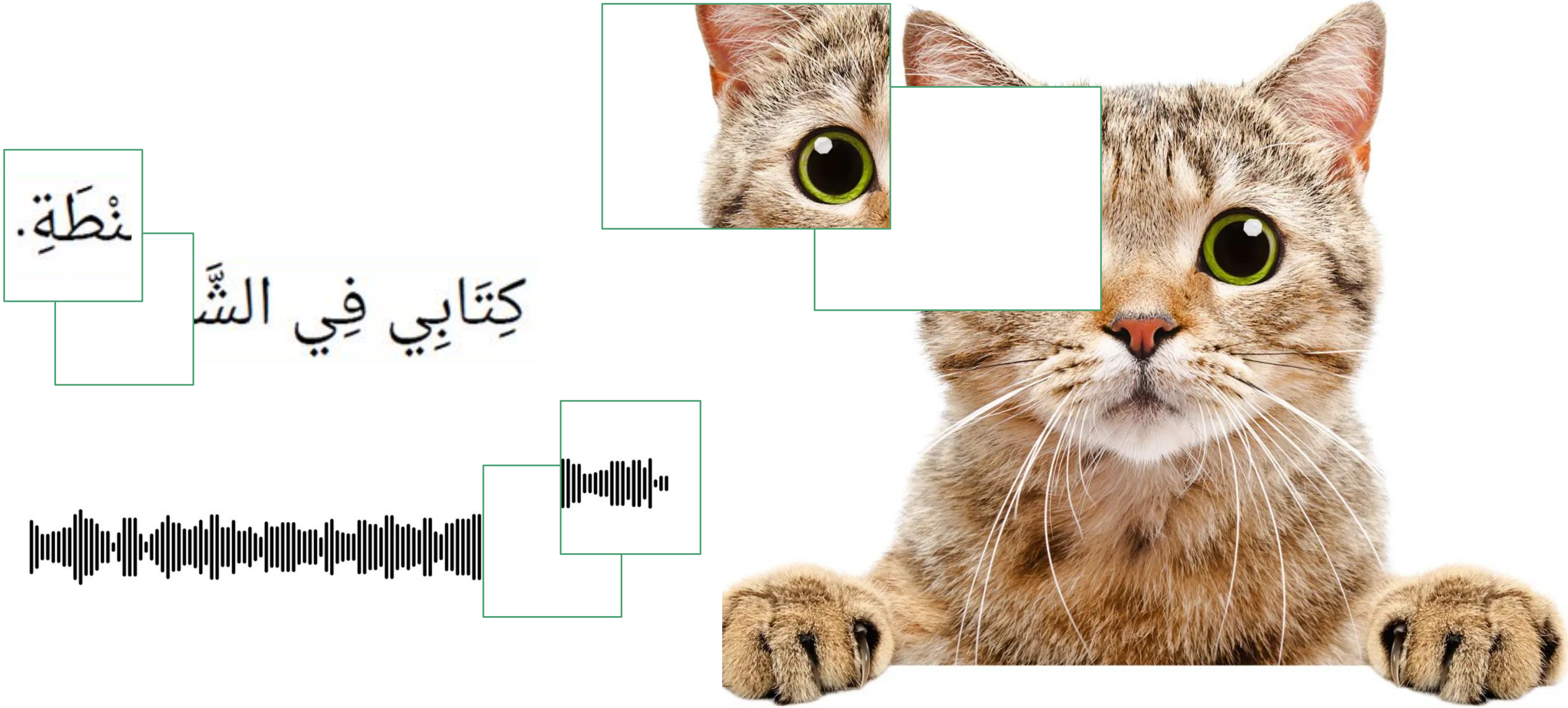
Hvorfra ved vi, hvad en sprogmodel ved?

