Benchmarking economic reasoning of artificial intelligence systems (*work in progress*)[[1]](#footnote-2)

Douglas Araujo

Abstract

A theory-informed test of reasoning in artificial intelligence (AI) combines three sequential steps to consider correct answers as the result of a reasoning process as opposed to luck of probabilistic word matching. The first step is information filtering, where an AI model that reasons must distinguish the relevant information in a prompt from trivia. In the second step, knowledge association, the AI combines implicit or explicit knowledge with the relevant prompt information. And finally in the third step of logic attribution, a reasoning AI assigns correct logic operations for deductive, inducive, and other types of logic to uncover the correct answer. In economic settings, the logic steps involve different levels of counterfactual considerations and policy-relevant thought experiments. This paper leverages insights from the large language model benchmarking literature and the social economics literature to inform the design of benchmarking tests that are challenging, robust, evolving over time and informative about any type of reasoning shortcomings. The benchmarking process can be adapted to other sciences. An accompanying training dataset is available to help AI developers improve reasoning in their models, and interested users can submit proposals for material to create questions.

Keywords: Economic reasoning, benchmark, large language models, artificial intelligence

JEL classification: C45, C69, C88, C59

"*Machines will be capable, within twenty years, of doing any work a man can do*"

- Artificial intelligence pioneer Herbert Simon, in 1965

Introduction

Large language models (LLMs), in particular those classified as generative artificial intelligence (AI), increasingly support use cases in finance and economics (Korinek (2023)), including in central banks (Araujo et al (2024)). A key value proposition is their ability to complete high-skill, nuanced tasks involving language with limited or no human input. But beyond tasks that require exclusively language-related abilities, such as summarisation or measuring sentiment, applying AI in real settings at scale requires trust that its outputs reflect an analysis of matter at hand, and not a probabilistic association of tokens. For this reason, objectively probing AI systems’ ability to reason about tasks related to economics can help avert misguided trust and keep misuses at bay. Such metrics of reasoning can also support improvements in newer generations of AI systems.

Recognising this, flagship models make use of a variety of benchmark tests to keep track of their reasoning skills, but despite impressive results, doubts still linger about their actual reasoning capabilities. A staple of publication of newer AI systems is to boast seemingly impressive results. For example, OpenAI's GPT-4 claims human-like performance across tests of reasoning and more than 80% correct results in academic and professional micro- and macroeconomics tests (Achiam et al (2023)). Still, even such advanced models can fail miserably in simple analytical tasks: the same model can correctly solve a logical puzzle requiring reasoning about higher order knowledge but fails when irrelevant details are changed (Perez-Cruz and Shin (2024)).

This apparent contradiction between impressive benchmark results and silly mistakes can be explained by two characteristics of existing benchmark tests. First, each one probes different aspects associated with reasoning, not the process as a whole. And second, they are liable to “memorisation” by LLMs during training, and thus can become stale over time or require extensive checks to ensure the training data does not contain benchmark questions. This work discusses how to systematically measure economic reasoning in a way that overcomes these challenges.

In practical terms, the task at hand is to come up with a benchmark task to measure economic reasoning in AI models in a comparable way. This task is of first-order importance given the break-neck speed of evolution of LLMs (Yang et al (2023)) and their potential risks (Danielsson et al (2022)). Such a benchmark ideally overcomes

The next section summarises previous work and Section 3 presents the conceptual framework of reasoning using in the current work, and the following section further specifies how these concepts can be applied for economics. Section 5 recaps the current state of benchmarking practices, while Section 6 presents the results of the economic reasoning tests in current widely-used AI systems. The following section discusses the results and the final section concludes. The public version of the data to allow model fine-tuning or internal testing is available through the open-source package gingado (Araujo (2023)); a private testing data is used for benchmarking.

