



Anders Søgaard (soegaard@di.ku.dk / 28348409)

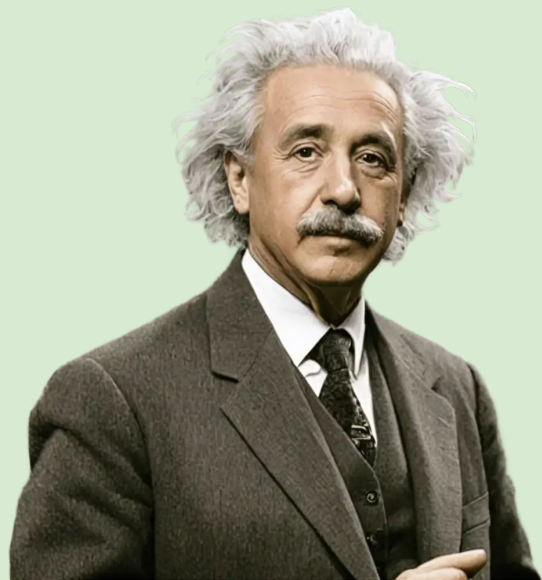
ZOMBIE INTERLOCUTORS: HOW HARD IS LANGUAGE?



CENTER FOR PHILOSOPHY OF
ARTIFICIAL INTELLIGENCE

‘The goal of natural-language processing research can be stated quite simply: to endow computers with human language capability’ (Shieber, 1984)

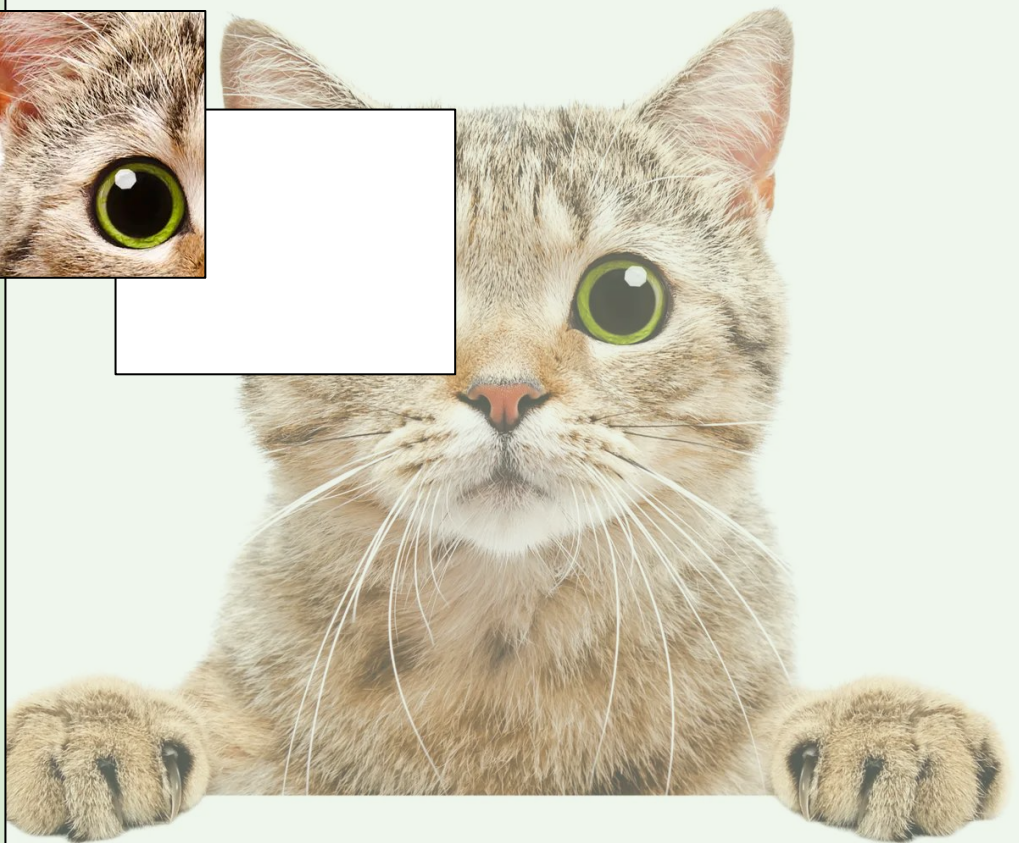
‘to develop robust methods for the understanding and generation of both written and spoken human language’ (Joshi et al., 1994)





نُطَّة.