Anders Søgaard

Forstår sprogmodeller?

$$f(x)$$

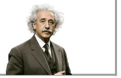


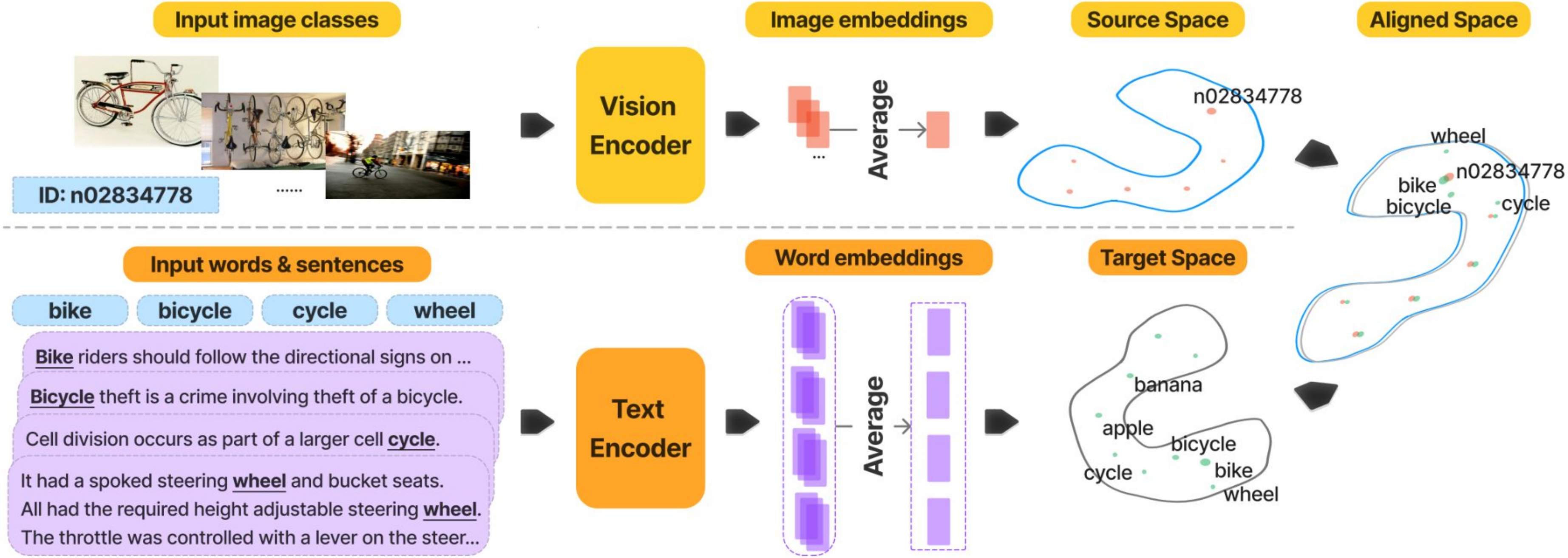
كِتَابِي فِي الشَّ

نْطَةِ.

Fakta: Hovedstaden i Rwanda er Kigali.

Spørgsmål	**Svar**
Hovedstaden i Rwanda er _____.	?
Hovedstaden i Sverige er _____.	?

Spørgsmål	**Svar**
Hovedstaden i Rwanda er _____.	Kigali
Hovedstaden i Sverige er _____.	Stockholm 

