Stop Making Unscientific AGI Performance Claims

Patrick Altmeyer, Andrew M. Demetriou, Antony Bartlett, Cynthia C. S. Liem (TU Delft)

Keywords — Machine Learning, Anthropomorphism, Artificial General Intelligence

I. Position

We therefore urge our fellow researchers to stop making unscientific AGI performance claims.

Current LLMs embed information. They don't „understand" anything. They are useful tools, but tools nonetheless.

- Meaningful patterns in embeddings are like doves in the sky.
- Humans are prone to seek patterns and anthropomorphize.
- Observed 'sparks' of Artificial General Intelligence are spurious.
- The academic community should exercise extra caution.
- Publishing incentives need to be adjusted.

II. Are Neural Networks Born with World Models?

- Llama-2 model tested in [1] has ingested huge amounts of publicly available data [2].
  - Geographical locations are literally in the training data: e.g. Wikipedia article for "London".
  - Where would this information be encoded if not in the embedding space $\mathcal{A}$? Is it surprising that $A_{\text{LDN}}=enc(\text{"London"}) \not\!\perp\!\!\!\perp \perp (\text{lat}_{\text{LDN}},\text{long}_{\text{LDN}})$?
- Figure 1 shows the predicted coordinates of a linear probe on the final-layer activations of an untrained neural network.

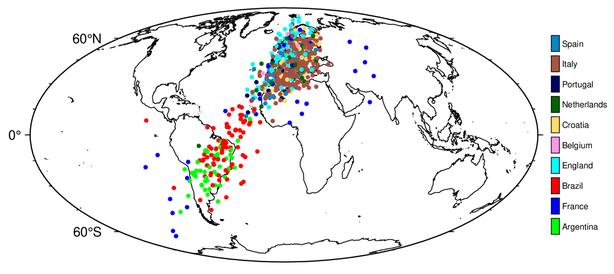

Figure 1: Predicted coordinate values (out-of-sample) from a linear probe on final-layer activations of an untrained neural network.

- Model has seen noisy coordinates plus d random features.
- Single hidden layer with $h < d$ hidden units.

III. PCA as a Yield Curve Interpreter

What are principal components if not model embeddings?

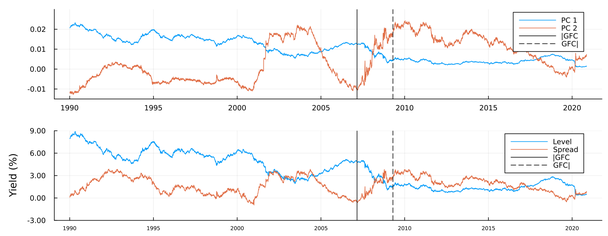

Figure 2: Top chart: The first two principal components of US Treasury yields over time at daily frequency. Bottom chart: Observed average level and 10yr-3mo spread of the yield curve. Vertical stalks roughly indicate the onset (IGFC) and the beginning of the aftermath (GFCI) of the Global Financial Crisis.

IV. Autoencoders as Economic Growth Predictors

- Yes, this can be used for feature extraction and forecasting:
  - Bottle-neck layer embeddings predict spread and level of the yield curve.

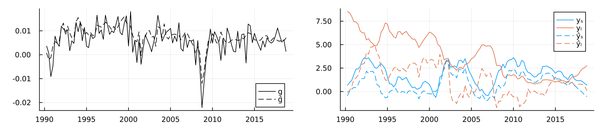

Figure 3: The left chart shows the actual GDP growth and fitted values from the autoencoder model. The right chart shows the observed average level and spread of the yield curve (solid) along with the predicted values (in-sample) from the linear probe based on the latent embeddings (dashed)

V. Embedding FOMC comms

- BERT-based model trained on FOMC minutes, speeches and press conferences to classify statements as hawkish or dovish (or neutral) [3].
- We linearly probe all layers to predict unseen economic indicators (CPI, PPI, UST yields).
- Predictive power increases with layer depth and probes outperform simple AR(p) models.

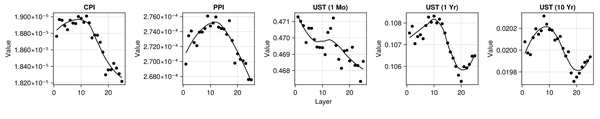

Figure 4: Out-of-sample root mean squared error (RMSE) for the linear probe plotted against FOMC-RoBERTa's n-th layer for different indicators.