كِتَابِي فِي الشَّ





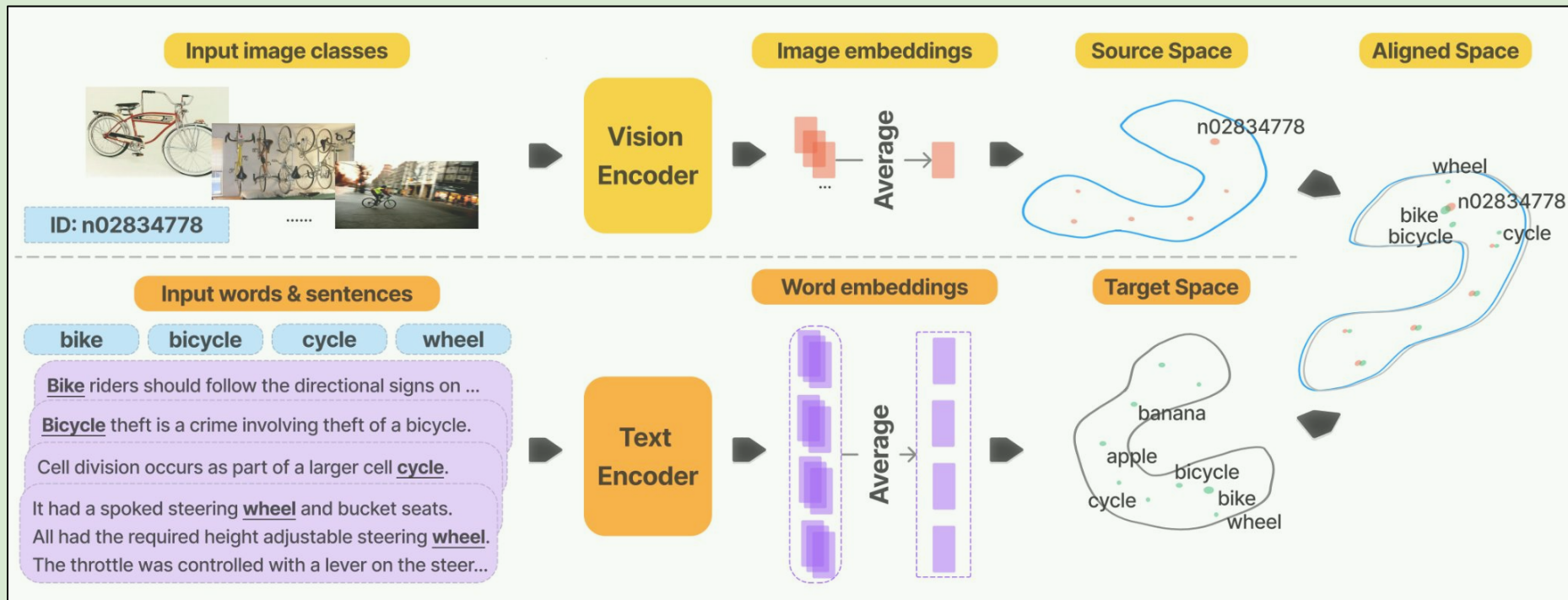
10^{20}

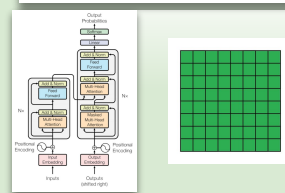
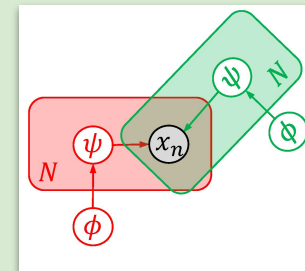
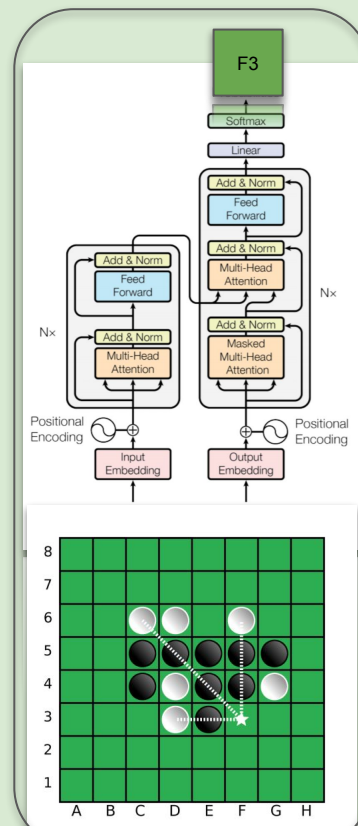
