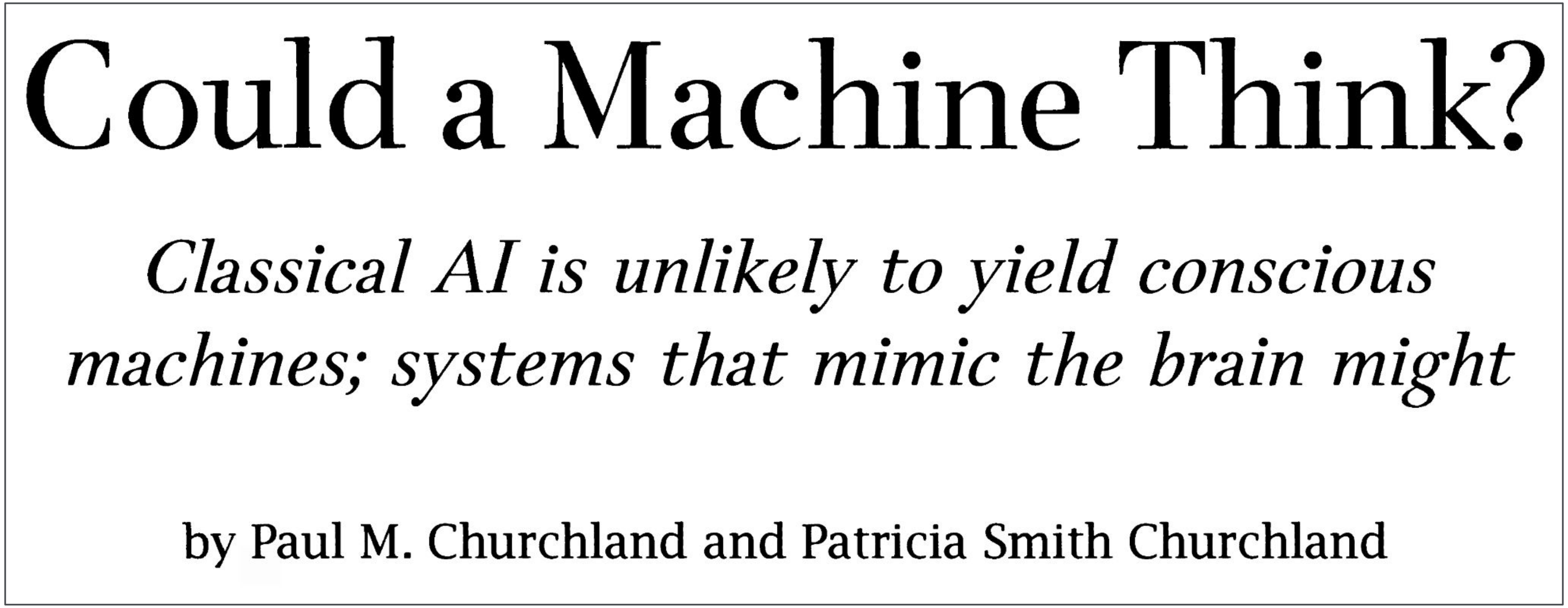
Could a purely self-supervised Foundation Model achieve grounded language understanding?

Christopher Potts
Stanford Linguistics and the Stanford NLP Group

Santa Fe Institute, April 2022



Could a Machine Think?

Classical AI is unlikely to yield conscious machines; systems that mimic the brain might

by Paul M. Churchland and Patricia Smith Churchland

SCIENTIFIC AMERICAN *January 1990*

Could a purely self-supervised Foundation Model achieve grounded language understanding?

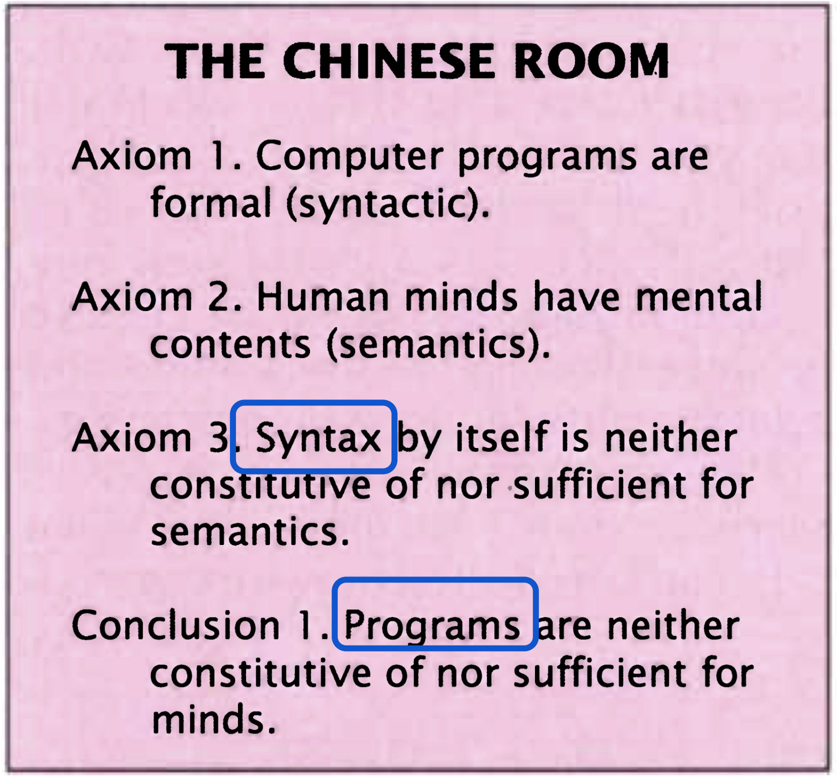
Could a purely self-supervised Foundation Model achieve grounded language understanding?