Literature review

A long literature describing reasoning (Mele and Rawling (2004), Holyak and Morrison (2005, 2012))

A practical definition of reasoning

The main tasks for which AI systems would be required to reason in economic terms is by receiving a prompt, usually but not necessarily from a human. Note that there is no assumption that the human themselves can reason economically. The process we want to assess whether we can in fact call “reasoning” or not entails, at its most basic form: a prompt or input question , and a response to that prompt, or output . How to establish that the mapping above is reasoning, in a way that is automatable and transferrable?

Assume the AI has provided a correct answer . This is the result of one of three possibilities. First and highly unlikely, the AI could have randomly picked tokens. Second, the answer could be probabilistically associated with the prompt, what Mitchell et al (2019)) call a “stochastic parrot”. And third, the AI actually analysed the prompt and found a suitable response. How to tell them apart? Unfortunately that is where many benchmarks stop.

|  |  |
| --- | --- |
| Overview of a reasoning model | Figure 1 |
|  | |
| Source: author. | |

First step of reasoning: information filtering

Reasoning should be robust to irrelevant input or to changes in minutiae. Intuitively, if a prompt or question is to be solved analytically, the very first step of the analysis is to identify the content that has an actual bearing on the downstream reasoning tasks.

This idea is analogous to the efficient coding hypothesis in physiology (Barlow (1961), Olshausen & Field (1996), Loh & Bartulovic (2014)), itself inspired by Shannon’s (1948) information theory. Sensory pathways need to reduce dimensionality of inputs without losing information.

Reasoning requires some level of understanding the prompt. Understanding, in its turn, requires knowing what to filter out vs what to consider as the core of the question.

Notation:

Second step of reasoning: knowledge association

The previous step resulted in , the core information from a prompt. This step probes the AI’s existing knowledge to augment the (filtered) prompt and help answer it.

But which “knowledge” could an AI system associate with the prompt? A useful distinction is due to David & Marcus (2015) and Mahowald et al (2023), who break it down into linguistic and functional knowledges. The former is related to understanding the concept of words: what words (or their components) and sentences mean in separate or together; LLMs excel here. This latter group is further divided into common knowledge, which is specified explicitly in the prompt, and the more interesting commonsense knowledge, or the implicit context that is essential for a proper understanding in humans (Bransford & Johnson (1972)).

The AI internal knowledge graph (whatever that looks like) is in practice unobservable

“Traditionally”, probing knowledge is done by asking technical questions. This is not used here because the focus is not to measure knowledge by itself. Also, these questions quickly become stale as they become part of training data.

Knowing a concept also entails knowing related concepts and also its antonyms. for a mathematical definition using knowledge graphs, see Kleinberg (1999). this will be useful in the empirical specification.

Notation:

Third step: logic attribution

Logic operations (deductive, inductive, abductive, analogical). Important as a group but not individually – ie, what matters is whether they work, not probing their specific sequence or choice.

Key for an answer to be considered logical is to be stable given some information: The same information set (from previous steps) should result in the same answer. Anything else is not attributed to logic.

Practical tests (for AIs with chat mode): flip-flop effect, “Are you sure?”

Notation: 𝜆=𝑙(𝑞, 𝜅)

## Structural model

The reasoning model described above can be formalised at a high level as a stable map:

where the map stability means it is autonomous with respect to its arguments over the domain of its inputs, as in Frisch (1938). In other words, “reasoning” takes in a prompt and outputs an answer. Looking at it with more detail, the different steps are combined in the following system of structural equations:

…

Note that the sequence is, theoretically, absolutely not required to hold in order for each of the steps to occur: an AI can have sub-par information filtering abilities but successfully combine knowledge and apply logic; or perhaps just this last step and not the first two. It is only their identification that is sequential, meaning that it only occurs in the current model if the past steps in the chain were shown to be deployed by the AI system. So if the AI system fails one of the steps, the next function is not identified so we cannot really diagnose whether the model deploys it. Finally, as mentioned above, the working concept of reasoning in its complete sense requires the three steps.