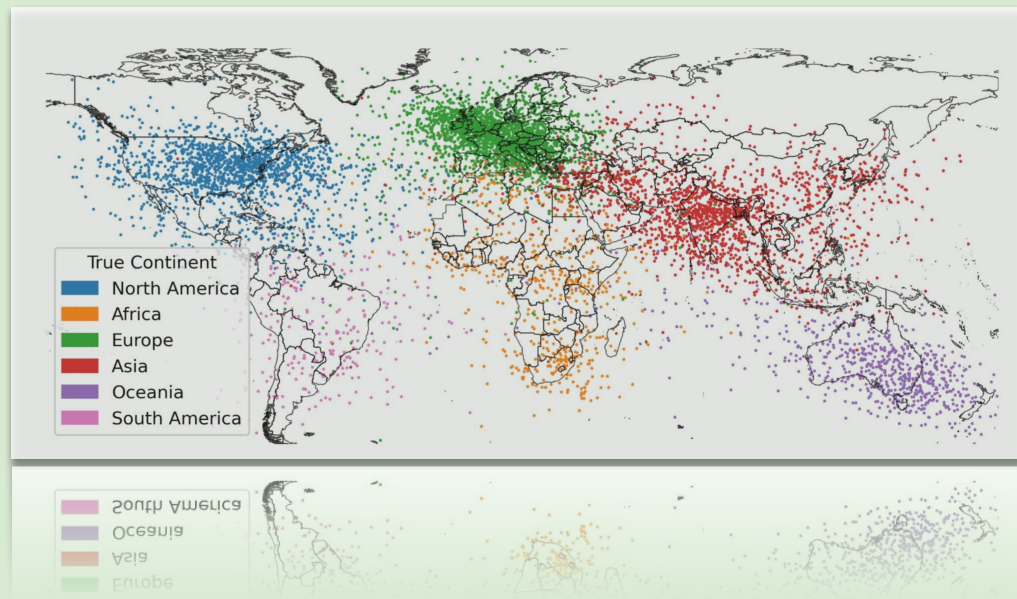


10^{11}



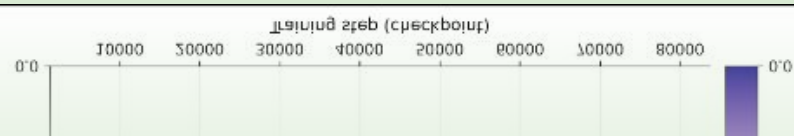
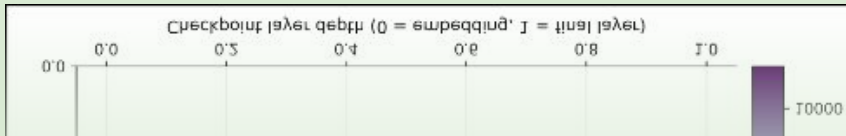
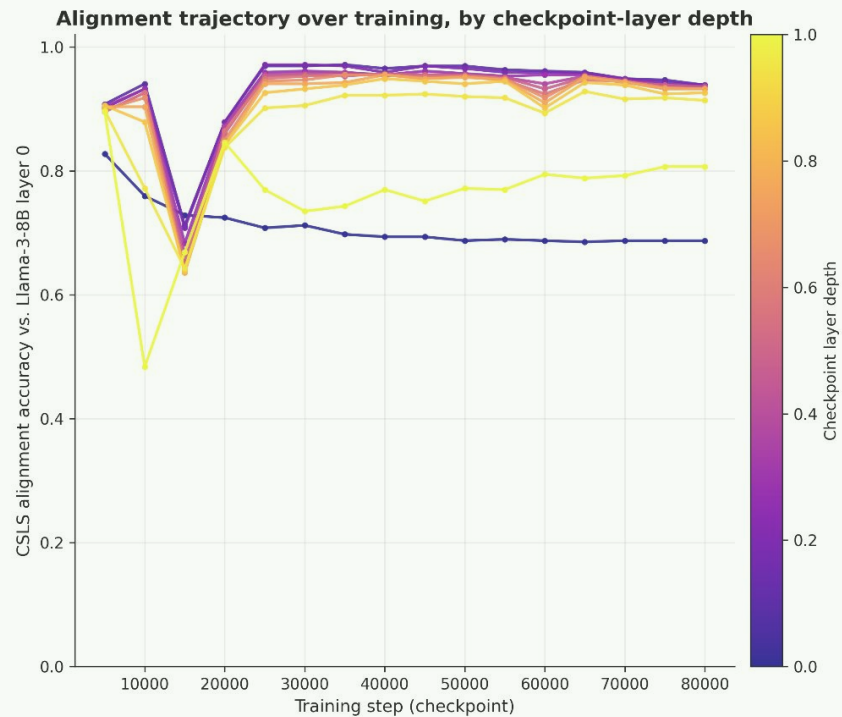
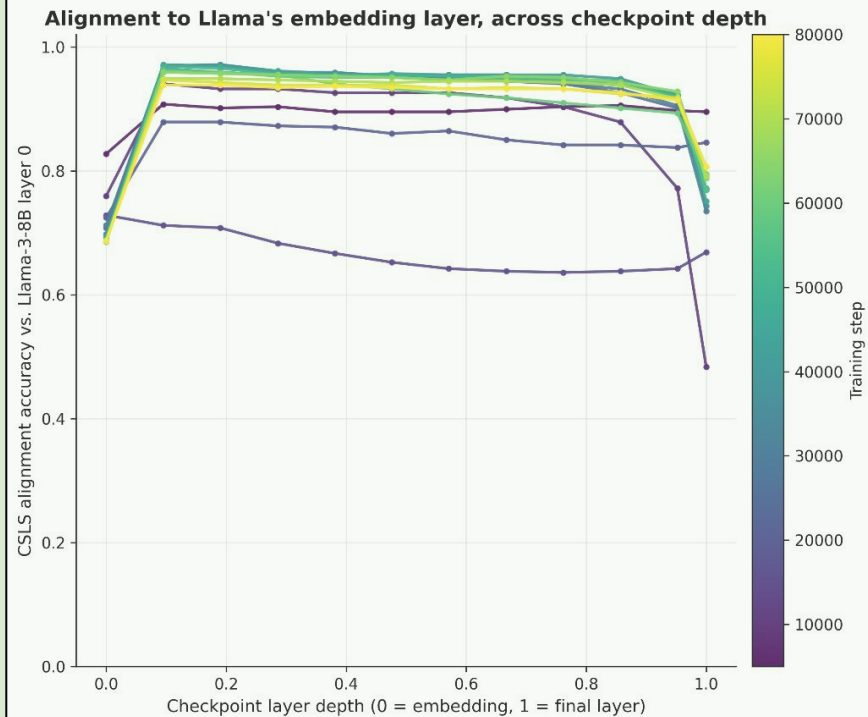
10^9





