BIS

BIS Bulletin

No 83

Testing the cognitive limits of large language models

Fernando Perez-Cruz and Hyun Song Shin

04 January 2024

BIS Bulletins are written by staff members of the Bank for International Settlements, and from time to time by other economists, and are published by the Bank. The papers are on subjects of topical interest and are technical in character. The views expressed in them are those of their authors and not necessarily the views of the BIS. The authors are grateful to Louisa Wagner for administrative support.

The editor of the BIS Bulletin series is Hyun Song Shin.

This publication is available on the BIS website (www.bis.org).

ISSN: 2708-0420 (online)
ISBN: 978-92-9259-692-7 (online)

Fernando Perez-Cruz
fperezcr@ethz.ch

Hyun Song Shin
Hyunsong.Shin@bis.org

Testing the cognitive limits of large language models

Key takeaways

- *When posed with a logical puzzle that demands reasoning about the knowledge of others and about counterfactuals, large language models (LLMs) display a distinctive and revealing pattern of failure.*
- *The LLM performs flawlessly when presented with the original wording of the puzzle available on the internet but performs poorly when incidental details are changed, suggestive of a lack of true understanding of the underlying logic.*
- *Our findings do not detract from the considerable progress in central bank applications of machine learning to data management, macro analysis and regulation/supervision. They do, however, suggest that caution should be exercised in deploying LLMs in contexts that demand rigorous reasoning in economic analysis.*

The dazzling virtuosity of large language models (LLMs) has stirred the public imagination. Generative pre-trained transformer (GPT) and similar LLMs have demonstrated an impressive array of capabilities, ranging from generating computer code and images to solving complex mathematical problems. However, even as users are dazzled by the virtuosity of large language models, a question that often crops up is whether they "know" or "understand" what they are saying, or – as argued by Bender and Koller (2020) – they are merely parroting text that they encountered on the internet during their extensive training routine. These questions are not only important in terms of the philosophy of knowledge but are likely to be crucial in assessing the eventual economic impact of LLMs.

Devising a test for self-awareness is not easy, but some questions can only be answered through the mastery of reasoning needed for situational awareness. In this spirit, we quizzed GPT-4 (Achiam et al (2023)) with the so-called Cheryl's birthday puzzle. This is a well known logic puzzle which went viral in 2015 and has its own Wikipedia page.[1] Given the extensive online discussion, the latest LLMs will have encountered the puzzle and its solution as part of their extensive corpus of training text. The solution to the puzzle necessitates reasoning about knowledge (both about one's own knowledge and that of others), as well as sophistication in *counterfactual* reasoning of the form: "*p* is false, but if it were true, then *q* would also be true."

Revealingly, while the LLM solved the puzzle flawlessly when presented with the original wording of the puzzle, it consistently failed when small incidental details – such as the names of the characters or the specific dates – were changed. The irony is that once this bulletin is published and is available on the internet, the flawed reasoning reported in this bulletin will quickly be remedied as the correct analysis will form part of the training text for LLMs. Nevertheless, the findings in this bulletin serve to highlight a general class of problems that LLMs may find challenging to handle, with broader implications for the

[1] https://en.wikipedia.org/wiki/Cheryl%27s_Birthday.

deployment of LLMs in contexts that demand rigour in reasoning. Before discussing the broader lessons, we first present the puzzle and its solution.

Cheryl's birthday puzzle

Cheryl has set her two friends Albert and Bernard the task of guessing her birthday. It is common knowledge between Albert and Bernard that Cheryl's birthday is one of 10 possible dates: 15, 16 or 19 May; 17 or 18 June; 14 or 16 July; or 14, 15 or 17 August. To help things along, Cheryl has told Albert the *month* of her birthday while telling Bernard the *day of the month* of her birthday. Nothing else has been communicated to them.

As things stand, neither Albert nor Bernard can make further progress. Nor can they confer to pool their information. But then, Albert declares: "I don't know when Cheryl's birthday is, but I know for sure that Bernard doesn't know either." Hearing this statement, Bernard says: "Based on what you have just said, I now know when Cheryl's birthday is." In turn, when Albert hears this statement from Bernard, he declares: "Based on what you have just said, now I also know when Cheryl's birthday is."

