Against Spurious Sparks — Dovelating Inflated AI Claims

ECONDAT Conference 2024[1]

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

Delft University of Technology

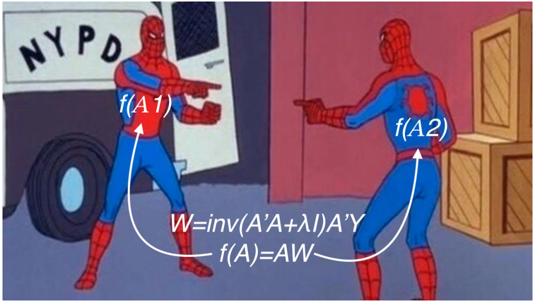
2024-05-07

[1] Upcoming position paper at ICML 2024.

Motivation

- Statement 1: „It is essential to bring inflation back to target to avoid drifting into deflation territory.“
- Statement 2: „It is essential to bring the numbers of doves back to target to avoid drifting into dovelation territory.“

Linear probe $f(A_1)$ on LLM embeddings: “They’re exactly the same.”



Position

Current LLMs embed knowledge. They don‘t „understand“ anything.

1. Finding meaningful patterns in LLM embeddings is like finding doves in the sky.
2. Developments in the field of AI in general, and Large Language Models (LLMs) in particular, have created a ‘perfect storm’ for observing ‘sparks’ of Artificial General Intelligence (AGI) that are spurious.
3. We therefore call for the academic community to exercise extra caution, and to be keenly aware of principles of academic integrity, in interpreting and communicating about AI research outcomes.

Outline

- **Experiments**: We probe models of varying degrees of sophistication including random projections, matrix decompositions, deep autoencoders and transformers.

Outline

- **Experiments**: We probe models of varying degrees of sophistication including random projections, matrix decompositions, deep autoencoders and transformers.
    - All of them successfully distill knowledge and yet none of them develop true understanding.

Outline

- **Experiments**: We probe models of varying degrees of sophistication including random projections, matrix decompositions, deep autoencoders and transformers.
  - All of them successfully distill knowledge and yet none of them develop true understanding.
- **Social Sciences review**: Humans are prone to seek patterns and anthropomorphize.

Outline

- **Experiments**: We probe models of varying degrees of sophistication including random projections, matrix decompositions, deep autoencoders and transformers.
    - All of them successfully distill knowledge and yet none of them develop true understanding.
- **Social Sciences review**: Humans are prone to seek patterns and anthropomorphize.
- **Conclusion and Outlook**: More caution at the individual level, and different incentives at the institutional level.

On the unsurprising nature of LLM embeddings

Questions?

Questions?

With thanks to my co-authors Mojtaba Farmanbar, Arie van Deursen and Cynthia C. S. Liem.

Code

The code used to run the analysis for this work is built on top of `CounterfactualExplanations.jl`.

There is also a corresponding paper, *Explaining Black-Box Models through Counterfactuals*, which has been published in JuliaCon Proceedings.

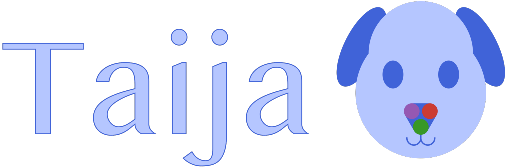


Figure 1: Trustworthy AI in Julia: github.com/JuliaTrustworthyAI

References