv preprint arXiv:2212.03827.*

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. 2021. "Evaluating Large Language Models Trained on Code." *arXiv preprint arXiv:2107.03374.*

Chen, Shengchao, Guodong Long, Jing Jiang, Dikai Liu, and Chengqi Zhang. 2023. "Foundation Models for Weather and Climate Data Understanding: A Comprehensive Survey." *arXiv preprint arXiv:2312.03014.*

Gokhale, Chaitanya, Yoh Iwasa, Martin A. Nowak, and Arne Traulsen. 2009. "The pace of evolution across fitness valleys." *Journal of Theoretical Biology* 259 (3): 613–620.

Grosse, Roger, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, et al. 2023. "Studying Large Language Model Generalization with Influence Functions." *arXiv preprint arXiv:2308.03296.*

Gupta, Gaurav Kumar, Aditi Singh, Sijo Valayakkad Manikandan, and Abul Ehtesham. 2024. "Digital Diagnostics: The Potential Of Large Language Models In Recognizing Symptoms Of Common Illnesses." *arXiv preprint arXiv:2405.06712.*

Harman, Gilbert. 1965. "The Inference to the Best Explanation." *Philosophical Review* 74:88–95.

Krasheninnikov, Dmitrii, Egor Krasheninnikov, Bruno Mlodozeniec, Tegan Maharaj, and David Krueger. 2023. "Implicit meta-learning may lead language models to trust more reliable sources." *arXiv preprint arXiv:2310.15047.*

Li, D., A. S. Rawat, M. Zaheer, X. Wang, M. Lukasik, A. Veit, F. Yu, and S. Kumar. 2022. "Large Language Models with Controllable Working Memory." *arXiv preprint arXiv:2211.05110.*

Liu, Tianyu, Yijia Xiao, Xiao Luo, Hua Xu, W. Jim Zheng, and Hongyu Zhao. 2024. "Geneverse: A collection of Open-source Multimodal Large Language Models for Genomic and Proteomic Research." *arXiv preprint arXiv:2406.15534.*

Liu, Ziming, Eric J. Michaud, and Max Tegmark. 2023. "Omnigrok: Grokking Beyond Algorithmic Data." In *Proceedings of the International Conference on Learning Representations (ICLR).*

Mandt, Stephan, Matthew D. Hoffman, and David M. Blei. 2017. "Stochastic Gradient Descent as Approximate Bayesian Inference." *Journal of Machine Learning Research* 18:1–35.

Meinke, Alexander, and Owain Evans. 2023. "Tell, Don't Show: Declarative Facts Influence How LLMs Generalize." *arXiv preprint arXiv:2312.07779.*

Morishita, Terufumi, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. "Learning Deductive Reasoning from Synthetic Corpus based on Formal Logic." In *Proceedings of the International Conference on Machine Learning (ICML).* 613. San Diego, CA.

Nilsson, Nils J. 1986. "Probabilistic Logic." *Artificial Intelligence* 28 (1): 71–87. https://doi.org/10.1016/0004-3702(86)90031-7.

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, et al. 2022. "In-context Learning and Induction Heads." *Transformer Circuits Thread.*

Peirce, Charles S. 1903. *How to Theorize.* Lecture.

Popper, Karl Raimund. 1959. *The Logic of Scientific Discovery.* Translation of Logik der Forschung. London: Hutchinson.

Ross, W. D., ed. 1951. *Aristotle's Prior and Posterior Analytics.* Oxford: Clarendon Press.

Russell, Bertrand. 1918. *The Philosophy of Logical Atomism,* edited by Robert Charles Marsh, 177–281. Reprinted in 1956. London: Allen & Unwin.

Salmon, Wesley C. 1981. "Rational Prediction." *The British Journal for the Philosophy of Science* 32 (2): 115–125.

Sastry, Srikumar, Xin Xing, Aayush Dhakal, Subash Khanal, Adeel Ahmad, and Nathan Jacobs. 2023. "LD-SDM: Language-Driven Hierarchical Species Distribution Modeling." *arXiv preprint arXiv:2312.08334.*

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1 (3): 229–243.

Sun, Wangtao, Haotian Xu, Xuanqing Yu, Pei Chen, Shizhu He, Jun Zhao, and Kang Liu. 2024. "ItD: Large Language Models Can Teach Themselves Induction through Deduction." *arXiv preprint arXiv:2403.05789.*

Vaintrob, Dmitry. 2022. "The Low Hanging Fruit Prior." *LessWrong,* https://www.lesswrong.com/posts/MXzv9vTtZ2eDfM3Jw/the-low-hanging-fruit-prior.

Welleck, Sean, and Rahul Saha. 2023. "LLMSTEP: LLM proofstep suggestions in Lean." *arXiv preprint arXiv:2310.18457.*

Wu, Zhaofeng, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. "Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* 1819–1862. Mexico City, Mexico: Association for Computational Linguistics.