VI. Sparks of Economic Understanding?

Premise: If probe results were indicative of some intrinsic 'understanding' of the economy, then the probe should not be sensitive to random sentences unrelated to economics.

1) *Parrot Test:*

1. Select the best-performing probe for each economic indicator.
2. Predict inflation levels for real (related) and perturbed (unrelated) sentences.

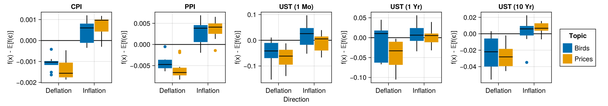

Figure 5: Probe predictions for sentences about inflation of prices (IP), deflation of prices (DP), inflation of birds (IB) and deflation of birds (DB). The vertical axis shows predicted inflation levels subtracted by the average predicted value of the probe for random noise.

As evidenced by Figure 5, the probe is easily fooled.

VII. Spurious Relationships

Definiton: Varies somewhat [4] but distinctly implies that the observation of correlations does not imply causation.

- Humans struggle to tell the difference between random and non-random sequences [5].
- Lack of expectation that randomness that hints towards a causal relationship will still appear at random.
- Even experts perceive correlations of inflated magnitude [6] and causal relationships where none exist [7].

VIII. Antropomorphism

Definition: Human tendency to attribute human-like characteristics to non-human agents and/or objects.

1. Experience as humans is an always-readily-available template to interpret the world [8].
2. Motivation to avoid loneliness may lead us to anthropomorphize inanimate objects [8], [9].
3. Motivation to be competent may lead us anthropomorphize opaque technologies like LLMs [8], [9].

IX. Confirmation Bias

Definition: Favoring interpretations of evidence that support existing beliefs or hypotheses [6].

- Hypotheses in present-day AI research are often implicit, often framed simply as a system being more accurate or efficient, compared to other systems.
  - Failing to articulate a sufficiently strong null hypothesis leading to a 'weak' experiment [10].
- Individuals may place greater emphasis on evidence in support of their hypothesis, and lesser emphasis on evidence that opposes it [6].

X. Conclusion and Outlook

- We call for the community to create explicit room for organized skepticism
  - Welcome negative results
  - Encouraging replication studies.

https://arxiv.org/abs/2402.03962 Forty-first International Conference on Machine Learning (ICML) p.altmeyer@tudelft.nl

- Move from authorship to contribution-based credit (see e.g. Liem and Demetriou, 2023 and Smith, 1997).
- Return to the Mertonian norms (communism, universalism, disinterestedness, organized skepticism) [11].

Bibliography

[1] W. Gurnee and M. Tegmark, “Language Models Represent Space and Time”, *arXiv preprint arXiv:2310.02207v2*, 2023.

[2] H. Touvron *et al.*, “LLaMA: Open and Efficient Foundation Language Models”. 2023.

[3] A. Shah, S. Paturi, and S. Chava, “Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis”, *arXiv preprint arXiv:2310.02207v1*, 2023.

[4] B. D. Haig, “What is a spurious correlation?”, *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, vol. 2, no. 2, pp. 125–132, 2003.

[5] R. Falk and C. Konold, “Making sense of randomness: Implicit encoding as a basis for judgment.”, *Psychological Review*, vol. 104, no. 2, p. 301, 1997.

[6] R. S. Nickerson, “Confirmation bias: A ubiquitous phenomenon in many guises”, *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[7] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska, “Investigating the effect of the multiple comparisons problem in visual analysis”, in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–12.

[8] N. Epley, A. Waytz, and J. T. Cacioppo, “On seeing human: a three-factor theory of anthropomorphism.”, *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[9] A. Waytz, N. Epley, and J. T. Cacioppo, “Social cognition unbound: Insights into anthropomorphism and dehumanization”, *Current Directions in Psychological Science*, vol. 19, no. 1, pp. 58–62, 2010.

[10] A. Claesen, D. Lakens, N. van Dongen, and others, “Severity and Crises in Science: Are We Getting It Right When We're Right and Wrong When We're Wrong?”, 2022.

[11] R. K. Merton and others, “Science and technology in a democratic order”, *Journal of legal and political sociology*, vol. 1, no. 1, pp. 115–126, 1942.