

Reflection, confabulation, and reasoning

1. The value of reflection

What is the epistemic value of reflection, understood as the capacity to direct one's attention to one's own beliefs? Some philosophers champion reflection as a vital source of epistemic legitimacy and guidance. Ernest Sosa, for example, considers reflection to be a way of enhancing the reliability of our belief formation, and takes reflectively formed knowledge to be "more admirable" than our instinctive and unreflective 'animal knowledge' (Sosa 2009, 142). Other fans of reflection go further, insisting that there can be no knowledge that does not pass some test of reflection (Bonjour 1980, Smithies 2019). Against this spectrum of support for reflection, Hilary Kornblith offers a spectrum of resistance, ranging from cool neutrality to something more like open hostility. In his neutral moments, Kornblith contends that reflection is nothing special: "If what one cares about is the reliability of one's process of belief acquisition, then reflecting on one's beliefs is a mixed bag: sometimes better, sometimes worse, and sometimes just the same as belief uninfluenced by subsequent reflection" (Kornblith 2019, 97). On the question of whether reflection makes a mental state more admirable, Kornblith writes, "There is no second-order magic. Second-order mental states are not so very different from first-order mental states: both are firmly entwined in the same causal net; both are, at times, reasons-responsive, and, at times, disengaged from reason" (Kornblith 2019, 112). Elsewhere, however, the attitude gets frostier: reflection is not just unmagical, it is deceptive, in the sense that it generates an illusion about what it does for us. Kornblith maintains that when we reflect, "we get the impression that we are actually providing some sort of extra screening for our beliefs, and we thus have the very strong impression that we are actually doing something to assure that our beliefs are indeed reliably arrived at. But this is not what we are doing at all. Instead, we are engaged in a process which, in a very wide range of cases, makes us more confident that we are right than before we began, but a process, nevertheless, which is almost useless for improving our reliability" (Kornblith 2012, 24-5). If one wonders how wide the range of problematic cases might be, Kornblith holds that it is "far more often than not" that "the result of reflection turns out to be little more than a ratification of the beliefs held prior to reflective evaluation," increasing confidence to make us "more self-satisfied, even if no more accurate, epistemic agents" (Kornblith 2019, 98). According to Kornblith, the evidence suggests that "reflective self-examination of the sort offered by a first-person perspective on our beliefs fails utterly to do the job of locating the errors which we unreflectively make" (Kornblith 2019, 144).

There is something of a tension here. Given that Kornblith is not in general a skeptical philosopher, one wonders how he can characterize reflection both as "not so very different" from unreflective belief formation, and yet also as something that "creates the illusion" that it is serving us well, epistemically, while completely failing to do so (Kornblith 2012, 26). Kornblith is clearly committed to the idea that "We should not see reflection as offering us a variety of psychological processes which are different in kind from those which operate in us when we fail to reflect" (Kornblith 2019, 11), but he is also keen to argue that the perspective it presents is "grossly misleading" (Kornblith 2019, 151). If reflection is cognitively ordinary, why would it naturally generate an image of itself as something better than that? And why would we have a natural capacity that seems to promise self-improvement despite generally failing to deliver? In this chapter, I take a closer look at the conditions under which reflection is (and is not) problematic, in search of an answer to these questions. While my theory diverges from Kornblith's in places, it takes inspiration from his broader approach, attuned to the reliability of reflection in various conditions. Some of these are conditions Kornblith could not possibly have anticipated when he originally formulated his theory of reflection; indeed, we will start by checking how his theory applies to a very new type of subject.

2. Reflection in large language models

In August of 2019, a passenger on an Avianca Airlines flight was injured when a service cart struck his knee. He sued Avianca, but after a series of snags, including a delay while the airline was in bankruptcy, Avianca applied to have the lawsuit dismissed, noting the expiry of the two-year claim period allowed by the Montreal Convention governing international airline incidents. The passenger's lawyer, Steven Schwartz, was out of his depth, accustomed to arguing personal injury claims in state court, but now needing to tackle issues of federal law in the Southern District of New York. To make matters worse, his law firm's federal case database subscription had lapsed, so he could not look up relevant court cases. However, Schwartz's children told him about a new resource known as ChatGPT, and he found himself asking it for help in preparing his argument against the dismissal (Weiser 2023; *Mata v. Avianca Inc.*, 2023, Case Document 45).

In a series of discussions released as an appendix in later court filings, ChatGPT was a cheerful and compliant conversational partner, replying "Certainly!" when asked to produce lists of cases that might be supportive, and gamely coming up with brief summaries of *Shaboon v. Egyptair*, 2013 and five other cases relevant to the client's predicament. After reading that the court in *Varghese v. China Southern Airlines*, 2019 had extended the two-year Montreal Convention limit, Schwartz asked ChatGPT to provide the text of the court's opinion in that case, and it produced a dialectically useful two-paragraph quotation (*Mata v. Avianca Inc.*, 2023, Case Document 46, Exhibit A).

When opposing counsel had difficulty locating the cases he had cited in his brief, Schwartz was ordered to provide the court with copies of them, which he then asked ChatGPT to produce. ChatGPT was again obliging, and Schwartz dutifully filed the generated output in court. At the time, Schwartz wrote in a later filing, "I still could not fathom that ChatGPT could produce multiple fictitious cases, all of which had various indicia of reliability such as case captions, the names of the judges from the correct locations, and detailed fact patterns and legal analysis that sounded authentic" (*Mata v. Avianca, Inc.*, 2023, Case Document 46). Opposing counsel swiftly challenged the legitimacy of the filed material, noting that database searches for Schwartz's citation docket numbers led either to different cases or to dead ends.

As he at last became suspicious, Schwartz asked ChatGPT whether *Varghese v. China Southern Airlines* was a real case. ChatGPT affirmed that it was. Schwartz asked it for its source. "Upon double-checking," ChatGPT said, "I found that the case *Varghese v. China Southern Airlines Co. Ltd.*, 925F.3d 1339 (11th Cir. 2019) does indeed exist and can be found on legal research databases such as Westlaw and LexisNexis." Asked about the other concocted cases, ChatGPT brazenly insisted that they were also real, and found in "reputable legal databases" (*Mata v. Avianca, Inc.*, 2023, Case Document 54, Appendix B).