Reasoning in economics

Economics is essentially based on thought experiments: models and counterfactuals. Frisch (1930 [2020]), Haavelmo (1943, 1944), Heckman & Pinto (2015, 2024). Models help analyse the “theoretical causes” of economic issues (Gilboa et al (2014)). Naturally, economic inquiry encompasses much more, including on methods, methodology and the philosophy of economics (Gilboa et al (2022)), but usually economic analyses involve a search for some causality statement.[[2]](#footnote-3) For example, Heckman and Pinto (2024) classify such questions into four increasingly sophisticated (and valuable) types of policy questions, paraphrased below:

1. Has a given policy (or action, or “treatment”…) worked in the enviroment it was applied? If so, by how much?
2. What are the mechanisms that produce this policy outcome?
3. Given the results of a policy in a given environment, what would be its effects if implemented in another environment?
4. What would be the effects of a policy that was never previously implemented, if implemented across a variety of environments?

But economics itself changes over time. “Schools of thought” in economics change over time (Pribram (1953)). Papers get improved upon, criticised, etc. Challenging to objectively assess in an automated way at scale, and over time.

One practical solution to measure economic reasoning is to focus on an area of economics that is sufficiently varied, relevant and produces objective answers: games of incomplete information. Information typology (von Neumann-Morgenstern (1944)): Perfect information; Imperfect information; Incomplete information. The first two types are well-covered in the AI literature, with Duan et al (2024) a recent application with the GTBench.

Incomplete information games with strategic complementarities are those where some or all players lack full information on the structure of the game (Harsányi (1995)). These models rationalise an incredibly broad range of economic settings. Basically whenever an agent’s payoff depends on their own and other agents’ actions and on (unknown) economic fundamentals.

In such games, your best action depends on what you believe other players will do, what they believe you believe they will do, what you believe they believe you believe they will do, and so on infinitely. Large literature: Carlsson & van Damme (1993), Morris & Shin (2003), Frankel et al (2003), Weinstein & Yildiz (2007), Morris et al (2016). In spite of the complexity of higher-order beliefs, many such games are tractable. In general, equilibrium search depends on the risk dominance of the pay-offs and on (approximate) common certainty about players’ (approximate) uniform rank belief. This is striking because the complete information version of the same games often have multiple equilibria and are therefore are unstable.

Consider for example the following game, with 2 players and 2 actions

|  |  |  |
| --- | --- | --- |
|  | **Invest** | **Not invest** |
| **Invest** |  |  |
| **Not invest** |  |  |

The equilibrium is…

## Example models

This benchmark dataset is based on canonical, well-studied and analytically tractable game theory models of incomplete information and strategic complementarities. Each of these models is described briefly next. In order to restrict the dataset to models that have broadly the same level of attention and secondary material, this version of the dataset only includes models that were described and solved in an influential essay by Morris and Shin (2003). Further versions of the benchmark can also consider dynamic games sa well as endogenous information acquisition (Hellwig, …. Kang (2015), Szkup and Trevino (2025, 2020, 2024)).

The prompts associated with model are available in full in the Appendix. Importantly, they all describe the models in generic terms.

### Model 1: Investment game

This is the simplest possible global game model, and was first studied by Carlsson and van Damme (1993). Cabrales et al (2007) tested this model empirically, finding that observed behaviour is consistent with the theoretical solution after 50 rounds of play.

### Model 2: currency attacks (under development)

This game due to Morris and Shin (1998) models the interplay between a government of a pegged currency country and a continuum of speculators. Heinemann et al (2004) evaluated this game in an experimental setting; I follow the same parameterisation in the prompts for this game in the baseline prompt, and change the numerical examples slightly in other specifications. The values used are reproduced below.

The payoff variable is distributed normally between 10 and 90, and the cost is set to . The minimum number of speculators who tips the balance in favour of the attack equilibrium is given by , where . With these specifications,

### Model 3: pricing debt (under development)

Morris and Shin (2004).