Left: Fine-tuning on Othello game transcripts. **Right:** Using Procrustes, to evaluate whether the Transformer models induce the board layout.

Pinned-reference alignment: LM1-8.7B checkpoint layers vs. Llama-3-8B's embedding layer (layer 0)



Position 1: Trust the linguists.
Position 2: Don't trust the linguists.

1948
Shannon's BSTJ paper

1986
D-PATR

1993
Brown: IBM Models
Penn Treebank

2013
Word2Vec

2017
Transformers

2020
GPT-3



Position 1: Trust the linguists.
~~Position 2: Don't trust the linguists.~~

1948
Shannon's BSTJ paper

1986
D-PATR

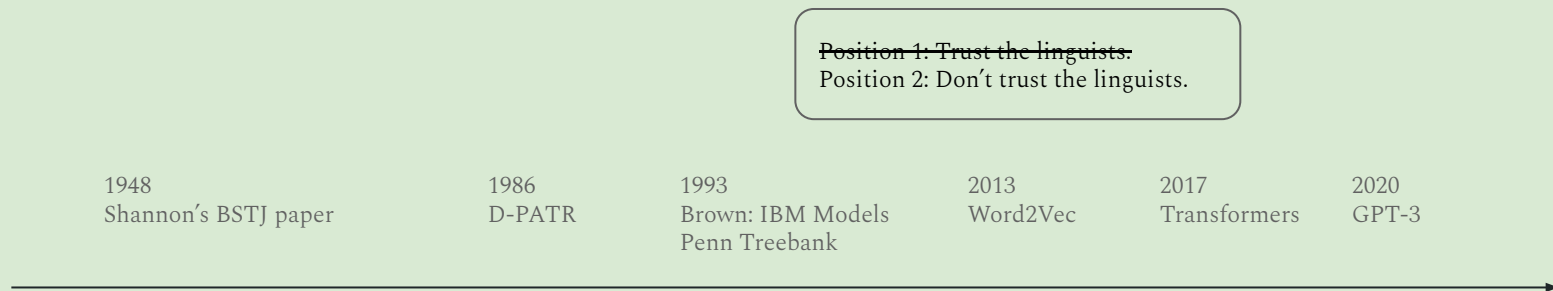
1993
Brown: IBM Models
Penn Treebank

2013
Word2Vec

2017
Transformers

2020
GPT-3







... but LLMs aren't AGI?



Sure, but that's not the game we're playing. AGI - whatever that means - is not the (historical) goal of NLP.



... but LLMs aren't AGI?