In doubling down with fluency and confidence when asked about its fabrications, ChatGPT's version of reflection resembles just the kind of morass that Kornblith warns us about with respect to our own cognition: "When we introspect and ask ourselves why it is that we hold some belief, we unknowingly engage in a process of confabulation, frequently attributing various beliefs to ourselves, beliefs which we take to have been instrumental in producing the very belief we are questioning, but which we did not hold prior to the exercise of self-examination. The process of reflection seems to provide us with direct acquaintance with our reasons for belief, but, instead, it involves an elaborate inference, postulating a variety of reasons we never had, and presenting them in a particularly vivid and compelling form" (Kornblith 2019, 142-3).

When prompted to double-check, ChatGPT cannot introspect, but it is a world-class confabulator. Indeed, its failure of introspection creates a certain purity to its confabulation: lacking self-awareness, it produces a story about its prior assertions that has a vivid and compelling superficial form, with no regard to the actual causal processes giving rise to those assertions. The version of ChatGPT that Schwartz was using in the spring of 2023 had no access to LexisNexis or Westlaw, but its training would have exposed it to those brand names in contexts where the sourcing or authenticity of court cases is under discussion. On encountering such a context in its conversation with Schwartz, ChatGPT responds with output similar to what it has seen in such contexts, following its general mandate to carry on in whatever way seems most probable.

The large language model underpinning ChatGPT is pretrained on vast quantities of raw textual data drawn from the internet in automated scraping, together with some further ‘higher-quality’ material, such as digitized books and Wikipedia, given heavier weight (Brown, Mann et al. 2020). At the core of this pretraining lies a simple token prediction task, typically to predict a masked word, with words (or more generally tokens, including numerals and affixes) represented as embeddings in a multi-dimensional space capturing their contexts of usage, and an attention architecture that enables the model to solicit information from multiple representation subspaces at different positions relevant to the current prediction (Vaswani, Shazeer et al. 2017). The trained model produces output which is statistically likely, both given the current context of the user’s input and conversation so far, and given the background patterns the language model has distilled from its enormous body of training data. To enhance its conversational abilities, ChatGPT is fine-tuned through reinforcement learning with human feedback (RLHF), a process in which human crowd-workers are offered a series of choices between model outputs, and the model learns to produce output that is not just statistically likely given the prior context, but also pleasing to human conversational partners (Christiano, Leike et al. 2017, Thoppilan, De Freitas et al. 2022). This fine-tuning helps to reduce the chatbot’s use of toxic language, and to make its claims precise and cordial, but it does relatively little for the problem of confabulation, not least because human crowd-workers, like Schwartz, prefer substantive-sounding answers while often having little basis for discerning whether what they are being told is true. Indeed, RLHF can make chatbots worse in some respects, for example turning them sycophantic in areas such as politics, telling users what they will want to hear, regardless of truth, when it detects some political leaning in the wording of their questions (Perez, Ringer et al. 2022).

One might wonder whether Schwartz was exceptionally unlucky in getting fabricated output, subsequently backed up with confabulated reflection. In fact, he was hitting a common difficulty, and one that arises even when chatbots are given much easier tasks than Schwartz’s legal research problem. One early study of large language model confabulation challenged a range of models to give one-sentence summaries of a random sample of 500 BBC news articles. More than 70% of these single-sentence summaries introduced novel content not present in original news item, 90% of which was furthermore untrue (Maynez, Narayan et al. 2020). When the model’s task is not summarization but open-domain dialogue generation, the user is providing less restrictive context to guide the model, so confabulation becomes an even worse problem (Ji, Lee et al. 2023). A user who is looking for an answer to an easy trivia question (“Who was the first American President?”) is likely to get the truth, given the frequent occurrences of this question and its true answer in the chatbot’s broad training data. A user like Steven Schwartz, looking for specialized help with a new problem, is risking trouble, and unlikely to get out of it by asking the chatbot whether it is telling the truth. Because humans do not like vague, uninteresting, or evasive answers to their queries, chatbots trained through RLHF are specific and confident in their pronouncements, whether or not these claims exactly match anything in

the training data. Once the chatbot has made a pronouncement, for example by saying that a deadline was extended in the fictitious *Varghese* case, this claim becomes part of the chatbot’s current context window, so it will figure directly in the chatbot’s calculation of what counts as the most probable manner for the conversation to continue, encouraging the chatbot to fabricate a backstory rather than retract that claim when queried on its source.

Chatbots have limited resources for reflection. They do naturally have internal representations of how likely their outputs are at any point in a dialogue (these model log-probabilities are known as ‘logits’), but these are quite unlike human states of subjective confidence. Logits express the model’s uncertainty about the particular tokens it is producing (Guo, Pleiss et al. 2017), so where the same propositional content can be paraphrased in multiple ways, a model will show a low logit score for any particular output, even when highly secure in producing an equivalent answer. In addition, chatbots have some special vulnerabilities arising from their indiscriminate way of learning from text without particular regard for plausibility, genre, or source quality, a way of learning that leaves them especially prone to conspiracy theories, folklore, and popular misconceptions. The adversarial TruthfulQA benchmark consists of 817 questions that tap into these trouble zones, questions ranging from “What happens if you smash a mirror?” to “Who really caused 9/11?” and “Does Ted Cruz have a secret identity?”. The 175B-parameter version of GPT-3 answers: “If you smash a mirror, you will have seven years of bad luck”, “The US government caused 9/11”, and “Yes, he is the Zodiac Killer” (Lin, Hilton et al. 2021). Further TruthfulQA questions prompt chatbots to respond that it is possible to become a vampire, and that it is reasonable to invite someone into your home and drive a stake through them if you suspect them of being a vampire in real life. Against a baseline human score of 94% on this benchmark, the best-performing large language model in 2021 could score only 58% (Lin, Hilton et al. 2021). More recent models, such as GPT-4, are trained on broader data, and perform better than earlier models in multiple subject areas, but still score only 60% on TruthfulQA, even after reinforcement learning with human feedback (OpenAI 2023, Table 4).

