Measuring the economic reasoning abilities of language models*

Douglas K. G. Araujo[1]

[1]Bank for International Settlements, douglas.araujo@bis.org

Abstract

Economic reasoning is represented as a state space function.

1 Introduction

Language models (LMs), in particular those classified as generative artificial intelligence (gen AI), are finding increasing uses in finance and economics. These models are usually tested for their ability to reason, and seem to do well: for example, OpenAI's GPT-4 boats 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. Perez-Cruz and Shin (2024) demonstrate how the same model can correctly solve a logical puzzle requiring reasoning about higher order knowledge, only to fail when irrelevant details are changed. Building on results such as this and other examples that clearly illustrate the limits of rationality assumptions on LMs, this work discusses how to systematically measure *economic* reasoning, combining literatures on economic thought and on computer science about gen AI benchmarking. In practical terms, the task at hand is to come up with testing mechanisms that estimate the level of economic reasoning of an LM by means of a prompt consisting of $n \geq 0$ examples and a question with multiple answers.

At its most essential form, testing for economic reasoning is the same as probing if the model is able to think in terms of logical operators. However, they can be subjective because (a) economic thought is always changing and (b) they are only as good as their abilites ot explain limited sets of reality (those that modern academics constantly see= rather than any other reality).

Similar to many other social disciplines, economics requires the analytical judgment referred to by Robbins (1932) in the analyses of events as a basis to extrapolate and predict, and this has a bearing on how economic reasoning should be

*This work represents my opinion and not necessarily that of the BIS.

benchmarked. Economic inference depends primarily on articulating unobservable quantities, theorecised and estimated on the basis of observable measures. This is unlike other major disciplines. For example, in human and veterinary medicine, all physiological and pathological variables of clinical importance are observable, even if that is not yet technologically feasible today. In the medical sciences, theoretical models merely fill in the gaps in the absence of a technologically feasible complete measurement. In contrast, many economically relevant quantities are latent variables that cannot by definition be observed, and always require a model applied to data to be estimated, implicit or not.

A quantitative test for economic reasoning must take this into account: selecting a correct answer in an economics question through reasoning will always depend on an unobserved transformation of the information received and the existing knowledge. This is important. LMs may also happen to choose the correct answer from either luck of through simple token probability. It is easy to see why a correct answer selected by chance is not informative about the reasoning abilities of a model. The second case requires more explanation: mathematically, LMs are trained to identify the most likely token θ in a vocabulary V given the tokens in its prompt. In practice, the function is inexcrutable so it is also considered an unobservable transformation. But a few characteristics allows us to distinguish reasoning from prediction. First, reasoning is robust to minutiae and other irrelevant detail. Mathematically, it would be analogous to applying a manifold transformation that retains only the relevant information in a prompt and then applies logic operations on top of them, and on them only. Second, reasoning is locally complete, meaning that an LM that can correctly deduce that A implies B also is able to understand that A' does not imply B, or that A does not imply B'. In other words, a reasoning that appears to be correct but whose obvious corolary is not achieved by an LM cannot be said to have been reasoned in the first place.

Knowledge: linguistic, common and commonsense.

Interpretation. information theory. Shannon.

The main intuition of this work is to combine a number of building blocks of evaluation.

- the benchmark must be challenging for machines: I use an adjusted version of adversarial filtering (Zellers et al. (2019)) to create answer candidates that are hard for LMs to guess
- the test must incorporate slow-moving evolutions in academic economic thought: evolving test set based on newly published academic work.
- results related to reasoning must be distinguished as best as possible from the ability to interpret the prompt or from knowledge (implicit or explicit) about economics, ie reasoning is a separate step: sets of perturbations in the spirit of Alzahrani et al. (2024) for each initial task.

the benchmark counts with a mathematical adjustment that takes into account performance across perturbations, penalising results that vary with ….

This benchmark evaluation addresses a poignant issue for the economics profession: the lack of publicly available data about how these benchmarks are created and any, and toasted.

