Stop Making Unscientific AGI Performance Claims

Patrick Altmeyer, Andrew M. Demetriou, Antony Bartlett, Cynthia C. S. Liem

Delft University of Technology, Department of Intelligent Systems

TLDR — We probe models of varying complexity including random projections, matrix decompositions, deep autoencoders and transformers: all of them successfully distill information that can be used to predict latent or external variables and yet none of them have previously been linked to AGI. We argue and empirically demonstrate that the finding of meaningful patterns in latent spaces of LLMs cannot be seen as evidence in favor of AGI. Additionally, we review literature from the social sciences that shows that humans are prone to seek such patterns and anthropomorphize.

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.
- The academic community should exercise extra caution.

II. Experiments

i. Are Neural Networks Born with World Models?

Llama-2 model tested in (Gurnee & Tegmark, 2023) has ingested huge amounts of data including Wikipedia dumps that contain geographical coordinates (Touvron et al., 2023): e.g. Wikipedia article for "London".

Where would this information be encoded if not in the embedding space $\mathcal{A}$? Is it really surprising that $A_{\text{LDN}} = enc(\text{"London"})$ predicts $(\text{lat}_{\text{LDN}}, \text{long}_{\text{LDN}})$?

A simple experiment:

- Model in Figure 1 has seen noisy coordinates of top-10 FIFA World Cup countries plus d random features.
- Randomly initialized single hidden layer with $h < d$ units.

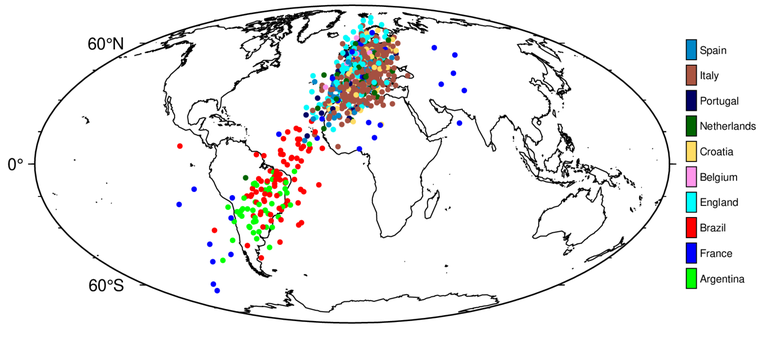

Figure 1: Predicted coordinate values (out-of-sample) from a linear probe on final-layer activations of an untrained neural network.

ii. PCA as a Yield Curve Interpreter

It is common practice to use principal component analysis (PCA) to extract meaningful latent features of yield curves (Crump & Gospodinov, n.d.).

What are principal components, if not model embeddings?

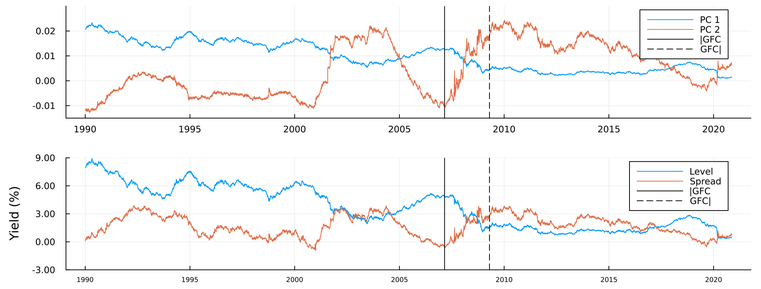

Figure 2: The first two principal components of US Treasury yields (top) and the observed average level and 10yr-3mo spread (bottom).

iii. Sparks of Economic Understanding?

If probe results were indicative of some intrinsic 'understanding' of the economy, then the probe should not be sensitive to unrelated sentences. As evidenced by Figure 4, probes are easily.

BERT-based model trained on FOMC minutes, speeches and press conferences to classify statements as hawkish or dovish (or neutral) (Shah et al., 2023).