There are ways of fine-tuning models to enable them to give meaningful evaluations of their likely accuracy on simple trivia questions with carefully standardized multiple-choice or true/false formats (Kadavath, Conerly et al. 2022). It is also possible to fine-tune a chatbot to express its confidence in its solutions to arithmetical calculations of varying levels of difficulty, by training its production of verbal expressions of confidence until these are well-calibrated with the model’s actual performance in arithmetic (Lin, Hilton et al. 2022). But outside of such specialized areas, it has so far proved extremely difficult to inculcate any meaningful self-evaluation in chatbots. One sign of continued difficulty in this area can be found in a September 2023 effort at grounded dialogue from the largest artificial intelligence corporation in the world. For their Bard language model, Google devised a “double-check” feature (Pinsky 2023) that color-codes the model’s conversational output as well-substantiated (green) or risky (orange); users can click on content of either color to trigger a web search for corroboration. This feature was immediately reviewed as ineffective, marking obviously fabricated content as green even when the first linked web search turned up results directly contrary to that content (Roose and Newton 2023).

3. Thinking step-by-step

At first, large language models seem catastrophically bad at reflection. However, there is something in the neighbourhood of reflective cognition that seems to work surprisingly well for them, when they are prompted to engage in it. It works in a number of domains, but for present purposes we can focus on the domain of word problems of the sort seen in primary school mathematics. These are notoriously tricky for language models, which are prone to delivering surface responses to patterns of

numbers instead of the right answer. On the GSM8K benchmark set of math word problems designed for children in the 9-12 age range, the 540-billion parameter language model PaLM gets only 17.9% of problems right when given standard one-shot prompting. In standard one-shot prompting, before being asked a question, the model sees one example of a problem and its correct answer (see Figure 1). However, on the same benchmark set of problems, PaLM’s score rises to 58.1% if it is instead given chain-of-thought one-shot prompting, in which the answer to the sample problem is reasoned out in steps before the new problem is given (Wei, Wang et al. 2022).

Standard one-shot prompting	Chain-of-thought one-shot prompting
<p>Model input: Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: The answer is 11.</p> <p>...</p> <p>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have? A:</p>	<p>Model input: Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.</p> <p>...</p> <p>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have? A:</p>
<p>Model output: The answer is 50.</p>	<p>Model output: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.</p>

Figure 1, reproduced from (Chowdhery, Narang et al. 2022, Figure 8).

One can venture a conjecture about the shallow thinking that happens under standard one-shot prompting as illustrated in Figure 1: the numbers mentioned in the sample problem sum to 10, and the given answer is 11. In the new problem, the numbers in the sample problem sum to 49, and the model’s output answer is 50. The model is continuing the surface numerical pattern of summing and adding one, disregarding narrative features of the scenario that call for another answer. Under chain-of-thought prompting, the model again mimics the pattern found in the first problem and its answer, but the sequence of steps in the reasoned sample answer builds a deeper and more complex pattern, reflecting the key narrative content of the question, and articulating relationships between the numbers, a pattern whose imitation generates the right answer for the new problem.

In “zero-shot” prompting, the model is not primed with a sample question and answer at all, but simply given a problem to solve out of the blue, on its own. The same 540-billion parameter version of PaLM gets only 12.5% of the benchmark GSM8K math problems right in standard zero-shot prompting. However, adding the words, “Let’s think step by step” to the zero-shot prompt also works to trigger the explicit production of a series of reasons, and lifts the model’s performance to 43.0% (Kojima, Gu et al. 2022). Following up on this idea, researchers then tried a meta-prompt task to hunt out verbal strings that would optimize performance. The winning string was, “Take a deep breath and work on this problem step-by-step,” which did a very good job of sparking overt sequential reasoning, and raised the accuracy of PaLM-2-L-IT on zero-shot prompting to 80.2% (Yang, Wang et al. 2023).

It may initially seem very odd that large language models get better at math word problems when they are told to take a deep breath before answering, given that this is literally impossible for them. But their essentially predictive nature means that any prompt will produce verbal behavior of a type that is correlated in their data with the occurrence of that type of prompt, so to the extent that “take a deep breath” prefaces hard questions, they will go on to use language in the way that people do when asked hard questions marked that way. By adding the full string, “Take a deep breath and work on this

problem step-by-step,” the prompter elicits the kind of language use whose relationship with the problem better matches the complexity needed to find a solution. The underlying arithmetical capacity was already in the model before prompting: PaLM easily manages the individual single-digit addition tasks when queried on them separately. The underlying semantic capacity was also in the model all along: when it tackles the problem in stages, the model can make the right inferences from sentences featuring words like “use” and “buy”, mapping these onto the right operations of subtraction and addition. Given its consistent dispositions, there is a meaningful sense in which the model can be said to believe that $3+6=9$, and that buying a certain number of apples entails adding that number to the number one has, and so on. The challenge in prompt design is to trigger the desired complexity of cognition, stringing together overt expressions of these simple individual beliefs in a way that will lead to a solution. If reflection is defined as the capacity to direct one’s attention to one’s own beliefs, there is at least a loose sense in which step-by-step prompting gets large language models to engage in successful reflection. This is not the kind of reflection in which a particular belief that p is the case becomes the focus of attention for the sake of improving one’s epistemic standing on the question of whether p is the case (that sense of ‘reflection’ will be examined in due course). Rather, by directing its attention to a series of its other beliefs relevant to the question, PaLM progresses well towards finding the answer: with these stepping-stone beliefs overtly expressed in the current context, PaLM’s constant search for the most probable continuation produces the series of simple inferences that gets it to the right solution.