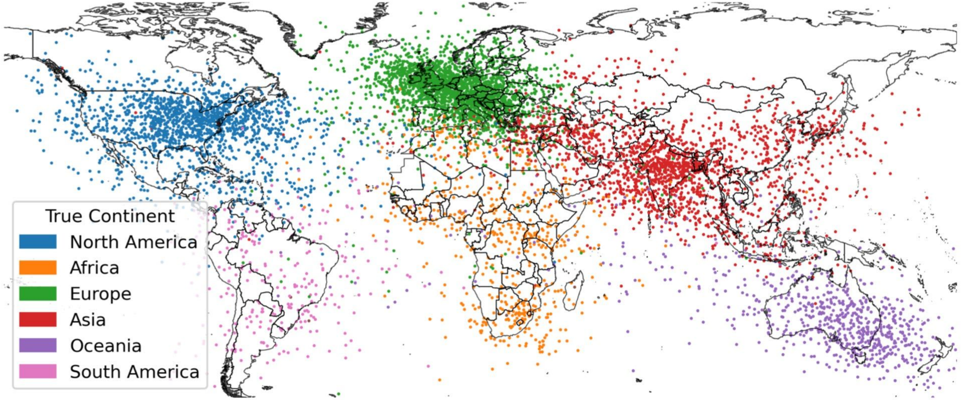


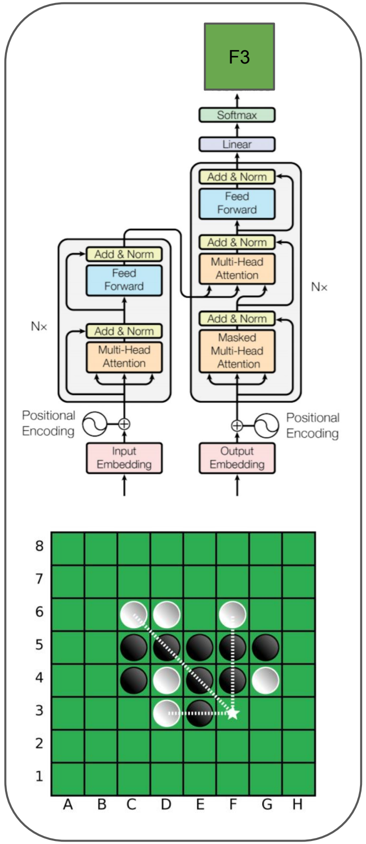
Bike riders should follow the directional signs on ...

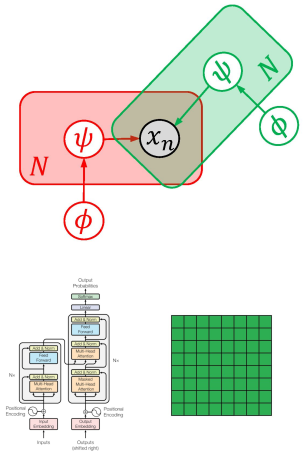
Bicycle theft is a crime involving theft of a bicycle.

Cell division occurs as part of a larger cell **cycle**.

It had a spoked steering **wheel** and bucket seats.
All had the required height adjustable steering **wheel**.
The throttle was controlled with a lever on the steer...







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

Objections from Extended Job Descriptions

Objections from Proper Functions

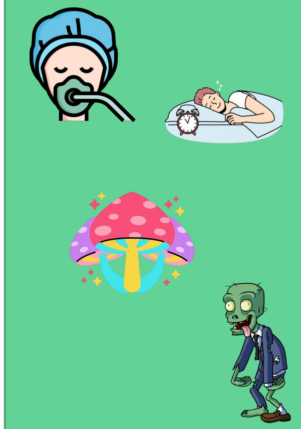
Objections from Insufficiency of Modeling

Hvad betyder det
at forstå ordet
‘filosof’?

Hvad betyder det
at forstå ordet
‘filosof’?

“although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by which means we may discover that **they did not act from knowledge,** but only from the disposition of their organs.”
(Descartes)

Mit svar:



“although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by which means we may discover that **they did not act from knowledge,** but only from the disposition of their organs.” (Descartes)

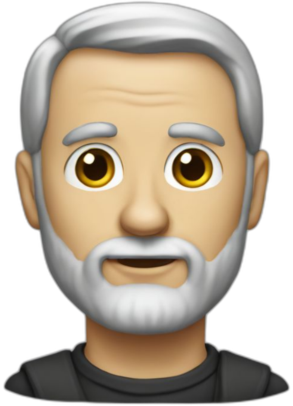
Hvad betyder det at forstå ordet ‘filosof’?

Mit svar: Extended Job Descriptions følger ikke vores intuitive forståelse af forståelse, giver os intet og komplicerer vores begreber.