A major inspiration in the design of the questions and how they can generate identifying variations is the social economics literature. A key reference is Stantcheva (2023). The idea here is that the design of the questionnaire itself can elicit responses that allow for insight into non-observable traits such as reasoning. Many of the insights of this literature carry over naturally to the machine space.[1]

A substantial body of work creates and discusses benchmarking models in general. A very useful reference is Storks, Gao, and Chai (2020). Literature on benchmarking economic reasoning appears to be new, although other works have touched upon the topic from different angles. An early foray into questions related to AI's ability to conduct economic reasoning is due to Parkes and Wellman (2015). But their angle is more on how AIs can be used to estimate synthetic economic agents - machina oeconomicus - ideal versions of purely rational agents, rather than on the measurement and the implications of AIs acquiring economic reasoning abilities. In any case, Parkes and Wellman (2015) see economic reasoning as the ability to understand and solve complex game-theoretical environments (eg, the poker example).

2 Lessons from human surveys

I use a considerable amount of specific advice on human surveys from Stantcheva (2023) to generate identifying variation in the questions. Specifically:

- coeteris paribus questions
- pre-testing
- including possibilities for blank, indifferent or even recognise that AI does not know
- avoiding jargon
- questions that check for "attention" and "effort" on the part of the respondent
- also including open ended questions (as in Ferrario and Stantcheva (2022))
  - including follow-up questions ("are thre any other reasons")
  - going beyond Ferrario and Stantcheva (2022), in this paper I use open ended questions that are similar in nature to closed end questions and deploy large language models to interpret them.
- question ordering

[1] Actually testing whether LMs *do not* parrot or "organically" exhibit biases or other behaviours that are assumed to be exclusively human would be an interesting line of research.

  - in particular, consideration is given to whether each question should be presented to a separate instance of the LM, or the full questionnaire could be shared in the same "chat".
- take due consideration of how to address the different types of bias associated with surveys (adapted for the machine context, naturally)

3 Desirable characteristics of a benchmark

3.1 Evolve over time

Economic reasoning evolves over time. For example, the Lucas critique (Lucas (1976)) was influential in shifting macroeconomic modelling, while the credibility revolution described in Angrist and Pischke (2008) was similarly influential in microeconomic work.

4 A model of economic reasoning

The result from existing benchmarks is largely, if not completely, directly related to the number of questions correctly answered. However, this measures only the model's ability to answer correctly, *not necessarily* its reasoning capabilities. The latter are part of a latent state space sitting between the input prompt and the answer. More concretely, for an input prompt X, which includes a question and any necessary explicit information, the language model is a function $\mathbf{M}$ that maps it to a given response: $\mathbf{M} : X \to y$. In order to show that it is done by reasoning, we need tests (and more specifically, measurements) that convey some information about the inner workings of this function.

4.1 Reasoning as an abstract of the input

- Input prompt X
- Transformed into $g(X, \kappa)$, a state space function that also takes the existing knowledge κ and associates it with the prompt to maps it to its abstract fundamentals (similar to manifold learning)
- Result based on $g(X)$.

4.2 A (very) simple model

This section builds on the intuition that in true reasoning, the result should be robust to minute perturbations, ie the model is a constant function over the domain of the input. Formally, both $\mathbf{M}(X) = y$ and $\mathbf{M}(X + \epsilon) = y$ for an infinitesimal ϵ. This implies the derivative with respect to the input prompt is zero. Using as an approachable example the simplest possible neural network, the logistic regression $\mathbf{N}(x) = \sigma(Wx + b)$, such robustness further implies that $\frac{d\mathbf{N}}{dx} = \sigma(Wx+b)(1-\sigma(Wx+b))W = 0$. Because W cannot be a zero vector in a

functioning network that is responsive to its inputs and $\sigma(Wx+b)(1-\sigma(Wx+b))=0$ has no solution because neither term is 0 or 1 in a sigmoid function with finite inputs, the neural network cannot be a constant function. This extremely simplified example, which holds for recursive architectures of similarly simple layers, does not bode well for the robustness of results given small perturbations in the input prompt.

5 Reasoning benchmarks in other fields

- Math
- Medical
- Biologia

6 A model of reasoning

This section develops a model of reasoning that fits naturally into both natural and artificial LMs. It will serve as the basis for the subsequent analyses and empirical creation of a reasoning benchmark.