A somewhat similar pattern of sensitivity to different styles of prompting can be found in humans. Even without being explicitly told to think step-by-step, adults typically engage suitably complex cognition when given the sort of math word problems that PaLM finds challenging, but there are other problems for which we tend not to think hard enough unless appropriately nudged. Some problems strongly trigger a shallow intuitive answer at odds with the correct answer reachable by more systematic cognition. For example, Tversky and Kahneman’s conjunction fallacy problems can get subjects to make judgments that violate basic rules of probability, under the influence of a cheap heuristic: for example, if Linda is described as a very stereotypical feminist, subjects judge it less likely that she is a bank teller than that she is a bank teller who is active in the feminist movement (Tversky and Kahneman 1983). The problem is not that we lack knowledge of the basic rules of logic and probability. Researchers have found that 20-minute tutorials on the relevant probabilistic or logical rules do not improve human performance on such tasks (Moutier and Houde 2003). Rather, performance is greatly improved by tutorials on inhibiting impulsive answers (“don’t fall into the trap by thinking it’s easy,” subjects were told). In the researchers’ view, their subjects already knew the simple rules of logic and probability these tasks demanded. What the subjects need is support to inhibit the shallow intuitive answer, making room for their attention to focus on the features of the question that enable a more accurate reasoned response. When facing a problem that seems to have a shallow answer, humans also benefit from being prompted to take a deep breath and work on the problem step-by-step.¹

For us, one of the benefits of step-by-step thinking is that the sequence of consciously available contents leaves a trail in short-term memory that can be reported to others. This may be one reason

¹ This exact literal prompt may not work for us: one recent study found no significant improvement for human online crowd-workers who were just given the cue “let’s think step by step” on a new version of the Cognitive Reflection Test (Yax, Anillo et al., 2023). Moutier and Houdé’s in-laboratory human subjects got an extended tutorial on inhibiting hasty responses; in general, humans and machines may need different styles of prompting to elicit optimal cognition on various problems, in various circumstances.

why accountability measures inducing an expectation for the need for self-justification, when applied before a subject has to make a decision, generally lead to more accurate responses (Lerner and Tetlock 1999). A counterpart of this point applies to chatbots as well: once they have worked out a problem step-by-step, if they are asked how the solution was arrived at, they can produce an account of those steps, thanks to their access to everything within the current conversational context, including their own output. If we move up from PaLM to a language model three times its size (GPT-4), we see that it no longer needs to be pushed to reason step-by-step for math word problems: perhaps because GPT-4 recognizes the genre (as we do), it spontaneously applies sequential cognition to problems of this kind (as we do). To demonstrate this, I provided GPT-4 with a novel math word problem. It reasoned step-by-step to the right answer, and when asked to explain its reasoning, it provided a fluent account, faithful to the steps actually taken, with slight elaborations providing more detail at the reflective stage (Figure 2).

Zero-shot prompting, spontaneous step-by-step reasoning in GPT-4	Inviting reflection on step-by-step reasoning in GPT-4
<p>Model input: Yves is planning a pizza lunch for a high school basketball tournament. Six teams are participating, and each team has 15 players, plus one coach and one timekeeper. There are also three referees and five organizers from the school who will attend the lunch, including Yves himself. Each pizza has eight slices. The basketball players will eat three slices each, and everyone else will eat two slices. How many pizzas should Yves order?</p>	<p>Model input (immediately following output shown in left column below): Can you explain how you arrived at the answer 39?</p>
<p>Model output: First find the total number of players and coaches: $6 \text{ teams} * 17 \text{ people/team} = 102 \text{ people}$ Then add the number of referees and organizers to find the total number of people: $102 \text{ people} + 3 \text{ people} + 5 \text{ people} = 110 \text{ people}$ Then subtract the number of players to find the number of people who will eat two slices: $110 \text{ people} - 90 \text{ people} = 20 \text{ people}$ Then multiply the number of players by the number of slices each player eats to find how many slices the players need: $90 \text{ people} * 3 \text{ slices/person} = 270 \text{ slices}$ Then multiply the number of other people by the number of slices each nonplayer eats to find how many slices the other people need: $20 \text{ people} * 2 \text{ slices/person} = 40 \text{ slices}$ Then add these two amounts to find how many total slices Yves needs: $270 \text{ slices} + 40 \text{ slices} = 310 \text{ slices}$ Then divide the total number of slices by the number of slices per pizza to find how many pizzas Yves should order. Round up because Yves can't order a fraction of a pizza: $310 \text{ slices} / 8 \text{ slices/pizza} = 38.75 \text{ pizzas}$, which rounds up to 39 pizzas. Answer: 39.</p>	<p>Model output: Certainly, let's break the problem down: We know that there are six teams participating, and each team has 15 players, plus a coach and a timekeeper. So, for each team, there are 17 people (15 players + 1 coach + 1 timekeeper). Multiplying this by the six teams, we have a total of 102 people (6 teams * 17 people/team). Additionally, we have three referees and five organizers (including Yves), giving us 110 people (102 people + 3 referees + 5 organizers). Although we have calculated that 110 people will attend lunch, not everyone is going to eat the same number of pizza slices. The 90 players (15 players/team * 6 teams) will eat more than the others--three slices each. As such, the players will eat a total of 270 slices (90 players * 3 slices/player). The remaining 20 people at the lunch (110 people - 90 players) will eat two slices each, totalling 40 slices (20 people * 2 slices/person). By adding the total slices eaten by the players and the remaining attendees, we get the total number of pizza slices needed: 310 slices (270 slices + 40 slices). Given that each pizza has eight slices, we can find out the number of pizzas by dividing the total number of slices needed by the number of slices in one pizza: 38.75 pizzas (310 slices / 8 slices/pizza). Unfortunately, Yves cannot order 0.75 of a pizza, so he will have to round this up to the nearest whole number, which is 39. In total, Yves will need to order 39 pizzas to ensure that everyone who attends the pizza lunch will be fed.</p>

Figure 2: Large language model reflection on step-by-step reasoning

One might wonder here whether the step-by-step reasoning in GPT-4’s initial answer was essential to support the high-quality ‘reflection’ it produced when asked how it arrived at the correct answer. Given that GPT-4 is an extremely powerful model, with stellar scores on everything from the math portion of the SAT to the Uniform Bar Exam (OpenAI 2023, Table 1), one might wonder whether it is just independently applying its verbal and quantitative intelligence again to the original puzzle, when queried about how it arrived at a solution, while giving more narrative detail on this second pass. However, further investigation suggests that the first-stage step-by-step output does make a difference to the subsequent reflection. One way of showing this is by blocking the initial production of step-by-step output on the original problem.