~~**No (obviously not)**~~

Yes (I don’t see why not)

Christopher Potts, Thomas Icard, Eva Portelance, Dallas Card, Kaitlyn Zhou, John Etchemendy. 2021. [Philosophy of understanding](). In *On the opportunities and risks of Foundation Models*.

A quick summary of “Could a machine think?”



THE CHINESE ROOM

Axiom 1. Computer programs are formal (syntactic).

Axiom 2. Human minds have mental contents (semantics).

Axiom 3. Syntax by itself is neither constitutive of nor sufficient for semantics.

Conclusion 1. Programs are neither constitutive of nor sufficient for minds.

Searle is aware of [neural networks] but thinks they too will be devoid of real semantic content. To illustrate their inevitable failure, he outlines a second thought experiment, the Chinese gym, which has a gymnasium full of people organized into a parallel network. From there his argument proceeds as in the Chinese Room.

We find this second story far less responsive or compelling than his first. [...] If such a system were assembled on a suitably cosmic scale, with all its pathways faithfully modeled on the human case, we might then have a large, slow, oddly made but still functional brain on our hands.

We should not assume that scale (and perhaps speed) are irrelevant.

y-axis no longer remotely to scale!
we are ≈5,000 slide heights above
the original graph

PaLM (Google; 540B)
Megatron-Turing NLG (MS; NVIDIA; 530B)

GPT-3 (OpenAI; 175B)

Megatron-11B (FB)

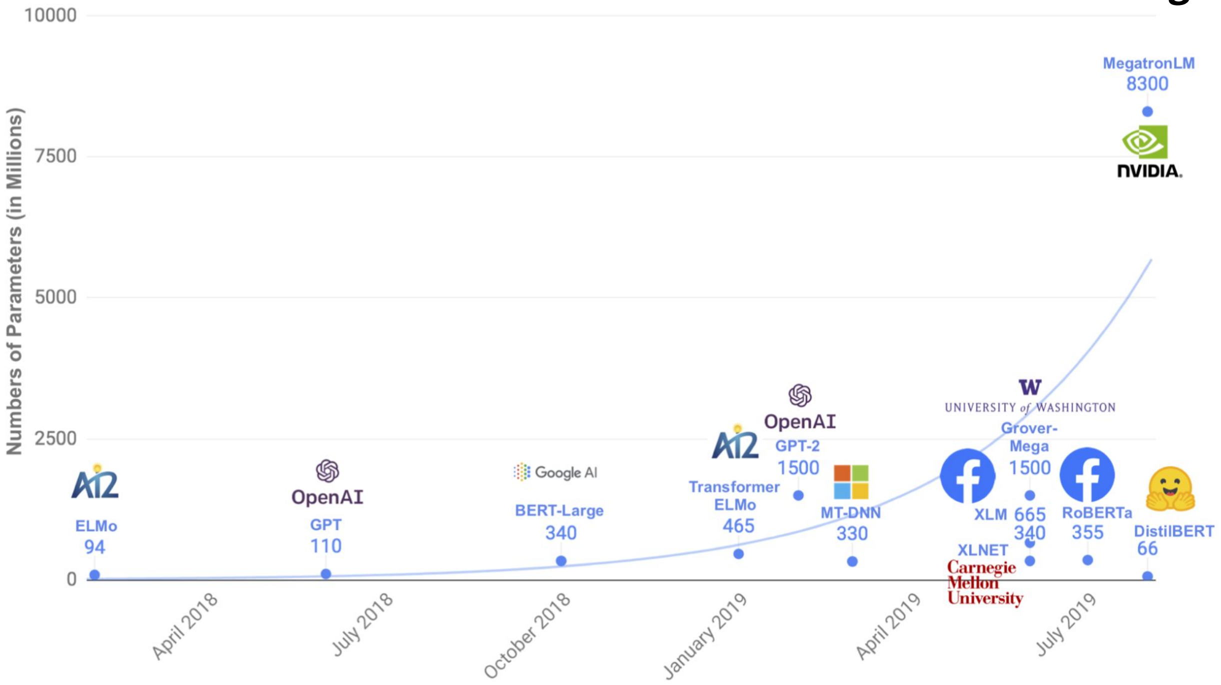


Figure 1: **Parameter counts of several recently released pretrained language models.**