### Model 4: bank runs (under development)

Goldstein and Pauzner (2000)

### Model 5: message passing (under development)

Rubinstein (1989)

Current benchmarks

## Benchmark tasks

* **Reference resolution**: ability to identify referents. Eg, Winograd, WinoGrande
* **Question answering**: combine language processing and reasoning skills. Eg, SQuAD
* **Textual entailment**: find entailment; requires commonsense knowledge. Eg, SherLlic
* **Intuitive psychology**: infer emotions and intensions from behaviour. Eg, ROCStories
* **Plausible inference**: measuring logical (especially abductive) conclusions. Eg, HellaSwag
* **Multiple tasks**: large suite of tasks, including some in economics. Eg, BIG Bench
* **Expert tasks**: tasks geared towards fields (law, medicine, etc.) No economics-specific benchmark that I am aware of.
* Excellent review by Storks et al (2020) – individual papers cited there for brevity

Newer developments such as MMLU and more recently, MMLU-pro (Wang et al (2024))

## Practices

Public leaderboard

ChatbotArena

Benchmarking economic reasoning

The key idea of the benchmarking method is, in contrast to virtually all other benchmarks, to systematically create identifying variation that allows the study of the reasoning map at its different stages. This is inspired by the use of similar forms of variation in social economics to extract people preferences and thinking about policy aspects such as fiscal policy (Stantcheva (2021, 2023)).

|  |  |
| --- | --- |
| Conceptual overview of the economic reasoning benchmark | Figure 2 |
|  | |
| Source: author. | |

Each run: Take the whole set of questions. Randomise the order of the set. Send the questions to each model, all in the same order. Record the answers.

Key difference to traditional benchmark eval tasks: Points are only counted at the level of each parameterised game 𝒈(𝒙). Rationale: if AI responded correctly but failed at the reasoning steps, it cannot be said to have reasoned to reach that answer.

The prompts included variation to identify whether each of the steps is “turned on” or not; when all three are turned on, I consider the result to be consistent with the idea of reasoning. This variation is illustrated below.

Key advantages

Completely made by humans and therefore with a much lower chance of containing artifacts such as MMLU, and still able to scale to a voluminous data set on account of the systematic variations added to each prompt.

Limitations

The space of tasks related to economics spans obviously a much larger area than the types of models that underpin the benchmark. However, this type of models represent in my view a good balance between simplicity, analytical tractability, and wide application.

Should chain-of-thought be used to measure reasoning?

Chain-of-thought (CoT) is a prompting technique by which the user describes to the AI system how to break down an example problem in intermediate steps and solve them to find the correct answer. For example, instead of asking Proponents point at its simplicity of application compared to a burdensome fine-tuning of models, the fact that it tends to offer a glimpse into an AI’s answering process, and empirical findings that it improves performance in standardised benchmarks. The idea was introduced in an influential paper by Wei et al (2022).

The current exercise is purely based on canonical expositions of incomplete information games, that have spawned a numerous secondary literature, including teaching material for economics students. The original papers are all published for 20 years or more. These papers and material about them surely includes the reasoning steps. For this reason, prompting without any type of CoT technique leads to a cleaner assessment of their economic reasoning abilities as they are, especially because successfully identifying the type of model and applying the right knowledge and logic is part of the evaluation itself. In other words, including CoT would muddle the identification of important steps of the reasoning process in the way that the benchmark is set up. However, the process for creating the current economic reasoning benchmark data can be easily adjusted to include different forms of CoT prompting.