Let a sentence $\mathbf{S} = (\theta_1, \theta_2, \theta_3, ...)$ be a sequence of token-location tuples $\theta_x = (\tau, x)$, with each $\tau \in \mathbf{V}$ belonging to a vocabulary $\mathbf{V}$ and $x \in \mathbb{N}^{d_{\text{model}}}$.[2] Create a function $\pi_{i,C} : \theta, \mathbf{S} \rightarrow \{-1, 0, 1\}$ that maps each token into one of three possibilities: the token's information can be considered a adversarial (-1), irrelevant (0) or relevant (1) with respect to the likelihood of individual (or LM) i uttering another sentence C. For example, take the following quote from the character Barf in the 1987 movie Spaceballs, organised as two sentences "I'm a mog. Half man, half dog." and "I'm my own best friend." With word-level tokenisation, $\mathbf{S} = \{("\text{I'm}", 1), ("\text{a}", 2), ("\text{mog}", 3), (".", 4), ("\text{Half}", 5), ("\text{man}", 6), (",", 7), ("\text{half}", 8), ("\text{dog}", 9), (".", 10)\}$ and $\mathbf{C}$ is similarly broken down. This example illustrates that even when there is not a logical connection grounded in truth, tokens in one sentence - even those made up like "mog", can have a bearing on the likelihood of tokens appearing in another sentence. This likelihood can differ depending on the location of the token, which also allows for situations where repeteating of a word τ is meant to convey different meaning. Another feature of this example is that all $\pi_{\text{Barf},C}(\theta) = 1 \forall \theta \in \mathbf{S}$. In the alternative sentence "I'm a mog. Half man, half dog. I am alive.", the new component is obviously irrelevant for $\mathbf{C}$: $\prod_{x \in [10,14]} \pi_{\text{Barf},C}(\theta_x) = 0$.

This exposition is important to delve into the reasoning aspect, entirely organised by function π. Since $\pi_{i,C}$ measures how informative a token is for individual

[2]The location is important because it helps define meaning, along with the actual letter (more generally, symbol) content of th token. Note that in this paper, white spaces are abstracted away for expositional simplicity.

i's $\mathbf{C}$, it constitutes the first aspect of reasoning: to recognise when a token is adversarial, irrelevant or relevant. This step is necessary before the application of any logical rules $l \in \mathcal{L}$ on the weighted token, $\pi_{i,C}(\theta_x)\theta_x$. The exact underpinnings of these logical rules are beyond the scope of this work - it can be approximated by a possibly non-linear function, g. What suffices in this work is to say that reasoning *depends* on correctly classifying the tokens: all relevant tokens must be so identified, lest they be either ignored as the irrelevant ones or taken with the opposite meaning. Similarly, if all relevant tokens are indeed diagnosed correctly but other tokens are also diagnosed as relevant when they are not, then this will cause problems for the correct reasoning. In other words, a first precondition for reasoning is to have a low categorical cross-entropy loss. Intuitively, a pre-condition of reasoning is to correctly interpret the inputs.

Use Taylor expansion on model since its derivative to perturbation should be zero. This gives us a head start in the Taylor expansion. Try to link the T-expanded equation to an estimating equation.

But what determines $\pi_{i,C}$? A combination of knowledges and logical relationships.

Knowledges: linguistic knowledge, common knowledge and commonsense knowledge

Rationales: reasoning from logic

Armed with the sentence-level categorical cross-entropy, the individual can establish chains of thought that will finally lead to reasoning. Again, for simplicity, the exact function is not discussed here, other than that it is a potentially simple or complex way to interact. What is important is to add the categorical cross-entropy to the estimation equation.

Benchmark testing mechanism...

7 Reasoning about economics

The model above allows us to estimate reasoning while also breaking down some of its components to better understand them. For example, we can estimate any errors in reasoning into an issue with **interpretation**, **knowledge** and **logical thinking**. The empirical estimation follows.