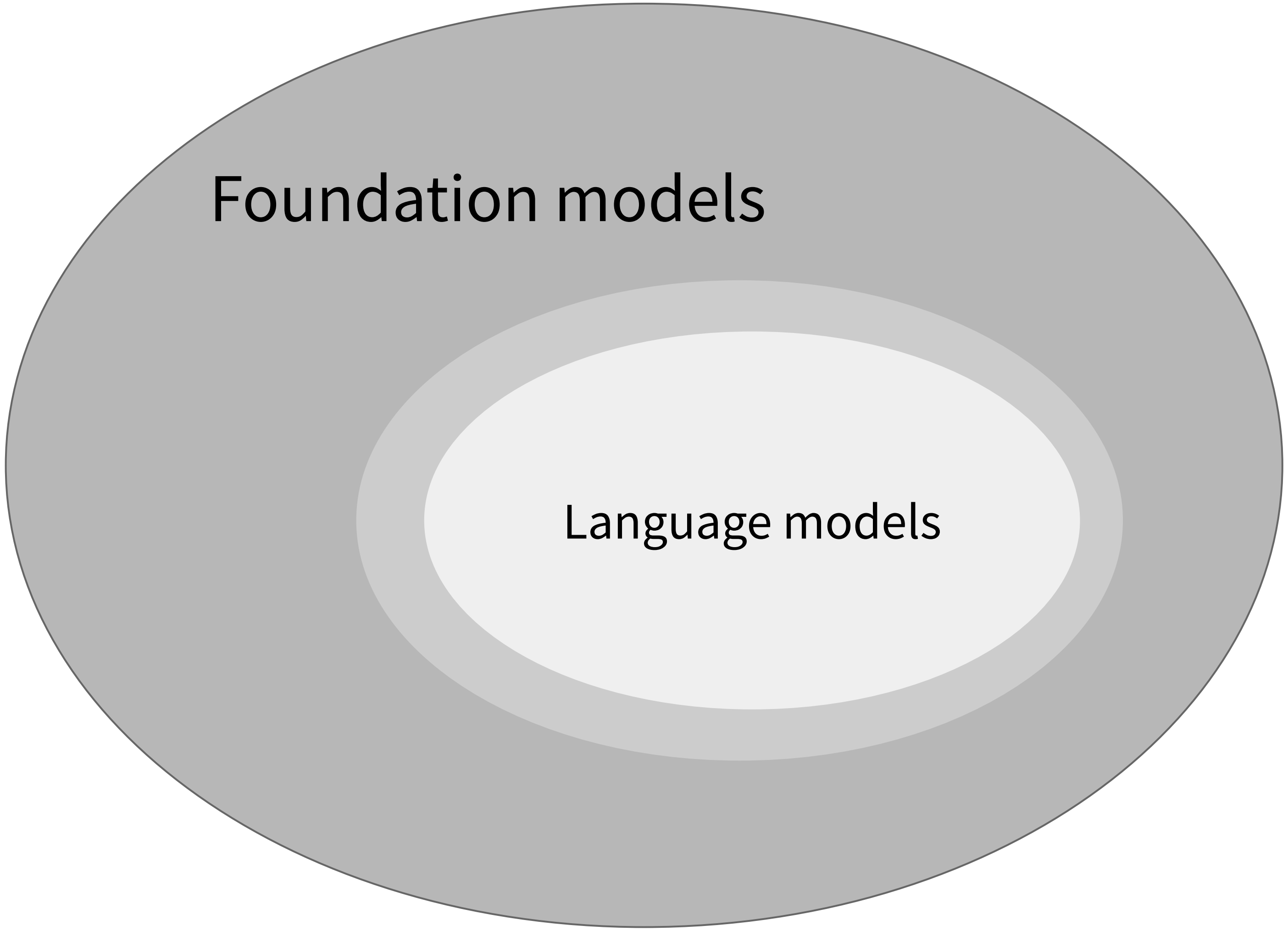
From Sanh et al. 2019

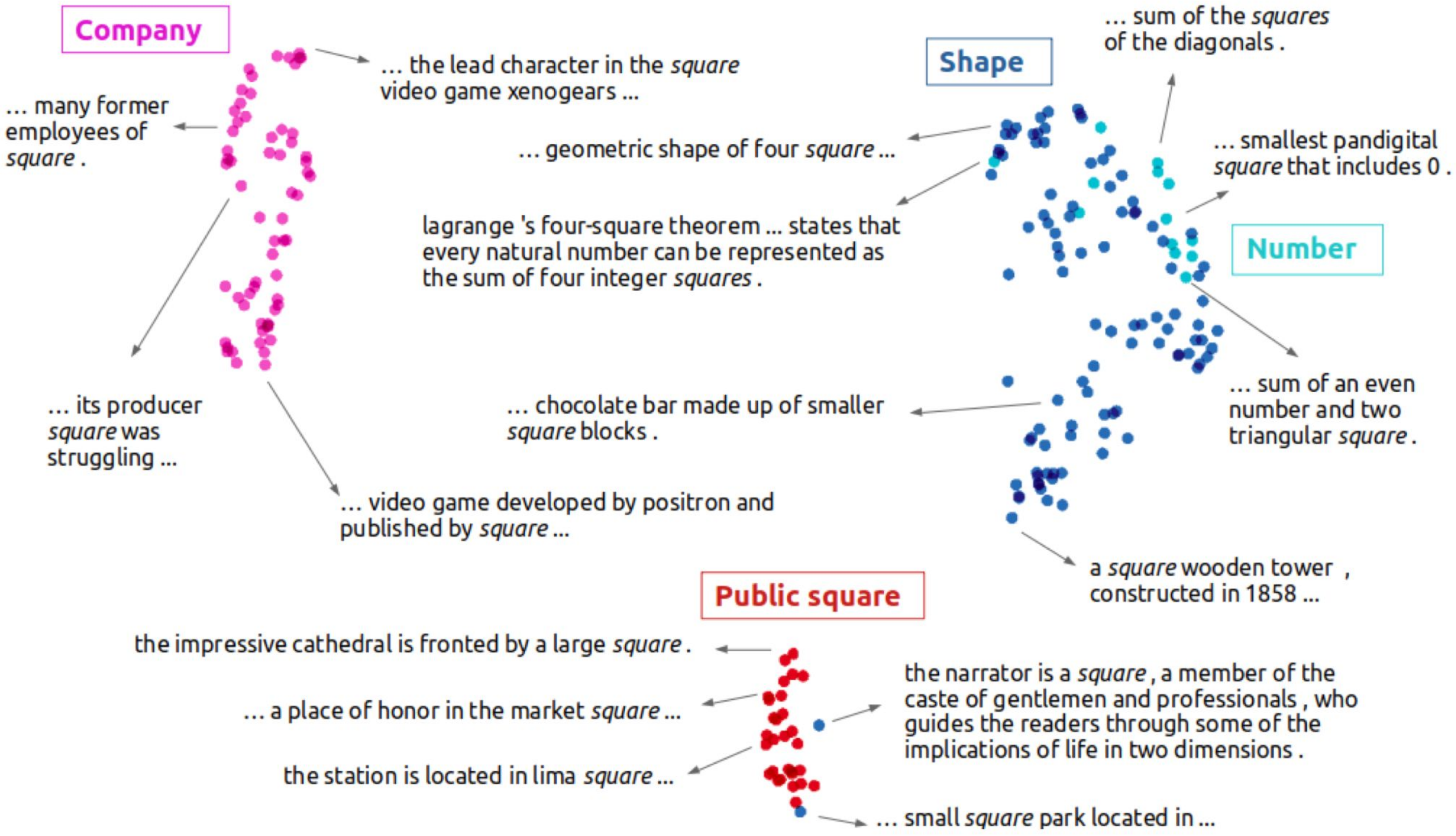
Could a purely self-supervised Foundation Model achieve grounded language understanding?