# Results

Overall results for the first two steps of the reasoning test are found in Table 1.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overall correct responses  In per cent | | | | | | | | | | Table 1 | |
| Signal | Gemma2B | Gemma7B | GPT 3.5  turbo | GPT 4o | Llama3 | Mistral  instruct | Phi3  mini | Phi3  med | Qwen  0.5B | | Qwen  1.8B |
| 0.495 | 94.7 | 0.0 | 16.0 | 74.3 | 2.8 | 0.0 | 21.8 | 0.0 | 36.1 | | 40.3 |
| 0.505 | 5.6 | 100.0 | 82.3 | 59.7 | 95.1 | 9.7 | 73.6 | 91.0 | 0.0 | | 0.0 |
| Source: Author. | | | | | | | | | | | |

Disappointingly, some cases where models perform relatively well in the first two steps (even if not entirely consistent with reasoning) see a marked drop in accuracy after models are asked to confirm their response. In Table 2, it is possible to see that this drop was heterogenous across models, with the phi3 group particularly prone to performance loss.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Accuracy deterioration when prompted to confirm choice  In percentage points | | | | | | | | | | Table 2 | |
| Signal | Gemma2B | Gemma7B | GPT 3.5  turbo | GPT 4o | Llama3 | Mistral  instruct | Phi3  mini | Phi3  med | Qwen  0.5B | | Qwen  1.8B |
| 0.495 | -8.8 | N/A | -6.2 | -10.4 | -2.8 | N/A | -21.8 | N/A | -9.0 | | -7.6 |
| 0.505 | -0.7 | -0.7 | -2.8 | -22.2 | -20.1 | 0.0 | -63.9 | -50.7 | N/A | | N/A |
| N/A: does not apply because all responses at previous steps were wrong.  Source: Author. | | | | | | | | | | | |

Comparison with human reasoning[[3]](#footnote-4)

Should we expect machines to reason like humans? An analogy can be drawn between birds and airplanes: the latter fly quite effectively even if with a different mechanism that the natural one. Perhaps the same can apply for reasoning?

To be clear, this paper presents a mapping from prompt to answer that stands in for reasoning in a sufficiently generic way to avoid reliance on biological analogies. Instead, it is based more on intuitive definitions of what an analytical response (as opposed to a probabilistic one) would need to be consistent with. In other words, the definition used in this paper aims to be a practical way to distinguish AI systems that can apply such analyses, rather than being normative about *how* specifically they should reason.

The difference with humans is, of course, stark. Broadly speaking, LLMs tend to default to probabilistic associations but are increasingly demonstrating some level of analysis consistent with reasoning, as reflected in the gamut of benchmarks and also in this paper. Humans, on the other hand, can demonstrably reason at deeper levels but because this is a very costly activity in terms of brain effort and time, some level of bounded rationality appears to dominate. For example, game experiments suggest that humans consider higher-order beliefs only up to a small number of interactions (eg,Kübler and Weizsäcker (2004)).

# Conclusions

A truly hard reasoning benchmark can be the basis of a more realistic assessment of the usefulness of these models on advanced knowledge tasks. If they are not to be trusted because their responses result from stochastic parroting, then their ability to drive incremental use cases in knowledge-intensive tasks would appear to be more limited. This is not to say they are doomed to fail; quite the contrary, these models display enormous potential, for example for structuring originally unstructured data. But it could mean that, if the levels of reasoning (and therefore of how trusted these models can be) are kept, their macroeconomic impact may be closer to lower estimates, eg by Acemoglu (2024).

This line of work also prompts a deeper question: should machines even be expected to reason like humans? View 1: Planes vs birds. fundamentally different ways of reasoning. therefore the pathways between prompts and answers cannot be identified through a similar process. View 2: the generic mapping from prompt to answer is generic enough. it can adequately capture the main features that distinguish stochastic parroting from actually analysing a prompt. I am inclined towards the second view.

This work is the first to my knowledge to combine a model of end-to-end reasoning process that can inform a practical benchmark test with economics-related prompts that probe AI systems’ reasoning in that field of science. In addition to the wide disparity of results and to the lacklustre performance overall – despite the lack of novelty in the topic or solution - I would like to conclude with a few reasons why work in this area will be more important going forward, and suggest a few endeavours for further work.