8 Empirical estimation

Each *task* $\theta \in \Theta$ can be asked in various different ways, each one being called a *question* $q \in \theta$. Questions vary with respect to their adversarial aspect; it is this variation within each question that allows the empirical estimation of the effects associated with interpretation or with knowledge. Most of the variations are originally those tested in Alzahrani et al. (2024). The variation in response

between the questions within each task will comprise the evaluation of the actual reasoning capabilities. As alluded to before, the variations are organised into those that measure the stability of a response to adversarial interpretation answers, and those that measure the stability across the knowledge dimension. In practice, each task has hundreds of different q. These groups are described in more detail next.

8.1 Variations related to interpretation

There are several classes of variations that can help test an LMs' interpretation.

8.1.1 Choice variations

Here the choices remain the same for a task but vary in their order across questions

- random choice order
- biased choice order
- uncommon answer choice symbols
- common but unordered answer choice symbols

8.1.2 Word variations

The main idea here is to introduce or change words that are irrelevant. This is along the lines of the test conducted by Perez-Cruz and Shin (2024).

Another one is to conduct random word repetition as if it were a typo

8.2 Variations related to knowledge

Changing key words related to field knowledge with other field knowledge words but that would not make a sense to an expert. This can be compared with just changing the same words into another generic word. Comparing responses between both should indicate the level of knowledge used by the model (should it? need to think more)

8.3 Estimation formula

The main formula is akin to the linear probability model since a_q is either zero or one:

$$a_q = \beta_\theta \theta + \beta_{\text{Interpretation}} \eta_q + \beta_{\text{Knowledge}} \kappa_q + \epsilon_q$$

Another idea to explore is whether these variations can actually instrument interpretation and knowledge. This would allow the formula to estimate the reasoning bit.

9 Operational characteristics

- avoid becoming part of training data

Some drawbacks of using academic papers include:

- bias to report only positive findings (and to do so in a way that is generous towards said findings)
- Also, academic papers suffer from false negatives: many contributions that are now considered classics have been previously rejected (Gans and Shepherd (1994)).

10 Conclusions

As economic agents and policymakers harness generative artificial intelligence (AI) to reap considerable efficiencies, and thus their societal footprint becomes larger, a benchmark for economic reasoning is needed. I suggest ways to implement such a benchmark, and measure the current performance of a selected list of LMs.

11 Annex 1: discussion of biases in human surveys and how they could affect LM questionnaires

- Section A-4 in Stantcheva (2023)

The goal of this annex is to list side-by-side the main human biases that affect survey responses and their corresponding machine version, if any (from a theoretical perspective - it would be interesting to test if LMs carry over some of these biases that are supposed to be only human, which could suggest they are parroting or in extremis developing sources of bias like shame, etc).

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. 2023. "Gpt-4 Technical Report." *arXiv Preprint arXiv:2303.08774*.

Alzahrani, Norah, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, et al. 2024. "When Benchmarks Are Targets: Revealing the Sensitivity of Large Language Model Leaderboards." *arXiv Preprint arXiv:2402.01781*.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ferrario, Beatrice, and Stefanie Stantcheva. 2022. "Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions." In *AEA*

Papers and Proceedings, 112:163–69. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Gans, Joshua S., and George B. Shepherd. 1994. "How Are the Mighty Fallen: Rejected Classic Articles by Leading Economists." *Journal of Economic Perspectives* 8 (1): 165–79. https://doi.org/10.1257/jep.8.1.165.

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." *Journal of Monetary Economics* 1 (2): 19–46.

Parkes, David C, and Michael P Wellman. 2015. "Economic Reasoning and Artificial Intelligence." *Science* 349 (6245): 267–72.

Perez-Cruz, Fernando, and Hyun Song Shin. 2024. "Testing the Cognitive Limits of Large Language Models." Bank for International Settlements.

Robbins, Lionel. 1932. *An Essay on the Nature and Significance of Economic Science*. Macmillan; Co., Limited.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15: 205–34.

Storks, Shane, Qiaozi Gao, and Joyce Y. Chai. 2020. "Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches." https://arxiv.org/abs/1904.01172.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. "Hellaswag: Can a Machine Really Finish Your Sentence?" *arXiv Preprint arXiv:1905.07830*.