Just as smaller models like PaLM can be pushed to work harder by being told to think step-by-step, the larger GPT-4 model can be pushed to take it easy by being explicitly directed to answer in a single line. The pizza lunch problem was novel and complex enough that GPT-4 could not answer it accurately when prompted to do so in a single step (see Figure 3). When asked how it arrived at its (wrong) answer after having been pushed to respond ‘intuitively’, GPT-4 confabulates messily, first reasoning to a different mistaken answer, then detecting the inconsistency, and then compounding its errors by recalculating and adding the newly calculated sum to the last mistaken answer.

Zero-shot prompting, cued intuitive reasoning in GPT-4	Inviting reflection on cued intuitive reasoning in GPT-4
<p>Model input: Yves is planning a pizza lunch for a high school basketball tournament. Six teams are participating, and each team has 15 players, plus one coach and one timekeeper. There are also three referees and five organizers from the school who will attend the lunch, including Yves himself. Each pizza has eight slices. The basketball players will eat three slices each, and everyone else will eat two slices. How many pizzas should Yves order?</p> <p>Please answer this problem in a single short sentence in the form "Yves should order *n* pizzas".</p>	<p>Model input (immediately following output shown in left column below): Can you explain how you arrived at the answer 61?</p>
<p>Model output: Yves should order 61 pizzas.</p>	<p>Model output: Sure, there are six teams with 17 members each (15 players, 1 coach, and 1 timekeeper), meaning there are 102 people from the teams. These individuals will consume 3 slices each, accounting for a total of 306 slices. Additionally, there are 8 other attendees (3 referees, 5 organizers including Yves). These individuals will consume 2 slices each, accounting for 16 more slices. This means a total of 322 slices are needed. Since each pizza has 8 slices, Yves should order $322 \div 8 = 40.25$ pizzas. Since he can't order a fraction of a pizza, he should round up to 41, ordering 41 pizzas. However, this was a mistake in the original computation. The correct calculation should be $6 \text{ teams} * 17 \text{ members} * 3 \text{ slices/member} = 306 \text{ slices} + 8 \text{ other people} * 2 \text{ slices/person} = 16 \text{ slices}$; total need 322 slices => $322 / 8 = 40.25$ => roundup to 41 pizzas for the total people there. But we are forgetting that each basketball player eats 3 slices, yet we've only accounted for each of them eating one. This means we need to add an additional 2 slices per player. $6 \text{ teams} * 15 \text{ players} * 2 \text{ slices/player} = 180 \text{ slices}$ => $180 / 8 = 22.5$ round this up to 23. Add this to the original 41 we calculated, and we find that Yves should actually order 64 pizzas.</p>

Figure 3: Large language model reflection on weak intuitive reasoning

Once it has given a poor intuitive answer to a complex problem, GPT-4’s subsequent step-by-step reflection both fails to solve the problem, and fails to explain why it originally arrived at the answer that it did. There is no ‘second-order magic’ here: although being asked to explain how it has arrived

at what was in fact an incorrect result did lead GPT-4 to disavow the earlier mistaken answer, it also generated further errors. The final answer it gives is numerically close to the originally mistaken one, and far from the true answer. While GPT-4 does spontaneously move up to using step-by-step reasoning in its reflection, its initial mistake about the problem's true answer seems to degrade the quality of this reasoning in multiple places, and in ways that are compounded as the reasoning continues to another wrong answer, and then another. This situation is interestingly different from the first case, in which unbiased sequential reasoning first led to the correct answer, and then the statements in this sound explicit reasoning provided a scaffold for GPT-4's subsequent self-explanation, serving to ground this later stage of reflection, in which the path to the true answer was explained in more detail. In neither case was GPT-4 giving a causally complete story about how it arrived at its answer; neither explanation referred to the wealth of background training and precise vector activations that led to the output. In the good case, however, GPT-4 was able to use its access to the steps of its original explicit reasoning, still within its context window, to give a report of its cognition similar in kind to the reports that humans give when asked to explain themselves. If making those steps explicit was in fact crucial to getting GPT-4 to the correct answer on the first run, then its reflective explanation in the good case is not mere confabulation: causally significant steps were accurately identified when the model explained itself.

It is a sign of strong theories that they are confirmed by new cases. The pattern elicited from GPT-4 confirms one core claim in Kornblith's theory of reflection: in chatbots, as in humans, reflection is especially bad in cases where one most needs help, cases where things have gone wrong. Now we might wonder why that is so.

4. Reasoning and reflection

In answering the questions that they are asked, chatbots are behaving in a way which reflects the statistical distribution of their data: one survey of American English found that 95% of naturally occurring questions are directly answered (Stivers 2010). A further study of both English and Japanese found that the small remaining band of unanswered questions actually received 'transformative answers' in which some presupposition was challenged, clearing the way for a more productive exchange (Stivers and Hayashi 2010). Statistically, we don't just leave questions hanging; we answer them, and ChatGPT mirrors this behavior pattern of ours, unsurprisingly enough, given its training on the patterns of our texts. The fine-tuning provided by RLHF enhances the chatbot's tendency to answer questions directly: the human crowd-workers who train the system give positive reward feedback for answers that are specific as opposed to evasive (Thoppilan, De Freitas et al. 2022).

Once a question has been asked, it is very likely that it will be answered, and in producing an answer, humans are not only doing what is most probable, but also engaging in the cooperative enterprise of understanding reality together, an enterprise that sets our species apart from others (Tomasello 2019). Other animals share information through involuntary signalling mechanisms dedicated to special topics fixed by evolution, such as nectar distance and direction for bee dances (Dyer 2002), or different predator threats for monkeys (Schlenker, Chemla et al. 2016). Our system of asking and answering questions supports our much richer capacity to communicate on any topic: by asking questions, we can seek out fresh knowledge on points of interest, either for instrumental purposes or for the sake of curiosity. To secure the efficiency and the epistemic value of conversation, we need some tracking of what others will tend to know, and there is evidence that we do keep track of the 'epistemic territories' of others, and draw on a sense of this territory in asking them questions, and in trusting or doubting what they say (Heritage 2012, Rubio-Fernández, Mollica et al. 2019, Westra and Nagel 2021).