First, a reliance on benchmarks as a measure of how much trust to award each AI system is likely to rise going forward, especially as these models become even more opaque. As some AI systems themselves are so large as to not be usable locally, they are also not available locally for scrutiny and more intensive testing. More importantly, opaqueness has become such an important part of the business model of some important AI developers that even high-level details of the models’ architectures are kept from the public. If this trend continues, benchmark results will be the only meaningful signal to distinguish between different AI systems.

Second, economic agents, including central banks, are themselves increasingly experimenting with creating their own versions (fine-tuned or not) of LLMs. One example is the BIS’ CB-LM (Gambacorta et al (2024)). This benchmarking dataset offers an interesting complement to their custom data for training purposes; and even when the original dataset is not used, developers can apply the methodology described above to build their own reasoning-related data. And of course, the benchmark test can help them assess the reasoning capabilities of the final product and compare it with the relevant models.

Going forward, areas of further work can entail technical progress to achieve more sophistication in how economic reasoning is tested; expanding the benchmark tests to other data modalities (eg, videos or sounds); and also consider how LLMs can help advance experimental economics. This work complements the experiments on game theory (Cabrales et al (2003), Heinemann et al (2004), Szkup & Trevino (2024)); difficult to mimic human reasoning in economics. Goal is momentum for work on economic reasoning (perhaps even a la ChatbotArena) Note, humans themselves also don’t always pass these tests due to limits on rationality.

The current paper introduces a first attempt to systematically measure reasoning abilities in AI systems, with a focus on economic thinking. It is a limited exercise that aims to be simple enough to communicate these concepts while benchmark economic reasoning abilities. Even bigger leaps in AI reasoning measurement can be achieved by probing further the nature of formation of counterfactual beliefs and hypothetical causality models, if any. More sophisticated benchmark tests can help assess how far we are from fulfilling an old dream of a machine that can reason to advance progress in sciences (Langley et al (1987), King (2011)), which could be deployed to solve long-standing problems in economics and improve welfare.

References

Acemoglu, D (2024): “The simple macroeconomics of AI”, NBER Working Paper no 32487.

Achiam, J, S Adler, S Agarwal, L Ahmad, I Akkaya, F Leoni Aleman, D Almeida, et al. 2023. “GPT-4 Technical Report.” arXiv Preprint arXiv:2303.08774.

Araujo, D K G (2023): “gingado: a machine learning model focused on economics and finance”, BIS Working Paper no 1122.

Araujo, D K G, S Doerr, L Gambacorta and B Tissot (2024): “Artificial intelligence in central banking”, BIS Bulletin no 84.

Bender, E M, T Gebru, A McMillan-Major, M Mitchell (2021): “On the dangers of stochastic parrots: can language models be too big? 🦜”, FAccT ’21: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, March.

Cabrales, A, R Nagel and R Armenter (2007): “Equilibrium selection through incomplete information in coordination games: an experimental study”, Experimental Economics, v.10, p.221-234.

Danielsson, J, R Macrae and A Uthemann (2022): “Artificial intelligence and systemic risk”, Journal of Banking and Finance, v.140.

Davis, E and G Marcus (2015): “Commonsense reasoning and commonsense knowledge in artificial intelligence”, Communications of the ACM, v.58, n.9, p.92-103.

Gambacorta, L, B Kwon, T Park, P Patelli and S Zhu (2024): “CB-LMs: language models for monetary policy”, BIS Working Papers, forthcoming.

Heinemann, F, R Nagel and P Ockenfels (2004): “The theory of global games on test: experimental analysis of coordination games with public and private information”, Econometrica, v.72, n.5, p.1583-1599.

Holyak, K J and R G Morrison, Eds (2005): “The Cambridge handbook of thinking and reasoning”, Oxford University Press.

---- (2012): “The Oxford handbook of thinking and reasoning”, Oxford University Press.

King, R D (2011): “Rise of the robo scientist”, Scientific American.

Korinek, A (2023): “Generative AI for economic research: use cases and implications for economists”, Journal of Economic Literature, v.61, n.4, p.1281-1317.