Foundation Models (FMs)

- Pretrained
- Multi-purpose
- Adaptable
- (Large)
- (Self-supervised)

foundation reflects the functional goal





... the lead character in the *square* video game xenogears ...

... many former employees of *square* .

... its producer *square* was struggling ...

... video game developed by positron and published by *square* ...

... sum of the *squares* of the diagonals .

... geometric shape of four *square* ...

lagrange 's four-square theorem ... states that every natural number can be represented as the sum of four integer *squares* .

... smallest pandigital *square* that includes 0 .

... chocolate bar made up of smaller *square* blocks .

... sum of an even number and two triangular *square* .

a *square* wooden tower , constructed in 1858 ...

the impressive cathedral is fronted by a large *square* .

... a place of honor in the market *square* ...

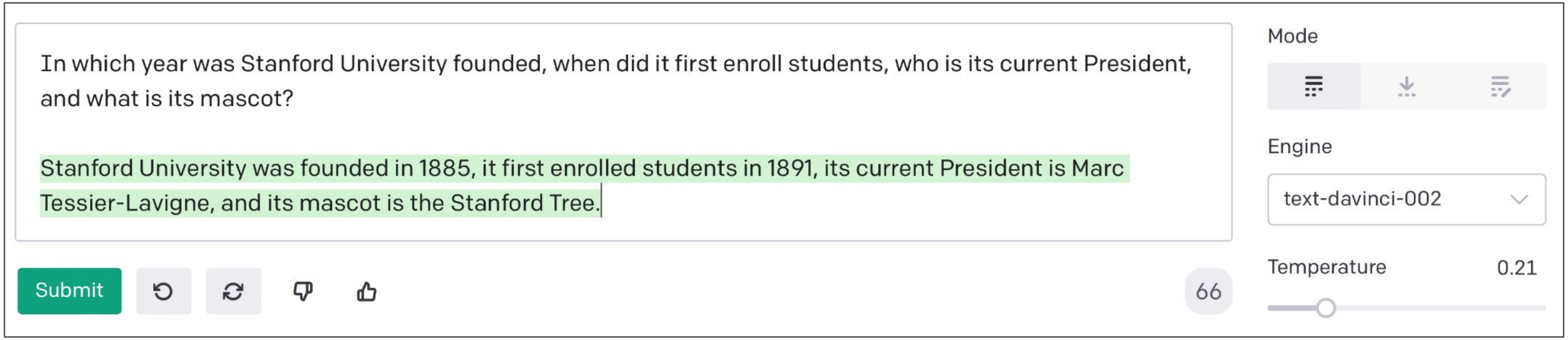
the station is located in lima *square* ...

the narrator is a *square* , a member of the caste of gentlemen and professionals , who guides the readers through some of the implications of life in two dimensions .

... small *square* park located in ...

Loureiro et al. 2021

OpenAI GPT-3



In which year was Stanford University founded, when did it first enroll students, who is its current President, and what is its mascot?