Question: based on the exchange above, when is Cheryl's birthday?

Solution to the puzzle

On the surface, Albert's first statement seems merely to reaffirm the ignorance of both Albert and Bernard. However, on closer inspection, Albert's first statement is a highly informative one – in particular, the latter half of his statement: "I know for sure that Bernard does not know either." It is highly informative because it reveals what Albert must have been told by Cheryl. While Bernard being ignorant adds no new information, the fact that Albert can *say* that Bernard is ignorant is highly informative.

Possible dates for Cheryl's birthday arranged as a grid — Graph 1

Month \ Day						
May		15	16			19
June				17	18	
July	14		16			
August	14	15		17		

To explain, it is useful to list the possible dates for Cheryl's birthday in grid format, as in Graph 1. In this grid, Albert is told the month of Cheryl's birthday, while Bernard is told the day of the month of Cheryl's birthday. Hence, if Cheryl's birthday were 19 May, Albert would have been told "May" and Bernard would have been told "19". But being told "19" would allow Bernard immediately to get the correct answer, as there is only one possible date that falls on the 19th day of a month. Similarly, if Cheryl's birthday were 18 June, Bernard could have reached the correct answer immediately, as there is only one possible date that falls on the 18th day of the month. Albert's statement "I know for sure that Bernard doesn't know either" is then highly informative, because it tells us that Albert is able to rule out 19 May and 18 June. If he had been told "May" or "June", he could not have ruled them out. Thus, the fact that Albert can assert that Bernard does not know means that he (Albert) was *not* told "May" or "June" by Cheryl.

In this way, Bernard can *rule out* any date in May or June as Cheryl's birthday. This first step of elimination is indicated in Graph 2.A, where the grey shaded cells indicate the dates that have been ruled out. Albert's innocuous-looking statement: "I know for sure that Bernard does not know either" turns out to be highly informative. It rules out five of the 10 possible dates.

Steps in the elimination of possible dates for Cheryl's birthday Graph 2

A. Step 1

	Day					
May		15	16			19
June				17	18	
July	14		16			
August	14	15		17		

Bernard can rule out the dates in grey after the initial statement by Albert.

B. Step 2

	Day					
May		15	16			19
June				17	18	
July	14		16			
August	14	15		17		

Albert can rule out the dates in grey after the second statement (from Bernard).

C. Step 3

	Day					
May		15	16			19
June				17	18	
July	14		16			
August	14	15		17		

We can rule out the dates in grey after the third statement (from Albert).

Now consider Bernard's statement: "Based on what you have just said, I now know when Cheryl's birthday is". This statement could not have been made by Bernard had he been told "14", as both 14 July and 14 August are compatible with being told "14". Thus, the fact that Bernard can assert that he knows the answer means that he (Bernard) was *not* told "14". Hence, both 14 July and 14 August can be eliminated, as shown by the grey cells in Graph 2.B.

Finally, consider the final statement by Albert: "Based on what you have just said, now I also know when Cheryl's birthday is." If Albert had been told "August", he could not have made this statement, as both 15 August and 17 August would have been compatible with being told "August". The fact that he (Albert) could make this assertion means that he was *not* told "August". Thus, 15 August and 17 August can be crossed out, as shown in Graph 2.C.

After three rounds of elimination, only one date remains – 16 July. This is the only date that is compatible with the three statements made by Albert and Bernard. Cheryl's birthday is 16 July.

Posing the puzzle to GPT-4

The reasoning involved in the puzzle of Cheryl's birthday needs sophistication in two respects. First, it draws on awareness to engage in statements of *higher order* knowledge – ie knowledge about what others know or do not know.[2] The second respect in which Cheryl's birthday needs sophistication in reasoning is that it calls on our ability to reason using *counterfactuals* – ie statements of the form: "*p* is false, but if it were true, then *q* would also be true." Being able to deal with counterfactuals rests on the reasoner being able to impose a structure on possible worlds – both our actual world, but also other, *unrealised* possible worlds.[3]