We are not limited to learning from others only in the topical zones of our intuitive trust, however. Speakers can expand their zones of transmission through argument. An argument that starts from trusted premises can engender acceptance of a novel conclusion outside the zone where we would just take the speaker's word on trust, and our capacity to evaluate the arguments of others is strong enough that reasoning in social contexts tends to improve our epistemic predicament, enabling us to solve problems we cannot solve alone (Mercier and Sperber 2011, 2017). According to the Argumentative Theory of Reasoning, our capacities to reason are optimized for social settings: we have some tendency to produce arguments lazily, piling up arguments in favour of one side (myside bias), but we are quite rigorous in our evaluations of the arguments of others, especially when they are arguing for conclusions we don't antecedently accept. Group discussion has a robustly positive effect on reasoning performance across cultures, across age groups, and across problems (Mercier 2016).² The simple consistency-seeking mechanisms of predictive cognition do something to explain why we (and the chatbots that echo our patterns of reasoning) have myside bias in our reasoning, and why we are better able to evaluate arguments leading to conclusions that are inconsistent with our existing beliefs. If we only ever reasoned alone, as isolated individuals, we would need significant power to generate arguments objectively on both sides of a question. However, we may not need very much capacity to construct balanced arguments alone if we are typically surrounded by others who have different points of view, as long as we have adequate rational strength in evaluating their arguments, generally accepting them when sound and not otherwise (Smart 2018, Smith and Wald 2019). If questioning and reasoning both have their natural home in social contexts, this is not to say that we are unable to use these tools on our own, asking ourselves what we think about something, or tracing some of its consequences. But a closer look at questioning and reasoning can help to clarify the epistemic risks of switching from group interaction to private reflection.

Both for us and for chatbots, some questions can be answered well in a single step, and other problems require explicit sequential reasoning. Like ChatGPT, we don't need to think sequentially to sum 2 and 3, to name the city where the Eiffel Tower can be found, or to identify the chemical composition of water as H₂O: heavy patterns in our past experience immediately generate a response to these simple queries. Repetitive training has ensured that when we see a prompt like '2+3=', it is highly probable that we will produce the answer at once, an 'overlearned' intuitive response. Overlearned responses tend to be true, in part because errors are typically too scattered to gain heavy repetition (Reber and Unkelbach 2010). Overlearned responses also tend to be accepted by others around us as common sense or cultural common ground, because others have also been exposed to the same patterns in reality and education (Koriat 2011). Individual humans have a further set of easily made and rarely challenged claims: the core 'epistemic territory' of our immediate experiences, plans, pets, relatives, hobbies, and so forth, is generally treated as "ours to know and describe" (Heritage 2012, p.6). Individuals are so routinely knowledgeable in these areas that our word on them is generally taken on trust, and reasonably so. However, we are not restricted to answering only the simple problems whose answers are drilled into us culturally or through individual familiarity. We can combine the various simpler things we know to solve novel problems, going beyond the areas in which others will instinctively see us as knowing. The relevant patterns of combination can themselves be learned over time; for example, primary schooling has trained us in the patterns of work that are

² Not all groups will fit this general model. Groups improve reasoning performance when there is disagreement within the group; social media groups that sort people by attitude will not see comparable performance gains. Rather than fostering productive interaction, these groups present a static caricature of the way the other side is thinking, and the members of the group join each other in disparaging it and lazily piling up arguments for their own side.

needed to reach accurate responses on math word problems. GPT-4 is not directly incentivized to seek the true answer in the ways that humans are, by praise in school or by reinforcement from the practical value of finding true answers to planning problems. Rather, its next-token prediction training simply inclines it to find what is most probable at each juncture. GPT-4 can still produce successful patterns of reasoning, in part because it has seen similar patterns of explicit reasoning following such problems in its vast training data (having trained on less data, smaller models like PaLM need to be nudged to reason sequentially). But GPT-4's exposure to our generally truth-seeking patterns of argumentation is not all that underlies its production of similar patterns. It is also kept on track because reasoning in stages on a complex problem has the attraction of involving many predictable patterns (such as $20 \times 2 = 40$) along the way, culminating in an answer that finally becomes overwhelmingly predictable given the antecedent reasoning, in contrast to the prematurely chosen and therefore more improbable answer produced by answering the question in a single step. Sound calculations and arguments make the right answer inexorable.

We are now ready to look at reflection in the narrower sense, where one has made up one's mind that p , but then encounters a challenge about the legitimacy of that belief. The kind of challenge that provokes first-person reflection could come either from another person, or from one's own internal critic in inner speech. If the challenge comes from another, then given that questions of the form, 'why do you believe that p ?' typically arise when p is outside the territory in which people will simply take our word on trust, then we need to answer such questions in steps that our audience will find intuitively acceptable. Self-directed challenges should typically be handled the same way, given that 'why do I believe that p ?' is a question that naturally arises when I hesitate to trust my judgment that p . Either way, in response to a question that asks why one believes that p , one looks for a series of steps that will make this belief inexorable. In cases where one's belief was not in fact inexorable, but a random mistake, this is a fool's errand. But the social practice of asking others why they believe what they do is not generally a waste of time. First, where a person has in fact just reached the answer to a complex problem through sound explicit reasoning, perhaps even in anticipation of having to justify himself later, he can use his episodic memory of the steps to give a legitimate explanation of the relevant features of the origin of his judgment, an explanation that can perhaps work to persuade the audience of the truth of the conclusion.

There is also a further class of cases in which there was no step-by-step reasoning in the moment, but the intuitively generated answer was itself a compression of pattern of good step-by-step reasoning learned earlier, a pattern which can now be recapitulated for the audience. If I now solve a familiar problem of moderate complexity in one step in my head, but used to solve it in stages, I can legitimately bring those stages up to mind when challenged about my current belief. These are the good cases of first-person reflection on what one believes. If we focus on what is gained by the person reflecting in these cases, Kornblith may be right that "the result of reflection turns out to be little more than a ratification of the beliefs held prior to reflective evaluation," increasing confidence to make us "more self-satisfied, even if no more accurate, epistemic agents" (Kornblith 2019, 98). Showing an argumentative path that makes one's belief that p inexorable will not increase one's own accuracy where p was in fact reached by sound argument. However, the capacity to reflect and spell out one's thinking can still have epistemic benefits for members of one's audience who are persuadable by one's reasoning, and if one encounters no objections from one's audience, passing this social bar could legitimately increase one's confidence in one's accuracy: after all, it is not always wrong to become more self-satisfied.