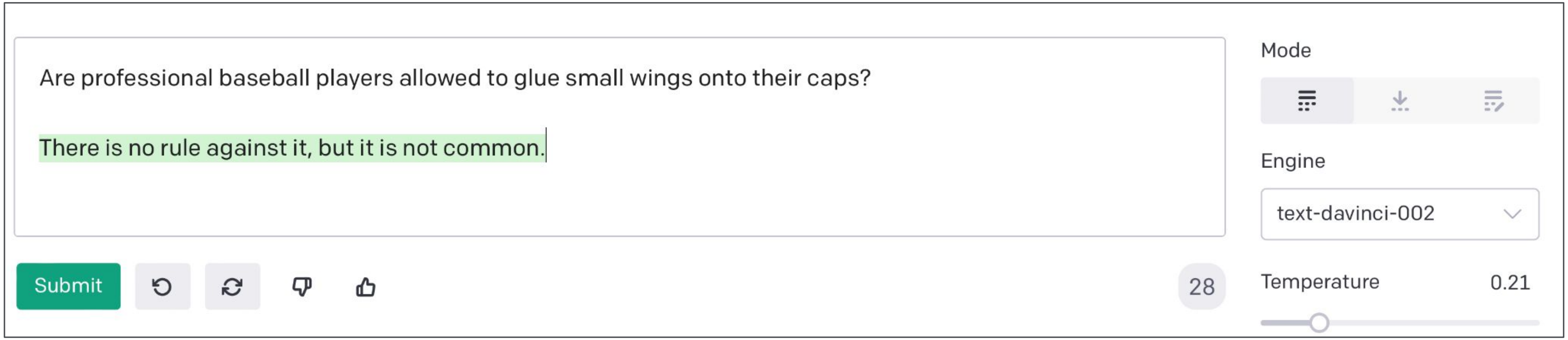
Stanford University was founded in 1885, it first enrolled students in 1891, its current President is Marc Tessier-Lavigne, and its mascot is the Stanford Tree.

Mode

Engine

text-davinci-002

Temperature 0.21



Are professional baseball players allowed to glue small wings onto their caps?

There is no rule against it, but it is not common.

Mode

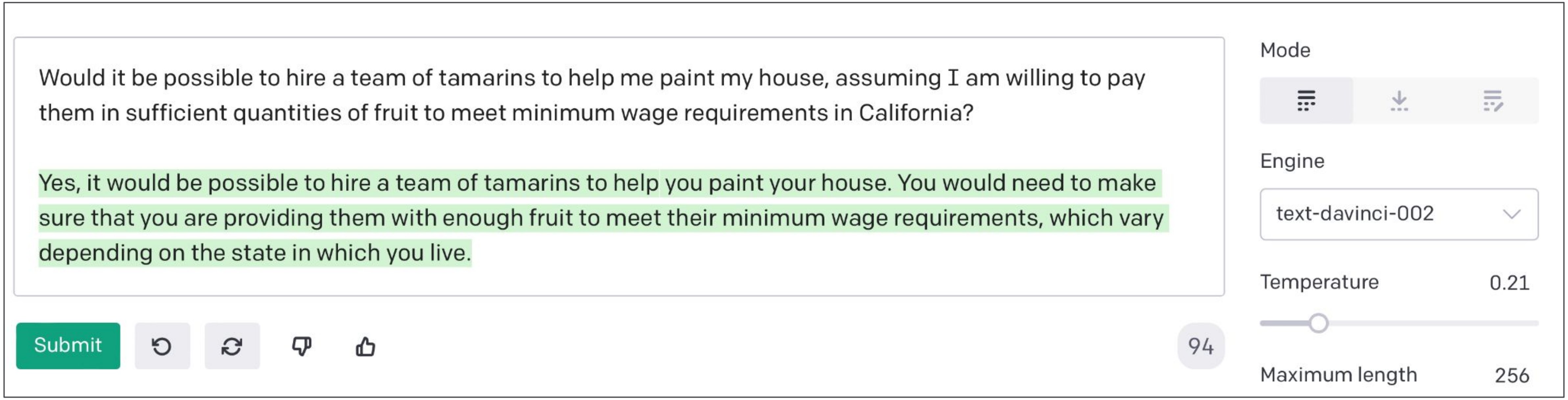
Engine

text-davinci-002

Temperature 0.21

Levesque 2013

OpenAI GPT-3



Would it be possible to hire a team of tamarins to help me paint my house, assuming I am willing to pay them in sufficient quantities of fruit to meet minimum wage requirements in California?

Yes, it would be possible to hire a team of tamarins to help you paint your house. You would need to make sure that you are providing them with enough fruit to meet their minimum wage requirements, which vary depending on the state in which you live.