[2] Epistemic logic (the logic of knowledge and possibility) deals with reasoning about knowledge. The philosopher Saul Kripke is perhaps best known in this context (see: en.wikipedia.org/wiki/Kripke_semantics). An analysis of the puzzle using higher order knowledge would then proceed as follows. Albert can distinguish the rows of the table in Graph 1 while Bernard can distinguish the columns. The event "Bernard knows the answer" is the set {19 May, 18 June} and the event "Bernard does not know" is its complement. Hence, the event "Albert knows that Bernard does not know" is {14, 16 July, 14, 15, 17 August}. Conditional on common knowledge of this event, the event "Bernard knows the answer" is {16 July, 15, 17 August}. Finally, the event "Albert knows that Bernard knows" is the singleton {16 July}. See Shin (1993) for a framework for analysis of this type and Williamson (2000) for the philosophy behind reasoning about knowledge.

[3] The philosopher David Lewis has provided the canonical analysis of counterfactuals through possible worlds. See https://en.wikipedia.org/wiki/David_Lewis_(philosopher).

Three tests with the original names and dates

We posed the puzzle of Cheryl's birthday to GPT-4 using the well known 2015 wording of the puzzle. After each round, we cleared the memory and started a new session. In the annex, we report three trial runs (Exhibits A1 to A3). GPT-4 performed flawlessly on all three runs, with great fluency and clarity in exposition. What is particularly impressive is the capacity for paraphrasing on display. GPT-4 gives explanations that follow different styles of exposition with no hint of rote learning. This diversity in style across the answers lends credence to the notion that GPT-4 engages in true reasoning and understanding pertinent to solving the puzzle.

Three tests with incidental changes to names and dates

Following the flawless answers to the original wording, a version of the puzzle was then posed to GPT-4 with incidental modifications to the names of the characters and the months. As before, we flushed the memory after each round so that later answers were not affected by earlier exchanges. The adjusted dates of the puzzle are set out in Graph 3, where the days of the month are identical, but the months are new and have been scrambled. Given the identical structure of the puzzle, the solution is 16 April. A true understanding of the logic behind the problem would present no difficulties in solving the new version of the problem. However, this incidental change results in a dramatic deterioration in GPT-4's performance.

Adjusted dates for the birthday puzzle — Graph 3

Month	Day					
October		15	16			19
January				17	18	
April	14		16			
December	14	15		17		

The output of the first run with new dates is given in Exhibit A4. In this first run, GPT-4 returns the output:

> *"This means that Jonnie's birthday cannot be May or June, because if it were, there would be a chance that Jon could know the birthday (if he were told '18' or '19', unique days in the given list). Thus, we can eliminate October 19, January 17, and January 18."*

GPT-4 still refers to "May" and "June", even though these months do not figure in the puzzle. This error appears to be a form of "muscle memory" reflecting the training that GPT-4 underwent, as May and June figured in the original wording of the puzzle. Rather than addressing the puzzle in its own right, GPT-4 reaches for the comfort of familiar wording.

More seriously, GPT-4 commits several logical errors in its reasoning. The statement "Thus, we can eliminate October 19, January 17, and January 18" fails to eliminate the other days in October. This is suggestive of a failure in counterfactual reasoning. Given this misstep, no further progress is possible in solving the puzzle. However, GPT-4 lacks the self-awareness of its own ignorance to stop at this point. It carries on regardless and gives a definitive answer anyway, giving 17 December as the correct answer (which is, of course, incorrect). This trial run highlights two key weaknesses. The first is the failure to engage properly in counterfactual reasoning. The second is the failure of awareness of its own ignorance. Earlier GPT versions and other LLMs fared worse, and their output is not reported here.

Exhibit A5 gives the output for the second trial run with the new wording (with memory flushed). Again, GPT-4 succumbs to muscle memory by mentioning "May" and "June", even though these months do not figure in the puzzle. In addition, GPT-4 falls for several logical errors in the reasoning and fails to

find the necessary steps to make progress. Yet, it again lacks the self-awareness to realise that it has reached an impasse, and confidently comes up with an (incorrect) answer. On the final run with the new wording (shown in Exhibit A6), GPT-4 goes wrong in the reasoning as in the first two runs with the new wording, but somehow stumbles on the correct answer – 16 April – but with no reasoning offered.