Sure, but that's not the game we're playing. AGI - whatever that means - is not the (historical) goal of NLP.

... but LLMs is not the right way of modeling language?

That, again, is a different game. It's called linguistics.



... but LLMs aren't AGI?



Sure, but that's not the game we're playing. AGI - whatever that means - is not the (historical) goal of NLP.

... but LLMs is not the right way of modeling language?

That, again, is a different game. It's called linguistics.

Also, LLMs are biased, unfair, and non-private.

You are moving the goalposts, and none of these problems are specific to NLP.



... but LLMs aren't AGI?



Sure, but that's not the game we're playing. AGI - whatever that means - is not the (historical) goal of NLP.

There's plenty of engineering left.

... but LLMs is not the right way of modeling language?

That, again, is a different game. It's called linguistics.

We can of course change the requirements and do another run, e.g., NLP for mobile devices.

Also, LLMs are biased, unfair, and non-private.

You are moving the goalposts, and none of these problems are specific to NLP.

NLP researchers are in a good position to engage with these problems, but they are not NLP.

Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

Newman (1928): 'It is trivial to establish an isomorphism between sets of equal size.'

Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

Goodman (1968) argues isomorphisms are symmetric, transitive, and reflexive, while representations are not.

Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

Goodman (1968) argues isomorphisms are symmetric, transitive, and reflexive, while representations are not.

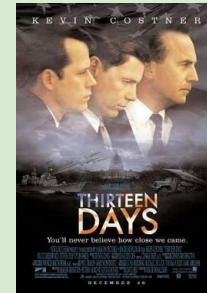
Symmetric: *Would anything prevent the boat from representing the model? Or the landscape from representing the map?*



Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

Goodman (1968) argues isomorphisms are symmetric, transitive, and reflexive, while representations are not.

Transitive: *'Thirteen Days' by JFK is a representation of (his perspective on) the Cuban Missile Crisis (1962) - but what about the movie, based on the book?*



Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

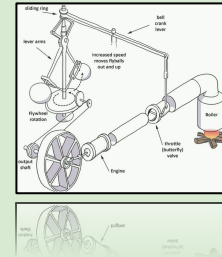
Goodman (1968) argues isomorphisms are symmetric, transitive, and reflexive, while representations are not.

Reflexive: *What, really, would prevent a map from being a representation of itself?*



Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

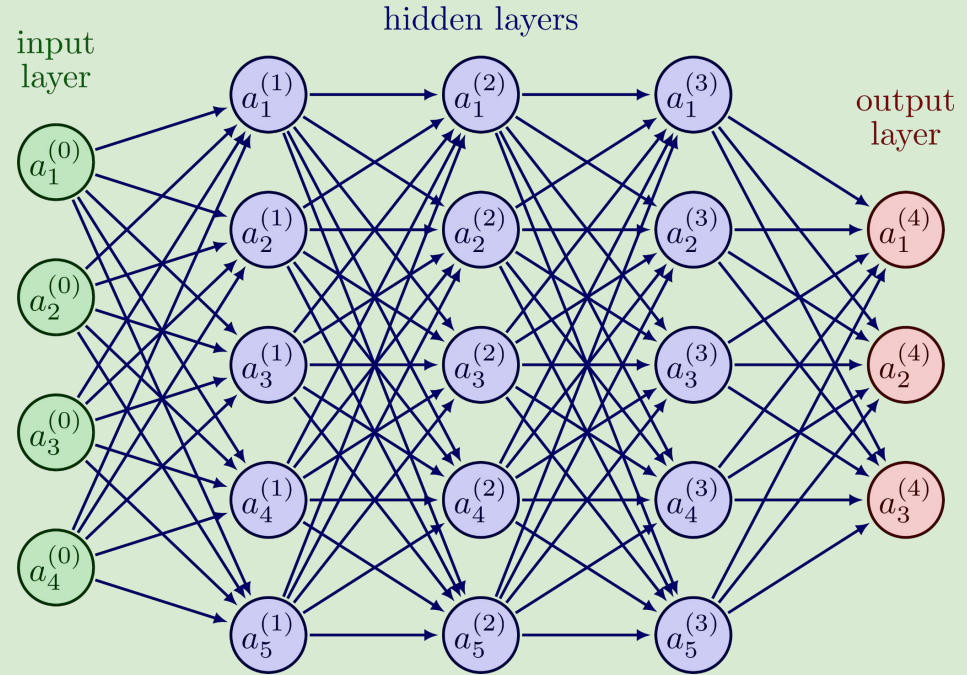
Watt's Governor: Is our theory too permissive?
Do the spindles of a centrifugal governor
understand speed?