- We linearly probe all layers to predict unseen economic indicators (CPI, PPI, UST yields).
- Predictive power increases with layer depth (Figure 3) and probes outperform simple AR(p) models.

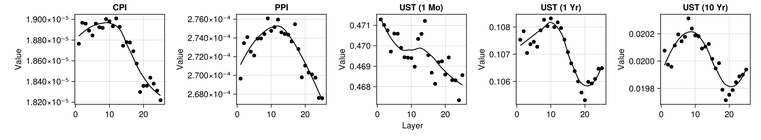

Figure 3: Out-of-sample root mean squared error (RMSE) for the linear probe plotted against FOMC-RoBERTa's n-th layer for different indicators.

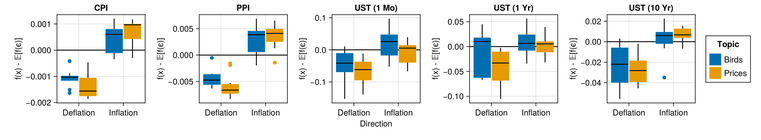

Figure 4: Probe predictions for sentences about inflation of prices (IP), deflation of prices (DP), inflation of birds (IB) and deflation of birds (DB). The vertical axis shows predicted inflation levels subtracted by the average predicted value of the probe for random noise.

III. Social Sciences Review

i. Spurious Relationships

Definiton: Varies somewhat (Haig, 2003) but distinctly implies that the observation of correlations does not imply causation.

- Humans struggle to tell the difference between random and non-random sequences (Falk & Konold, 1997).
- Lack of expectation that randomness that hints towards a causal relationship will still appear at random.
- Even experts perceive correlations of inflated magnitude (Nickerson, 1998) and causal relationships where none exist (Zgraggen et al., 2018).

ii. Anthropomorphism

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 (Epley et al., 2007).
2. Anthropomorphize inanimate objects to avoid loneliness (Epley et al., 2007), (Waytz et al., 2010).
3. Anthropomorphize opaque technologies like LLMs to be competent (Epley et al., 2007), (Waytz et al., 2010).

iii. Confirmation Bias

Definition: Favoring interpretations of evidence that support existing beliefs or hypotheses (Nickerson, 1998).

- 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 (Claesen et al., 2022).
- Individuals may place greater emphasis on evidence in support of their hypothesis, and lesser emphasis on evidence that opposes it (Nickerson, 1998).

IV. Conclusion and Outlook

Concrete recommendations for future research

- (*Acknowledge Human Bias*) Be explicit about risks of human bias and anthropomorphization.
- (*Stronger Testing*) Refrain from premature AGI conclusions.
- (*Epistemologically Robust Standards*) Define terms like 'intelligence' and 'AGI' precisely.

Furthermore: create explicit room for organized skepticism; welcome negative results; encourage replication studies; move from authorship to contribution-based credit (see e.g. Liem and Demetriou, 2023 and Smith, 1997).

Bibliography

Claesen, A., Lakens, D., Dongen, N. van, & others. (2022). *Severity and Crises in Science: Are We Getting It Right When We're Right and Wrong When We're Wrong?*.

Crump, R. K., & Gospodinov, N. *Deconstructing the yield curve*.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301.

Gurnee, W., & Tegmark, M. (2023). Language Models Represent Space and Time. *Arxiv Preprint Arxiv:2310.02207v2*.

Haig, B. D. (2003). What is a spurious correlation?. *Understanding Statistics: Statistical Issues in Psychology, Education, And the Social Sciences*, *2*(2), 125–132.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Shah, A., Paturi, S., & Chava, S. (2023). Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. *Arxiv Preprint Arxiv:2310.02207v1*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*.

Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, *19*(1), 58–62.

Zgraggen, E., Zhao, Z., Zeleznik, R., & Kraska, T. (2018). Investigating the effect of the multiple comparisons problem in visual analysis. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.