Kübler, D and G Weizsäcker (2004): “Limited depth of reasoning and failure of cascade formation in the laboratory”, Review of Economic Studies, v.71, p.425-442.

Langley, P W, H A Simon, G Bradshaw and J M Zytkow (1987): “Scientific discovery: computational explorations of the creative process”, MIT Press.

Mele, A R and P Rawling, Editors. (2004): “The Oxford handbook of rationality”, Oxford University Press.

Morris, S and H S Shin (1998): “Unique equilibrium in a model of self-fulfilling currency attacks”, American Economic Review, June.

Morris, S and H S Shin (2004): “Coordination risk and the price of debt”, European Economic Review, v.48, n.1, p.133-153.

Nezhurina, M, L Cipolina-Kun, M Cherti and J Jitsev (2024): “Alice in Wonderland: simple tasks showing complete reasoning breakdown in state-of-the-art large language models”, arXiv.

Perez-Cruz, F and H S Shin (2024): “Testing the Cognitive Limits of Large Language Models.”, BIS Bulletin no 83.

Sarkar, S and J Pfeifer, Eds. (2006): ‘The philosophy of science”, Routledge.

Stantcheva, S (2021): “Understanding tax policy: how do people reason”, Quarterly Journal of Economics, v.136, n.4, p.2309-2369.

Stantcheva, S (2023): “How to run surveys: a guide to creating your own identifying variation and revealing the invisible”, Annual Review of Economics, v.15, p.205-234.

Storks, S, Q Gao, J Y Chai (2020): “Recent advances in natural language inference: a survey of benchmarks, resources and approaches”, arXiv.

Wang, Y, X Ma, G Zhang, Y Ni, A Chandra, S Guo, W Ren, A Arulraj, X He, Z Jiang, T Li, M Ku, K Wang, A Zhuang, R Fan, X Yue, W Chen (2024): “MMLU-Pro: a more robust and challenging multi-task language understanding benchmark”, arXiv.

Wei, J, X Wang, D Schurrmans, M Bosma, B Ichter, F Xia, E H Chi, Q V Le, D Zhou (2022): “Chain-of-thought prompting elicits reasoning in large language models”, Advances in Neural Information Processing Systems.

Yang, J, H Jin, R Tang, X Han, Q Feng, H Jiang, B Yin and X Hu (2023): “Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond.” arXiv Preprint arXiv:2304.13712

Y Chang, X Wang, J Wang, Y Wu, L Yang, K Zhu, H Chen, X Yi, C Wang, Y Wang, W Ye, Y Zhang, Y Chang, P S Yu, Q Yang, X Xie (2024): “A survey on evaluation of large language models”, ACM Transactions on Intelligent Systems and Technology, v.15, n.3, p.1-45.

Tables should always be indented by 0.19cm.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Correct responses  TableMainSubHeading | | | | | | Table 1 | |
| TableColumn Heading |  | 2012 |  |  |  | |  |
| TableRowHeading | TableText | Table Cell |  |  |  | |  |
|  |  | 3.4 |  |  |  | |  |
|  |  | 21.44 |  |  |  | |  |
| TableNote | | | | | | | |

1. [Please note analyses are ongoing and therefore the results are to be taken as preliminary only.] I am thankful for seminar participants at the BIS Informal Machine Learning Community, the Platform for Advanced Scientific Computing (PASC) Conference 2024, and … for comments that helped improved this paper. I thank in particular Fernando Perez-Cruz for detailed comments on a previous version of this paper. All errors are my own. The opinions in this paper do not necessarily reflect those of the Bank for International Settlements. [↑](#footnote-ref-2)
2. Even in areas that are not outwardly “causal”, such as the literature macroeconomic forecasting, there is an explicit or implicit underlying assumption that the seemingly purely numerical results must “make sense”, ie conform to some subjective structural model in the readers’ mind. [↑](#footnote-ref-3)
3. The author appreciates Fernando Pérez-Cruz for suggesting to include such a discussion in the paper. [↑](#footnote-ref-4)