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.

The main intuition of this work is to combine a number of building blocks of evaluation. First, adversarial filtering as in HellaSwag. Second, evolving test set based on newly published academic work. Third, perturbations in the spirit of Alzahrani et al. (2024). And fourth, 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.

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

2 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.

2.1 Reasoning as an abstract of the input

- Input prompt X
- Transformed into $g(X)$, a state space function that maps it to its abstract fundamentals (similar to manifold learning)
- Result based on $g(X)$.

2.2 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 does not bode well for the robustness of results given small perturbations in the input prompt.

3 Reasoning benchmarks in other fields

- Math
- Medical
- Biologia

4 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 tokens-location tuples $\theta_x = (\tau, x)$, with each $\tau \in \mathbf{V}$ belonging to a vocabulary $\mathbf{V}$ and $x \in \mathbb{N}^{d_{\text{model}}}$.[1] 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, $\mathbf{S} = \{(\text{"I'm"}, 1), (\text{"a"}, 2), (\text{"mog"}, 3), (\text{"."}, 4), (\text{"Half"}, 5), (\text{"man"}, 6), (\text{","}, 7), (\text{"half"}, 8), (\text{"dog"}, 9), (\text{"."}, 10)\}$ 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 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.

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

Knowledges: linguistic knowledge, common knowledge and commonsense knowledge

[1] 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.

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.

5 Reasoning about economics

6 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. Following Alzahrani et al. (2024), the variations q are:

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

In practice, each task has hundreds of different q. The variation in response between the questions within each task will comprise the evaluation of the actual reasoning capabilities.

In addition, two more tests are added: the changing of irrelevant detail (a la Perez-Cruz and Shin (2024)) and random word repetition as if it were a typo.

7 Operational characteristics

- avoid becoming part of training data

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*.

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