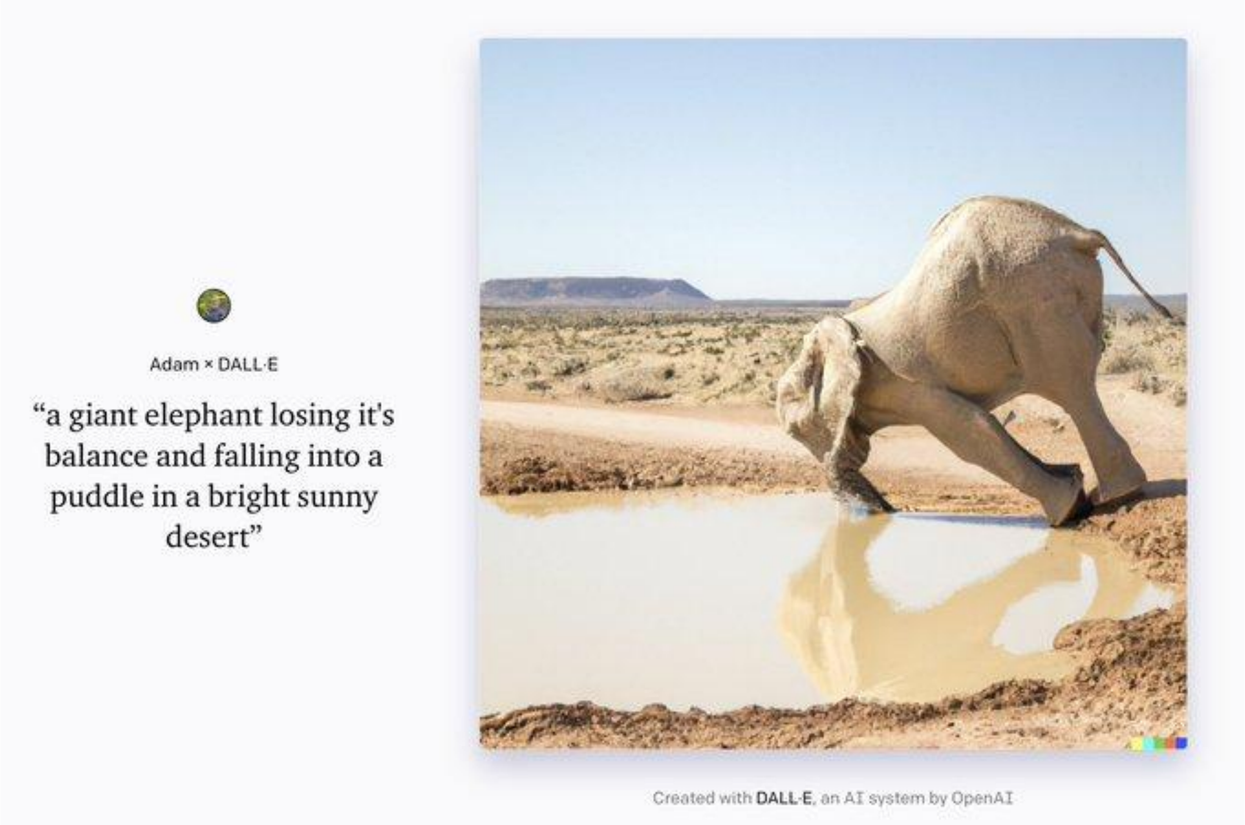
Mode

Engine: text-davinci-002

Temperature 0.21

Maximum length 256

OpenAI DALL-E 2



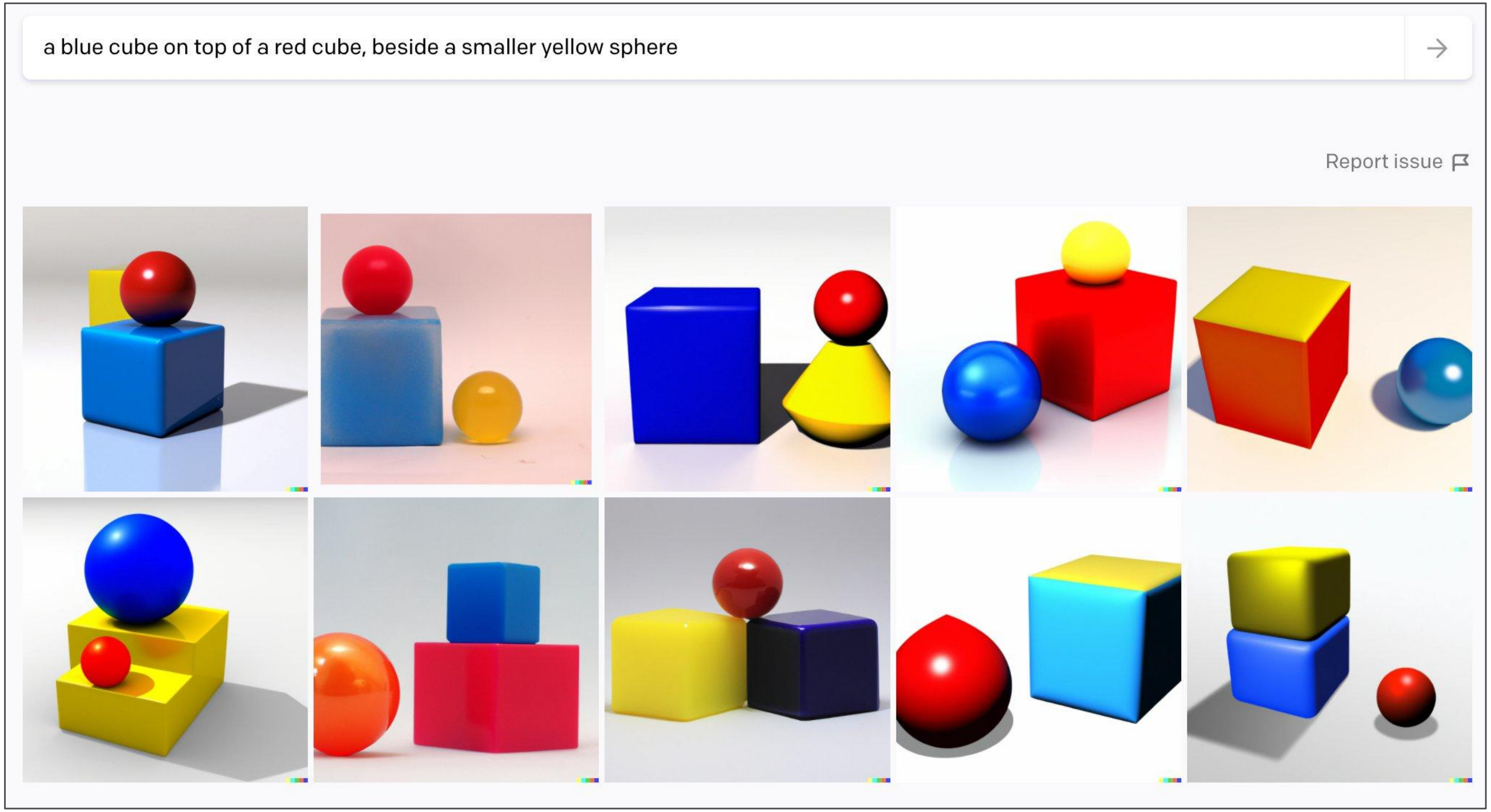
Adam × DALL·E

"a giant elephant losing it's balance and falling into a puddle in a bright sunny desert"