Objections from Insufficiency of Modeling	Objections from Scientific Knowledge
Objections from Proper Functions	Objections from Extended Job Descriptions

1. Why do neural networks predict what they do, given the input data?
2. How do neural networks settle on the functions they settle on, given the training data?

-  Inference-opacity
-  Training-opacity



Incompleteness

Since (unfathomable) DNNs can be computationally irreducible, it follows that fathomable, faithful transition theories do not always exist.

This means we are bound to accept Popperian trade-offs.

Hooke's Law enables us to predict the force of a compressed spring:

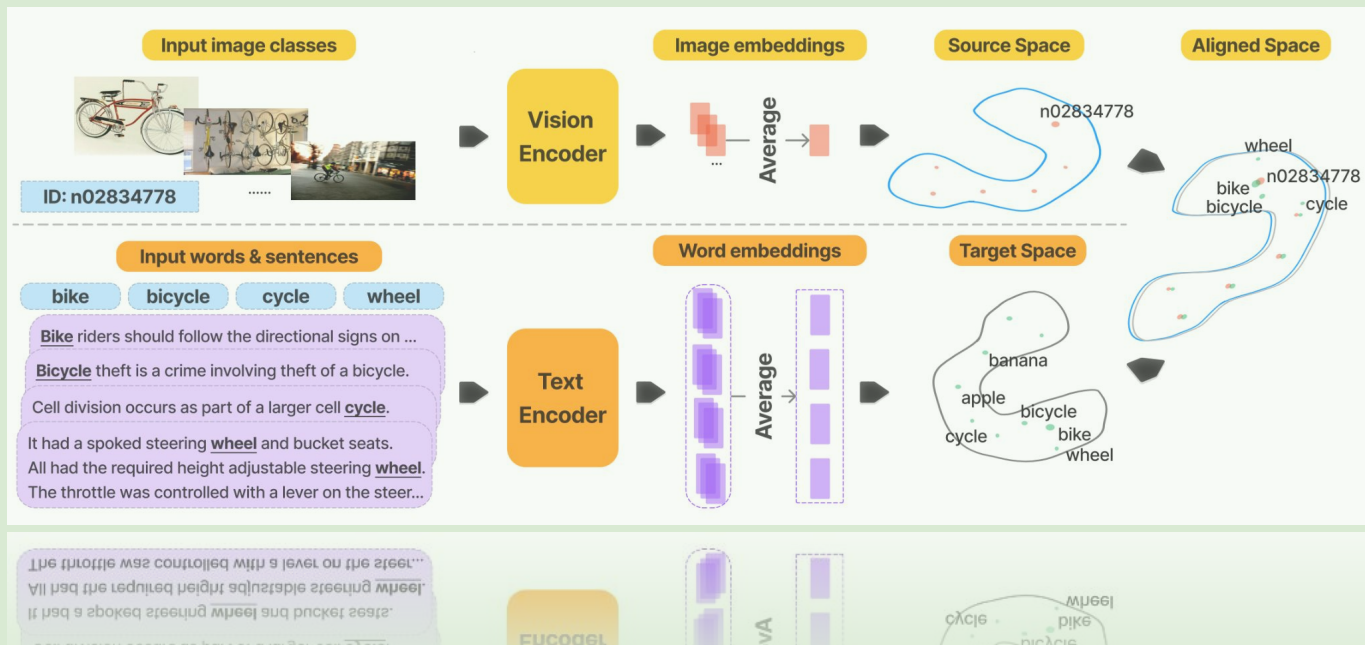
$$F = -k\Delta x$$

but a **property theory** saying that bones are spring-like, because flexibility is adaptive and prevents fracture, also has explanatory value.

	Object	Status	Motivation
Local XAI instance, e.g., a linear model trained using LIME	DNN Inference	A transition theory of a DNN inference.	Prevent harm, provide explanation.
Global XAI instance, e.g., uptrained model	DNN Inference Distribution	A transition theory of a DNN's probable inferences.	Prevent harm.
XAI method, e.g., LIME or LRP	Family of DNNs Family of distributions	A meta-theory or methodology	Develop tools for preventing harm.
Most things MI	DNN Inference Distribution (DNN)	A property probe	Prevent harm.

Example: Need for compression facilitates representation

Shea (2018): Presentation turns on unmediated exploitable structural similarity. This is a concept.



