Counterfactual Training: Teaching Models Plausible and Actionable Explanations

Patrick Altmeyer, Aleksander Buszydlik, Arie van Deursen, Cynthia C.S. Liem
(Delft University of Technology)

KEYWORDS — Counterfactual Explanations, Algorithmic Recourse, Adversarial ML

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.

EXPERIMENTS

i. Are Neural Networks Born with World Models?

Llama-2 model tested in W. Gurnee and M. Tegmark [1] has ingested huge amounts of data including Wikipedia dumps that contain geographical coordinates [2]: 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 ?@fig-map has seen noisy coordinates of top-10 FIFA World Cup countries plus d random features.
- Randomly initialized single hidden layer with $h < d$ units.

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 [3].

What are principal components, if not model embeddings?

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 ?@fig-attack, probes are easily.

BERT-based model trained on FOMC minutes, speeches and press conferences to classify statements as hawkish or dovish (or neutral) [4].

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

SOCIAL SCIENCES REVIEW

i. Spurious Relationships

Definiton: Varies somewhat [5] but distinctly implies that the observation of correlations does not imply causation.

- Humans struggle to tell the difference between random and non-random sequences [6].
- Lack of expectation that randomness that hints towards a causal relationship will still appear at random.
- Even experts perceive correlations of inflated magnitude [7] and causal relationships where none exist [8].

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 [9].
2. Anthropomorphize inanimate objects to avoid loneliness [9], [10].
3. Anthropomorphize opaque technologies like LLMs to be competent [9], [10].

iii. Confirmation Bias

Definition: Favoring interpretations of evidence that support existing beliefs or hypotheses [7].

- 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 [11].
- Individuals may place greater emphasis on evidence in support of their hypothesis, and lesser emphasis on evidence that opposes it [7].

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

[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] R. K. Crump and N. Gospodinov, “Deconstructing the yield curve.”

[4] A. Shah, S. Paturi, and S. Chava, “Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis.” 2023.

[5] 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.

[6] R. Falk and C. Konold, “Making sense of randomness: Implicit encoding as a basis for judgment.,” *Psychological Review*, vol. 104, no. 2, p. 301–302, 1997.

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

[8] 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.

[9] N. Epley, A. Waytz, and J. T. Cacioppo, “On seeing human: a three-factor theory of anthropomorphism.,” *Psychological review*, vol. 114, no. 4, p. 864–865, 2007.

[10] 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.

[11] 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.

p.altmeyer@tudelft.nl