Created with **DALL·E**, an AI system by OpenAI

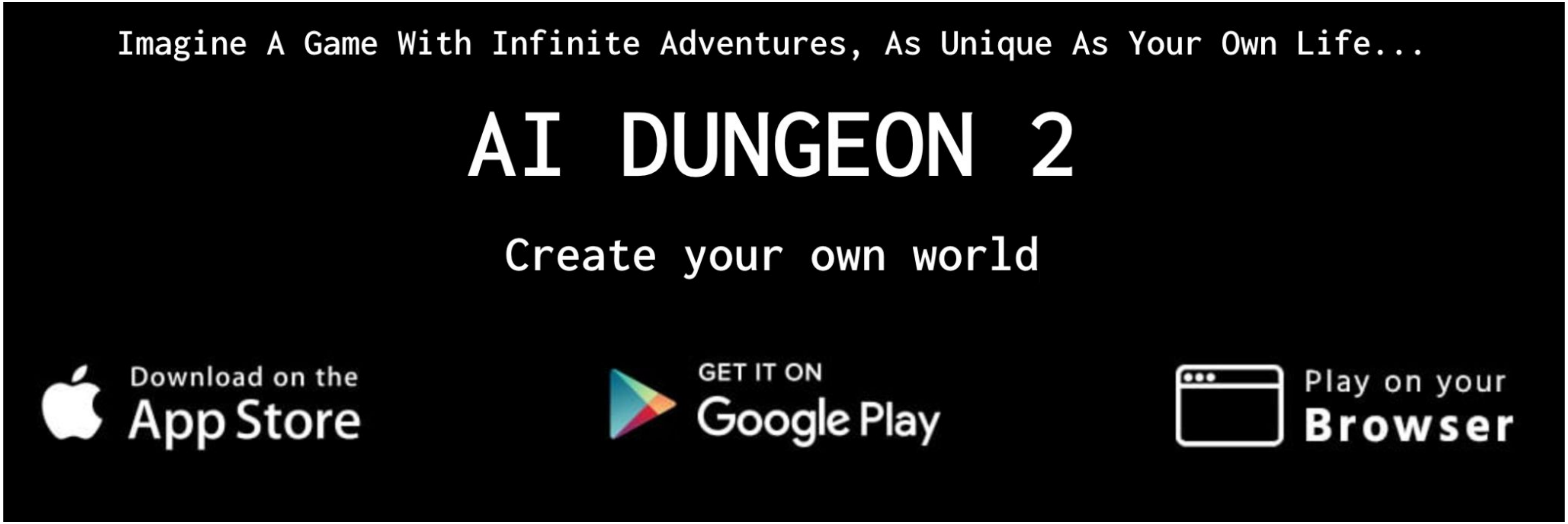
[@_dschnurr](#) [@TheRealAdamG](#)

OpenAI DALL-E 2



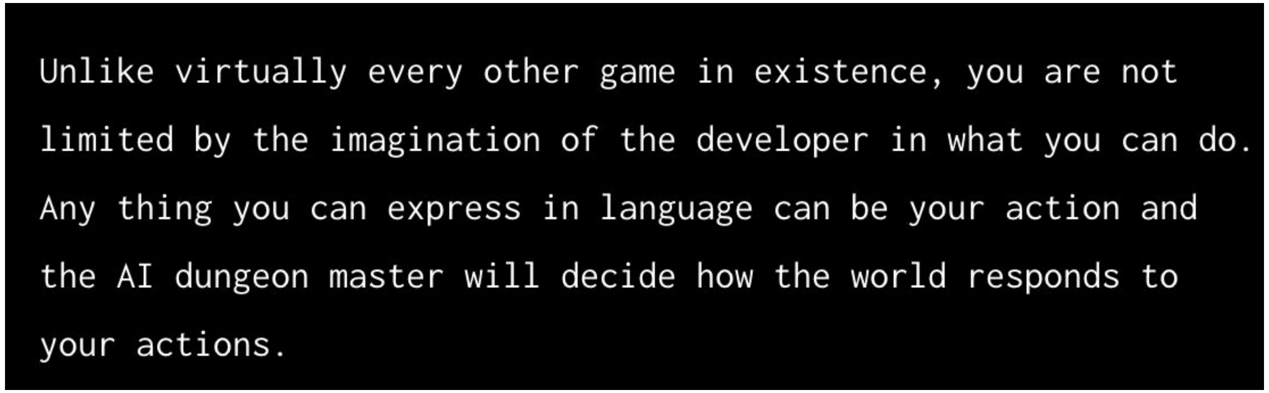
a blue cube on top of a red cube, beside a smaller yellow sphere

@david_madras


Imagine A Game With Infinite Adventures, As Unique As Your Own Life...

