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

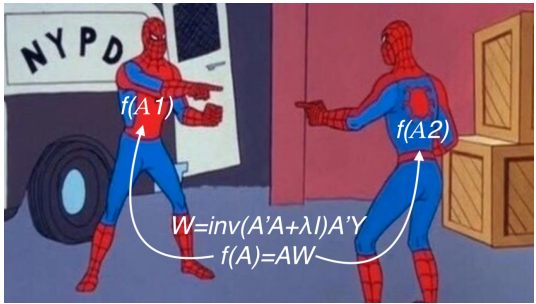
2024-05-07

[1] Upcoming position paper at ICML 2024.

Motivation

- A_1: „It is essential to bring inflation back to target to avoid drifting into deflation territory.“
- A_2: „It is essential to bring the numbers of doves back to target to avoid drifting into dovelation territory.“

“They’re exactly the same.”
— Linear probe $\widehat{cpi} = f(A)$



Position

Current LLMs embed knowledge. They don‘t „understand“ anything. They are useful tools, but tools nonetheless.

- Meaningful patterns in embeddings are like doves in the sky.

Position

Current LLMs embed knowledge. 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.

Position

Current LLMs embed knowledge. 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.

Position

Current LLMs embed knowledge. 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.

Position

Current LLMs embed knowledge. 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.

Outline

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

Outline

- **Experiments**: We probe models of varying complexity 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 complexity 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 complexity 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.

There! It's sentient!

There! It's sentient!

The Holy Grail

Achievement of Artificial General Intelligence (AGI) has become a grand challenge, and in some cases, an explicit business goal.

Definition

The definition of AGI itself is not as clear-cut or consistent:

- (loosely) a phenomenon contrasting with 'narrow AI' systems, that were trained for specific tasks (Goertzel 2014).

Practice

Researchers have sought to show that AI models generalize to different (and possibly unseen) tasks or show performance considered 'surprising' to humans.

- For example, Google DeepMind claimed their AlphaGeometry model (Trinh et al. 2024) reached a 'milestone' towards AGI.

A Perfect Storm

Recent developments in the field have created a ‘perfect storm’ for inflated claims:

- Early sharing of preprints and code.

A Perfect Storm

Recent developments in the field have created a ‘perfect storm’ for inflated claims:

- Early sharing of preprints and code.
- Volume of publishable work has exploded.

A Perfect Storm

Recent developments in the field have created a ‘perfect storm’ for inflated claims:

- Early sharing of preprints and code.
- Volume of publishable work has exploded.
- Social media influencers start playing a role in article discovery and citeability (Weissburg et al. 2024).

A Perfect Storm

Recent developments in the field have created a 'perfect storm' for inflated claims:

- Early sharing of preprints and code.
- Volume of publishable work has exploded.
- Social media influencers start playing a role in article discovery and citeability (Weissburg et al. 2024).
- Complexity is increasing because it is incentivized (Birhane et al. 2022).

“Not Mere Stochastic Parrots”

- We consider a recently viral work (Gurnee and Tegmark 2023a), in which claims about the learning of world models by LLMs were made.
  - Linear probes (ridge regression) were successfully used to predict geographical locations from LLM embeddings.
- Claims on X that this indicates that LLMs are not mere ‘stochastic parrots’ (Bender et al. 2021).
- Reactions on X seemed to largely exhibit excitement and surprise at the authors’ findings.

On the unsurprising nature of latent embeddings