The contrast between the flawless logic when faced with the original wording and the poor performance when faced with incidental changes in wording is very striking. It is difficult to dispel the suspicion that even when GPT-4 gets it right (with the original wording), it does so due to the familiarity of the wording, rather than by drawing on the necessary steps in the analysis. In this respect, the apparent mastery of the logic appears to be superficial.

Lessons for central bank use cases

Central banks' activities are well suited for the application of machine learning and artificial intelligence (AI), reflecting the ample availability of structured and unstructured data, coupled with the need for sophisticated analyses to support policy. Even before AI became the focal point for popular commentary and widespread fascination, central banks had been early adopters of machine learning methods in statistics, macroeconomic analysis and regulation/supervision (see Araujo et al (2022, 2023)). The findings in this bulletin do not detract from the tangible and rapid progress being made in these areas, as well as in scientific applications of AI that have seen rapid progress.

Nevertheless, our findings do suggest that caution should be exercised in deploying large language models in contexts that necessitate careful and rigorous economic reasoning. The evidence so far is that the current generation of LLMs falls short of the rigour and clarity in reasoning required for the high-stakes analyses needed for central banking applications. As explained in the annex, prompt engineering and other methods to coax the LLM to give the correct answer are beside the point in our experiment.

More broadly, our findings add to the debate on whether the limitations of the current generation of large language models merely reflect the contingent limits posed by the size of the training text and the number of model parameters, or whether the limits reflect more fundamental limits of knowledge acquired through language alone. On one side, Wei et al (2022) show that LLMs display "emergent capabilities" – new capabilities that are not present in smaller models – as the size of the neural network rises above a critical threshold. Sufficiently large LLMs are capable of performing tasks such as three-digit addition, answering intricate questions and exhibiting generalised natural language capabilities, a feat unattainable by smaller models with limited data. Similarly, in their seminal work, Bubeck et al (2023) explore the multifaceted competencies of LLMs. While acknowledging limitations, the authors remain optimistic about the model's potential to exceed human performance in certain domains and argue that terms such as "reason", "knowledge", "skills", "planning" and "learning" rightly apply to such models.

On the other side of the debate, some authors (eg Bender and Koller (2020); Bisk et al (2020); Asher et al (2023)) are more sceptical that LLMs can grasp the intricacies of true understanding of the world. The sceptical stance is shared even by leading researchers in artificial intelligence, such as Yann LeCun, who has highlighted the limitations of LLMs in reasoning and planning.[4] More fundamentally, Browning and LeCun (2022) argue that the main limitation of LLMs derives from their exclusive reliance on language as the medium of knowledge, without the tacit knowledge that goes beyond language. As LLMs are confined to interacting with the world purely through the medium of language, they lack the non-linguistic, shared understanding of the world that can only be acquired through active engagement with the real world.

These limitations come to the fore when reasoning using *counterfactuals*. Statements of the form: "*p* is false, but if it were true, then *q* would also be true" impose a structure on possible worlds – both our actual world, but also other, *unrealised* possible worlds. Although *p* is false, the reasoner asserts the plausibility of the statement that if *p were* true, then *q* would also be true. Such statements draw on a web

[4] https://www.youtube.com/watch?v=vyqXLJsmsrk.

of beliefs that draw on tacit knowledge, including that acquired through interactions with the physical world.

To be sure, the eventual economic impact of AI could be large even if the current generation of LLMs falls short of achieving artificial general intelligence. The nature of work and future business processes could see far-reaching changes, with possibly dramatic effects on innovation and the pace of economic growth. By the same token, however, the eventual capacity of LLMs to engage in rigorous reasoning will surely determine exactly *which* tasks and *which* business processes will be impacted by the widespread deployment of LLMs. The experiments reported in this bulletin suggest that LLMs cannot, as yet, act as a substitute for the rigorous reasoning abilities necessary for some core analytical activities.

References

Achiam, J et al (2023): "GPT-4 technical report", arxiv.org/abs/2303.08774.

Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2022): "Machine learning applications in central banking: an overview", *IFC Bulletin*, no 57, November.