Bevidsthed

Fælles opmærksomhed

Fælles opmærksomhed



Objections from Extended Job Descriptions

Objections from Proper Functions

Objections from Insufficiency of Modeling

Mechanical philosophers

Logical positivists

Cybernetics, GOFAI

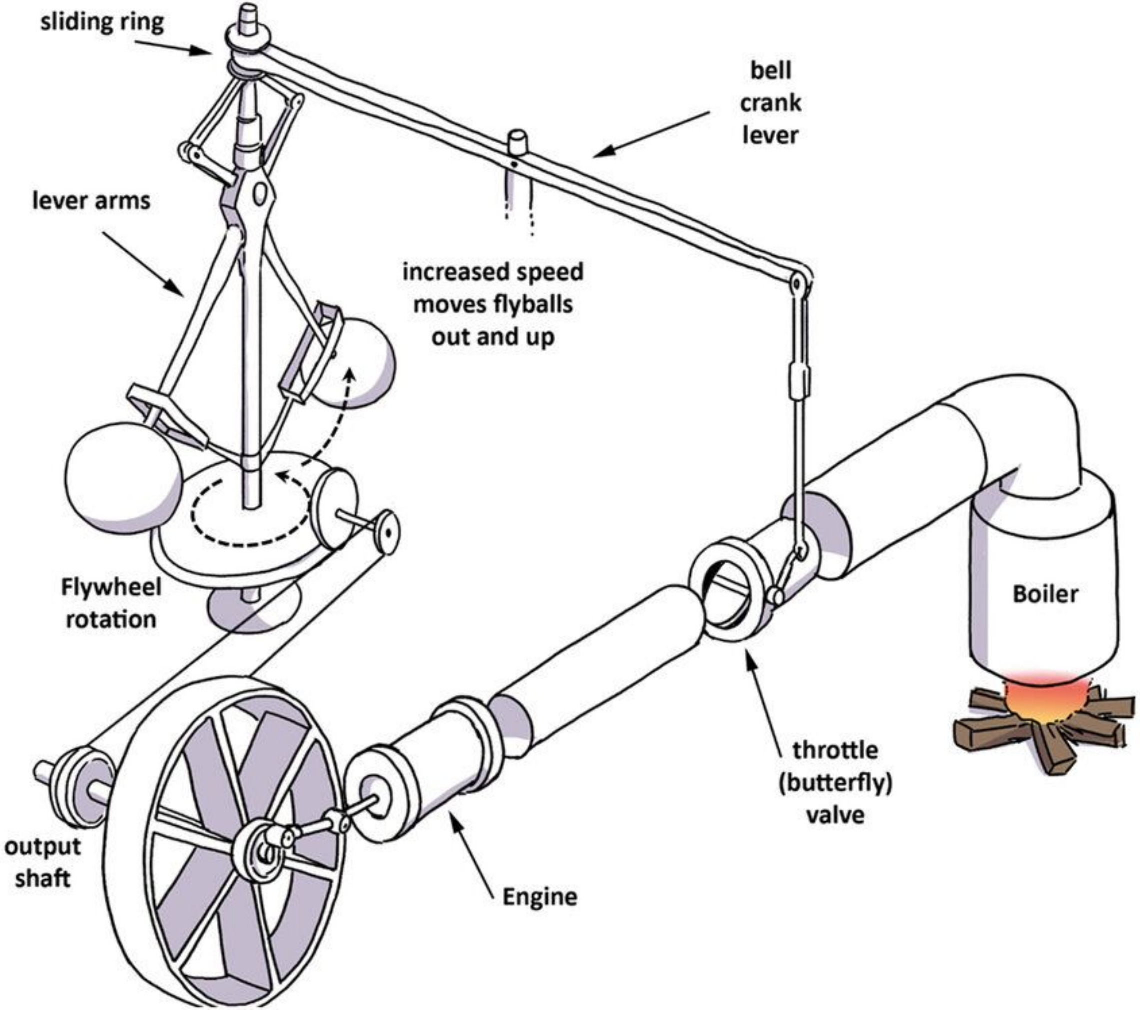
LLMs

Watt's centrifugal-regulator

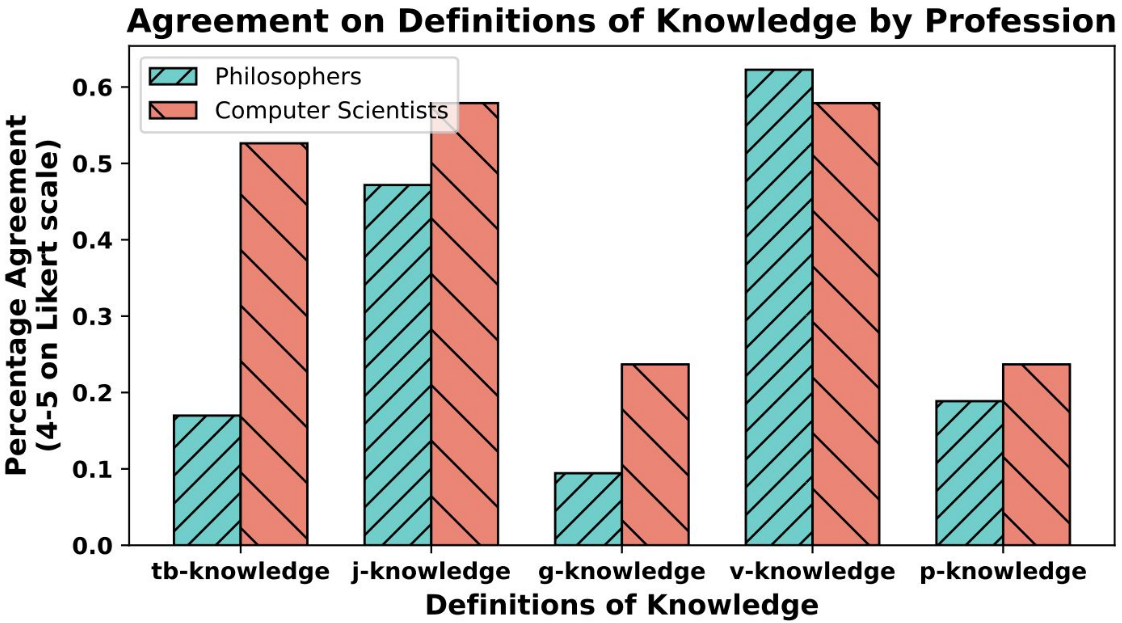
Er vi nødt til så at sige, at alt fra vandure til centrifugal-regulatorer *forstår*?

Mit svar:

Nej: Centrifugalregulatoren har ikke en verdensmodel. Måske en model, men ikke en verdensmodel.



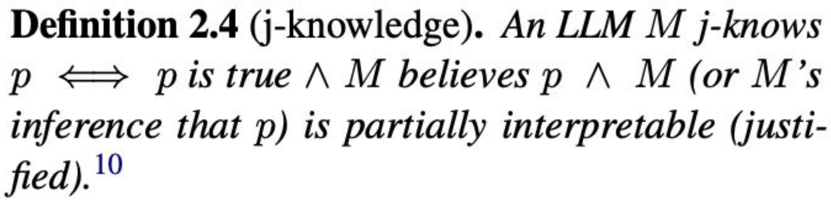
Ved sprogmodeller?

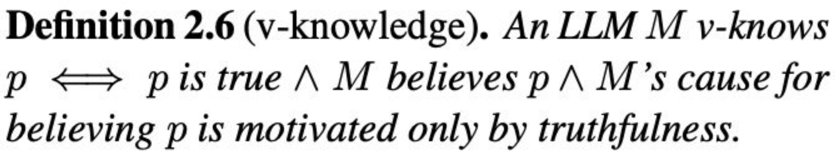


Sprogmodellers j-viden

Hase et al. (2023): belief consistency under paraphrasing and entailment.

K and Søgaard (2023): training data attribution.


Definition 2.4 (j-knowledge). *An LLM M j-knows $p \iff p$ is true $\wedge$ M believes p $\wedge$ M (or M's inference that p) is partially interpretable (justified).*[10]


Definition 2.6 (v-knowledge). *An LLM M v-knows $p \iff p$ is true $\wedge$ M believes $p \wedge M$'s cause for believing p is motivated only by truthfulness.*

?

Søgaard, Anders. 2023. **Grounding the Vector Space of an Octopus: Word Meaning from Raw Text**. **Minds and Machines** 33 (1).

Li, Jiaang; Kementchedjhieva, Yova; Fierro, Constanza; Søgaard, Anders. 2024. **Do Vision and Language Models Share Concepts? A Vector Space Alignment Study**. Transactions of the Association for Computational Linguistics (**TACL**).

Fierro, Constanza; Dhar, Ruchira; Stamatiou, Filippos; Garneau, Nicolas; Søgaard, Anders. 2024. **Defining Knowledge: Bridging Epistemology and Large Language Models**. Conference on Empirical Methods in Natural Language Processing (**EMNLP**) 2024. Miami, Florida.