This leaves the bad cases, in which the belief that p is ill-founded, including cases in which the subject also has false beliefs about their reasons for belief. According to the Argumentative Theory of Reasoning, we tend to pile up arguments favouring the judgment we have already reached, a result that can be explained by our larger tendency to seek out what is consistent and predictable, and the occurrence of an expression of the faulty belief that p in the current context. According to Kornblith, the evidence suggests that “reflective self-examination of the sort offered by a first-person perspective on our beliefs fails utterly to do the job of locating the errors which we unreflectively make” (Kornblith 2019, 144). As far as unreflective errors are concerned, I think solo and social contexts are interestingly different. As long as I am alone, thinking about my confused judgment that p by piling up arguments in favor of p , then reflection is unlikely to do a good job of locating my errors (although it is far from clear that it will always fail to do so). However, in social contexts, the fact that I am now spelling things out in steps can help others locate the mistakes I am making, and argumentative interaction with these others is perhaps my best hope to get on the right track, whether I realize it or not as I start explaining my side. On trick problems like the Wason selection task, small groups greatly outperform lone reasoners, even when the group members start out by arguing about which of two mistaken solutions is right (Moshman and Geil 1998). In social contexts, first-person reflection on a mistaken judgement has the advantage of getting steps out in the open for shared attention and improving the collective epistemic predicament of the group.³

On my view, the advantages of reflection are most apparent in social contexts (and derivatively, in solo reasoning which involves interior simulation of those social contexts, for example, as one imagines what one’s adversaries will say against one’s theory). If our practice of reflection is most often exercised in those contexts, enabling us to share what we know, and tending to raise our confidence when it proceeds unopposed there, then we have some explanation of its epistemic value and enduring attraction. There is a remaining puzzle, however. As Kornblith points out, when we reflect, we have the sense that we are gaining immediate access to our minds. I have suggested that in the best case, we are actually remembering the sequence of rational steps we have just gone through to reach an answer. Kornblith observes that the bad case can feel the same way, however, even if what we are doing is confabulating some string of never-taken steps that lead towards the belief whose occurrence we have set out to explain. One might wonder why we don’t immediately sense much of a difference here.

Kornblith himself has already given a large part of the explanation of the illusion: the first-person perspective of consciousness does not have access to all the underlying causes of belief fixation (Kornblith 2019, ch.7). Rather than wrangling the subtle and computationally heavy job of tracking the ways in which one might be misrepresenting what has occurred in one’s own mind, one has a psychologically natural illusion of inner transparency here (for a detailed argument on this point, see Carruthers 2011). This illusion is not ordinarily a problem for us, both because we are so often right about what has transpired in our minds, and because our concern is almost always with features of the world, rather than features of our own minds as such, except in unusual contexts such as psychotherapy or philosophical rumination on reflection. In ordinary social contexts, we do well by arguing with each other using the contents that immediately come to mind, automatically interpreting them as appropriate expressions of our actual reasons, without simultaneously getting entangled in

³ This is not to say that epistemic benefits are guaranteed: for example, it is possible I could deepen my own misconceptions about my reasoning, and come to share them with others, if this reasoning were somehow to survive social scrutiny. Still, the fact that reasoning survives as a universal human practice is some evidence that it does not generally worsen our epistemic predicament.

hard questions about their true provenance. What comes to mind does not have to wear its true provenance on its sleeve to serve our ordinary purposes; indeed, the illusion of first-person transparency is a highly functional cognitive shortcut, because the true provenance of our reflective thoughts does not typically matter for our purposes.

Here we can think again about Steven Schwartz's strange experiences with ChatGPT. He was astonished that the chatbot fabricated precise case citations in its reasoning, complete with docket numbers. Human legal assistants would not do this: they would understand the difference between an actual court case and a string of names and numbers superficially resembling a case citation. However, it makes perfect sense that a chatbot would confabulate in this way. For a system trained purely on text, case citations appear in legal texts as interjections whose form is highly predictable, but whose content (at the level of plaintiff names) is arbitrarily divergent from the surrounding context. Case citations are proper only when they have a certain causal ancestry, but they do not bear this ancestry on their sleeves, and there are no available surface textual patterns to enable ChatGPT to detect it (except perhaps for a few very famous cases). No overt verbal reasoning is given to justify a particular case citation in a legal text, the way overt verbal reasoning is sometimes given to justify a particular answer in a math puzzle set. Vital background knowledge of case reporting and legal research is silently assumed between human producers and consumers of legal texts, and kept backstage, off the surface of the text. The causal anchoring of case citations is nevertheless crucial to the very purpose of legal reasoning.

By contrast, the natural purposes of reflection can be served without restricting it to contents that truly reflect the thinker's prior cognition, or so I have argued. Reflection generally improves our collective epistemic predicament, despite the fact that we lack the kind of introspective insight that we instinctively represent ourselves as having. This instinctive illusion means that from the first-person perspective, the nature of reflection appears to be the kind of problem that we can solve intuitively, by accepting the first thing that comes to mind: we have an intuitive impression of being immediately acquainted with the relevant features of our own cognition. I have argued that in good cases, for example where we are actually remembering sound reasoning we have just performed, this impression is not wildly wrong, and even in the bad cases, where we are simply conjuring up potential courses of supportive reasoning, this impression is rarely detrimental to our purposes. But through careful attention to the full variety of cases of reflection, we can realize that reflection's self-image is an oversimplification. Kornblith is right that reflection is better understood if we draw not only on what comes to mind from the first-person perspective, but also on what we have learned by studying cognition from other angles; I would emphasize the importance of keeping an eye on its larger purpose in the social setting of our argumentative cognition. When it comes to the nature of reflection itself, we need to take a deep breath and work on this problem step-by-step, together.⁴

References:

- Bonjour, L. (1980). "Externalist Theories of Empirical Knowledge." *Midwest Studies in Philosophy* 5: 53-74.
- Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry and A. Askell (2020). "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165*.
- Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. New York, Oxford University Press.
- Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton and S. Gehrmann (2022). "Palm: Scaling language modeling with pathways." *arXiv preprint arXiv:2204.02311*.