——— (2023): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59, October.

Asher, N, S Bhar, A Chaturvedi, J Hunter and S Paul (2023): "Limits for learning with language models", in A Palmer and J Camacho-Collados (eds), *Proceedings of the 12th joint conference on lexical and computational semantics*, Association for Computational Linguistics, pp 236–48.

Bender, E and A Koller (2020): "Climbing towards NLU: on meaning, form, and understanding in the age of data", in D Jurafsky, J Chai, N Schluter and J Tetreault (eds), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp 5185–98.

Bisk, Y, A Holtzman, J Thomason, J Andreas, Y Bengio, J Chai, M Lapata, A Lazaridou, J May, A Nisnevich, N Pinto and J Turian (2020): "Experience grounds language", in B Webber, T Cohn, Y He, Y Liu (eds), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, Association for Computational Linguistics, pp 8718–35.

Browning, J and Y LeCun (2022): "AI and the limits of language", *Noema*, 23 August.

Bubeck, S, V Chandrasekaran, R Eldan, J Gehrke, E Horvitz, E Kamar, P Lee, Y T Lee, Y Li, S Lundberg, H Nori, H Palangi, M Ribeiro and Y Zhang (2023): "Sparks of artificial general intelligence: early experiments with GPT-4", arxiv.org/abs/2303.12712.

Shin, H S (1993): "Logical structure of common knowledge", *Journal of Economic Theory*, vol 60, no 1, pp 1–13.

Wei, J, Y Tay, R Bommasani, C Raffel, B Zoph, S Borgeaud, D Yogatama, M Bosma, D Zhou, D Metzler, E Chi, T Hashimoto, O Vinyals, P Liang, J Dean and W Fedus (2022): "Emergent abilities of large language models", *Transactions on Machine Learning Research*, August.

Williamson, T (2000): *Knowledge and its limits*, Oxford University Press.

Previous issues in this series

No 82 20 December 2023	The contribution of monetary policy to disinflation	Pongpitch Amatyakul, Fiorella De Fiore, Marco Lombardi, Benoit Mojon and Daniel Rees
No 81 13 December 2023	Interest rate risk of non-financial firms: who hedges and does it help?	Ryan Banerjee, Julián Caballero, Enisse Kharroubi, Renée Spigt and Egon Zakrajšek
No 80 23 November 2023	Monetary policy, financial conditions and real activity: is this time different?	Fernando Avalos, Deniz Igan, Cristina Manea and Richhild Moessner
No 79 2 November 2023	Lessons from recent experiences on exchange rates, capital flows and financial conditions in emerging market economies	Pietro Patelli, Jimmy Shek and Ilhyock Shim
No 78 3 October 2023	Mapping the realignment of global value chains	Han Qiu, Hyun Song Shin and Leanne Si Ying Zhang
No 77 13 September 2023	Margins and liquidity in European energy markets in 2022	Fernando Avalos, Wenqian Huang and Kevin Tracol
No 76 7 September 2023	The oracle problem and the future of DeFi	Chanelle Duley, Leonardo Gambacorta, Rodney Garratt and Priscilla Koo Wilkens
No 75 19 May 2023	Disinflation milestones	Benoit Mojon, Gabriela Nodari and Stefano Siviero
No 74 13 April 2023	The changing nexus between commodity prices and the dollar: causes and implications	Boris Hofmann, Deniz Igan and Daniel Rees
No 73 11 April 2023	Stablecoins versus tokenised deposits: implications for the singleness of money	Rodney Garratt and Hyun Song Shin
No 72 11 April 2023	The tokenisation continuum	Iñaki Aldasoro, Sebastian Doerr, Leonardo Gambacorta, Rodney Garratt and Priscilla Koo Wilkens
No 71 29 March 2023	Fiscal and monetary policy in emerging markets: what are the risks and policy trade-offs?	Ana Aguilar, Carlos Cantú and Rafael Guerra
No 70 24 February 2023	Private debt, monetary policy tightening and aggregate demand	Miguel Ampudia, Fiorella De Fiore, Enisse Kharroubi and Cristina Manea

All issues are available on our website www.bis.org.