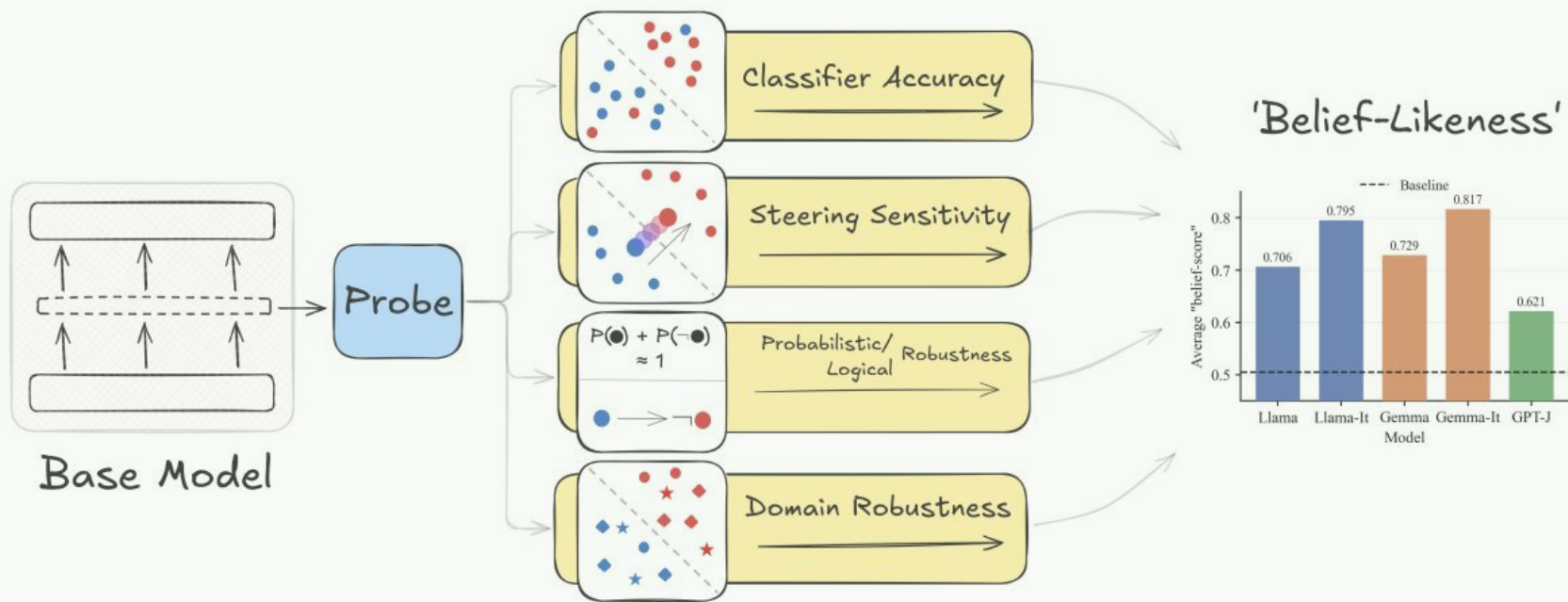
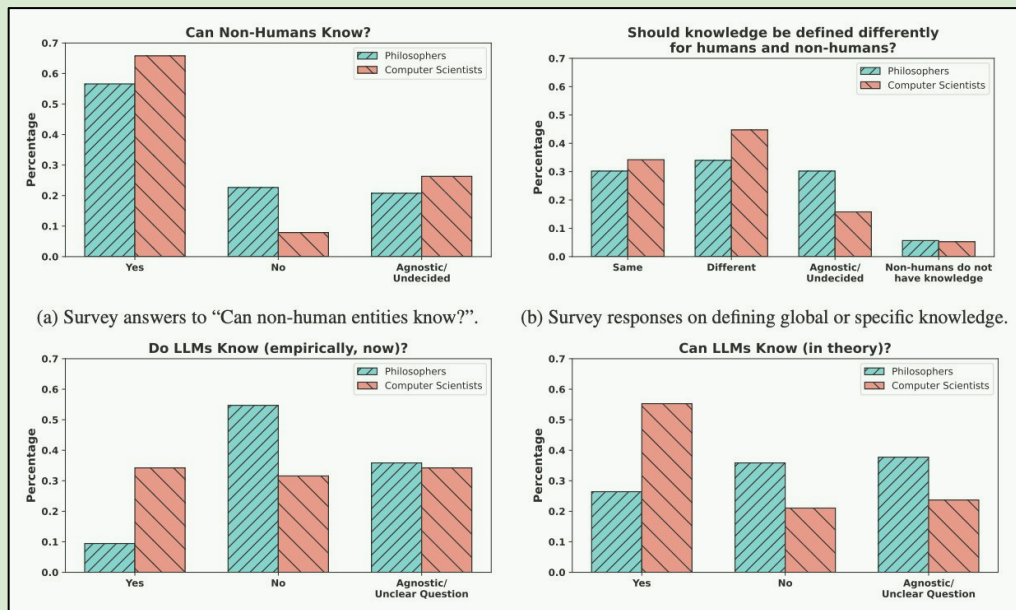