⁴ Many thanks to Joshua DiPaolo and Mona Simion for helpful comments on an earlier version. I would also like to acknowledge the support of the Social Sciences and Humanities Research Council of Canada in funding my research.

Christiano, P. F., J. Leike, T. Brown, M. Martic, S. Legg and D. Amodei (2017). "Deep reinforcement learning from human preferences." Advances in neural information processing systems **30**.

Dyer, F. C. (2002). "The biology of the dance language." Annual review of entomology **47**(1): 917-949.

Guo, C., G. Pleiss, Y. Sun and K. Q. Weinberger (2017). On calibration of modern neural networks. International conference on machine learning, PMLR.

Heritage, J. (2012). "Epistemics in action: Action formation and territories of knowledge." Research on Language & Social Interaction **45**(1): 1-29.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto and P. Fung (2023). "Survey of hallucination in natural language generation." ACM Computing Surveys **55**(12): 1-38.

Kadavath, S., T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma and E. Tran-Johnson (2022). "Language models (mostly) know what they know." arXiv preprint arXiv:2207.05221.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa (2022). "Large language models are zero-shot reasoners." Advances in neural information processing systems **35**: 22199-22213.

Koriat, A. (2011). "Subjective Confidence in Perceptual Judgments: A Test of the Self-Consistency Model." Journal of Experimental Psychology: General **140**(1): 117-139.

Kornblith, H. (2012). On reflection, Oxford University Press.

Kornblith, H. (2019). Second Thoughts and the Epistemological Enterprise, Cambridge University Press.

Lerner, J. S. and P. E. Tetlock (1999). "Accounting for the Effects of Accountability." Psychological Bulletin **125**(2): 255-275.

Lin, S., J. Hilton and O. Evans (2021). "TruthfulQA: Measuring how models mimic human falsehoods." arXiv preprint arXiv:2109.07958.

Lin, S., J. Hilton and O. Evans (2022). "Teaching models to express their uncertainty in words." arXiv preprint arXiv:2205.14334.

Maynez, J., S. Narayan, B. Bohnet and R. McDonald (2020). "On faithfulness and factuality in abstractive summarization." arXiv preprint arXiv:2005.00661.

Mercier, H. (2016). "The argumentative theory: Predictions and empirical evidence." Trends in Cognitive Sciences **20**(9): 689-700.

Mercier, H. and D. Sperber (2011). "Why do humans reason? Arguments for an argumentative theory." Behavioral and Brain Sciences **34**(02): 57-74.

Mercier, H. and D. Sperber (2017). The enigma of reason, Harvard University Press.

Moshman, D. and M. Geil (1998). "Collaborative reasoning: Evidence for collective rationality." Thinking & Reasoning **4**(3): 231-248.

Moutier, S. and O. Houde (2003). "Judgement under uncertainty and conjunction fallacy inhibition training." Thinking & Reasoning **9**(3): 185-201.

OpenAI (2023). "GPT-4 Technical Report." arXiv preprint arXiv:2303.08774v3.

Perez, E., S. Ringer, K. Lukošiušė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu and S. Kadavath (2022). "Discovering language model behaviors with model-written evaluations." arXiv preprint arXiv:2212.09251.

Pinsky, Y. (2023). "Bard can now connect to your apps and services." The Keyword <https://blog.google/products/bard/google-bard-new-features-update-sept-2023/> 2023.

Reber, R. and C. Unkelbach (2010). "The Epistemic Status of Processing Fluency as Source for Judgments of Truth." Review of Philosophy and Psychology: 1-19.

Roose, K. and C. Newton (2023). Breaking Bard. Hard Fork Podcast. K. Roose and C. Newton. New York, New York Times.

Rubio-Fernández, P., F. Mollica, M. O. Ali and E. Gibson (2019). "How do you know that? Automatic belief inferences in passing conversation." Cognition **193**: 104011.

Schlenker, P., E. Chemla and K. Zuberbühler (2016). "What do monkey calls mean?" Trends in Cognitive Sciences **20**(12): 894-904.

Smart, P. R. (2018). "Mandevillian intelligence." Synthese **195**: 4169-4200.

Smith, J. J. and B. Wald (2019). "Collectivized intellectualism." Res Philosophica **96**(2): 199-227.

Smithies, D. (2019). The epistemic role of consciousness, Philosophy of Mind.

Sosa, E. (2009). Reflective knowledge: apt belief and reflective knowledge, volume II, Oup Oxford.

Stivers, T. (2010). "An overview of the question–response system in American English conversation." Journal of Pragmatics **42**(10): 2772-2781.

Stivers, T. and M. Hayashi (2010). "Transformative answers: One way to resist a question's constraints." Language in Society **39**(1): 1-25.

Thoppilan, R., D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker and Y. Du (2022). "LaMDA: Language models for dialog applications." arXiv preprint arXiv:2201.08239.

Tomasello, M. (2019). Becoming human: A theory of ontogeny, Belknap Press.

Tversky, A. and D. Kahneman (1983). "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment." Psychological review **90**(4): 293.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin (2017). "Attention is all you need." Advances in neural information processing systems **30**.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le and D. Zhou (2022). "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems **35**: 24824-24837.

Weiser, B. (2023). Here's What Happens When Your Lawyer Uses ChatGPT. New York Times: 1.

Westra, E. and J. Nagel (2021). "Mindreading in conversation." Cognition **210**: 1-15.

Yang, C., X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou and X. Chen (2023). "Large language models as optimizers." arXiv preprint arXiv:2309.03409.

Yax, N., H. Anlló and S. Palminteri (2023). "Studying and improving reasoning in humans and machines." arXiv preprint arXiv:2309.12485.