Figure 1: *Standards for Belief Representations* (Herrmann and Levinstein, 2025). To take the role of beliefs, intermediate activations should be distinctly separable by a classifier (*Accuracy*), causally implied in logit computation (*Use*), robust logically and probabilistically (*Coherence*), and uniform across different domains (*Uniformity*).



tb-knowledge p is true, and p is believed+ (Sartwell, 1992)

j-knowledge p is true, p is believed, and p is justified (Nozick, 2000)

g-knowledge p is known *sui generis* (Williamson, 2005)

v-knowledge p is inferred with intellectual virtue (Zagzebski, 1999)

p-knowledge p is believed and facilitates correct predictions (Austin, 2000)

	For	Against
Directive function	During token generation, FVs, OFs and PFs cause the outcomes that they represent. • The directive role of FVs, OFs and PFs is plausibly selected for in reinforcement-tuned systems.	The directive role of FVs, OFs and PFs is not selected for in purely pretrained systems.
Commitment	The self-attention mechanism means representations can be carried forward “across time” (from earlier to later token positions).	PFs and OFs often occur alongside other PFs/OFs representing alternative (inconsistent) outcomes, which also causally influence model output behaviour.

20..

Agency and consciousness

Some deflationary accounts permit protocols, but things are hand-wavy at best..

20..

Knowledge and intentions

For some definitions, we have protocols and some evidence.

2026

Beliefs

For some definitions, we have protocols and solid evidence.

2023

Understanding

Concept geometries emerge from next-token prediction. Robust structural similarities across languages and modalities, and with human representations.

Note: NLP only asked for understanding.

20..

Agency and consciousness

Some deflationary accounts permit protocols, but things are hand-wavy at best..

20..

Knowledge and intentions

For some definitions, we have protocols and some evidence.

2026

Beliefs

For some definitions, we have protocols and solid evidence.

2023

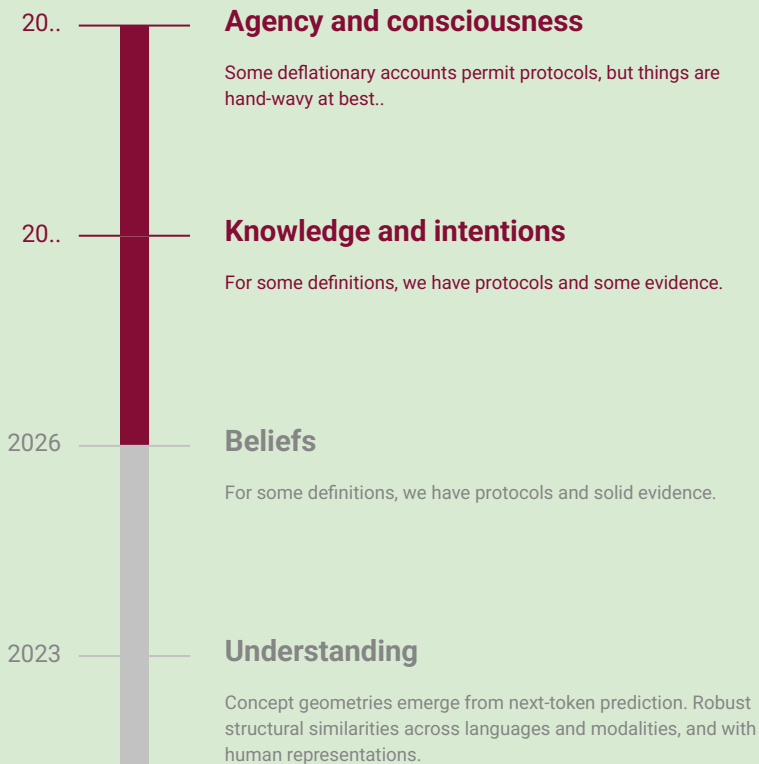
Understanding

Concept geometries emerge from next-token prediction. Robust structural similarities across languages and modalities, and with human representations.

‘The goal of natural-language processing research can be stated quite simply: to endow computers with human language capability’ (Shieber, 1984)

‘to develop robust methods for the understanding and generation of both written and spoken human language’ (Joshi et al., 1994)

Note: NLP only asked for understanding.





Anders Søgaard (soegaard@di.ku.dk / 28348409)

ZOMBIE INTERLOCUTORS: HOW HARD IS LANGUAGE?



CENTER FOR PHILOSOPHY OF
ARTIFICIAL INTELLIGENCE