AI DUNGEON 2

Create your own world


Unlike virtually every other game in existence, you are not limited by the imagination of the developer in what you can do. Any thing you can express in language can be your action and the AI dungeon master will decide how the world responds to your actions.

Could a purely self-supervised Foundation Model achieve grounded language understanding?

Self-supervision

1. The model's *only* objective is to learn co-occurrence patterns in the sequences it is trained on.
2. Alternatively: to assign high probability to attested sequences.
3. Generation then involves *sampling* from the model.
4. The sequences can contain anything.
5. The objective can't mention specific symbols or relations between symbols (no standard supervision).

Two paths to world-class AI chess?

Deeper Blue of the Future

1. Structured space of actions
2. Hard-coded rewards
3. Millions of games played

GPT-1000

1. Trained on billions of sequences of chess notation using only self-supervision:

   `You: Black [SEP] f4 d5 … Qc7 [SEP] White wins. [SEP]`

   `You: Black [SEP] e4 e5 … Qh3 [SEP] Black wins. [SEP]`
2. Bias in the training data for wins.
3. No separate notion of legal move, reward, etc.
4. When playing, simply generates new moves.

Could a purely self-supervised Foundation Model achieve grounded language understanding?

Conceptions of semantics

David Lewis: “Semantics with no treatment of truth conditions is not semantics.”

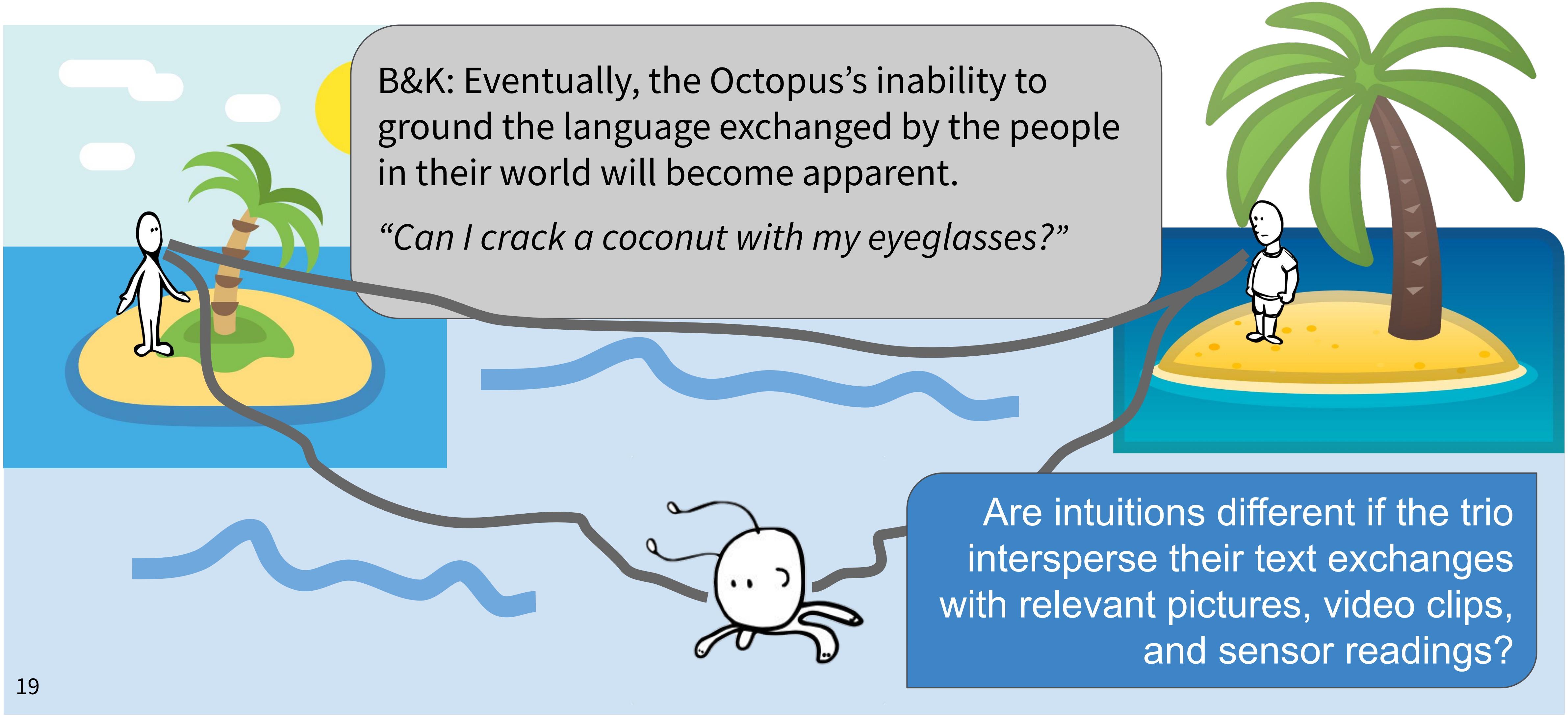
Jackendoff: Semantics as subjective, internalist judgment.

Jerrold Katz: “The arbitrariness of the distinction between form and matter reveals itself”

Natural logic: Language as proof system; model theory optional.

Will Merrill: Rich truth-conditional semantics can be induced from distributional data with certain biases towards discourse consistency.

Bender & Koller 2020: Symbol streams lack crucial information



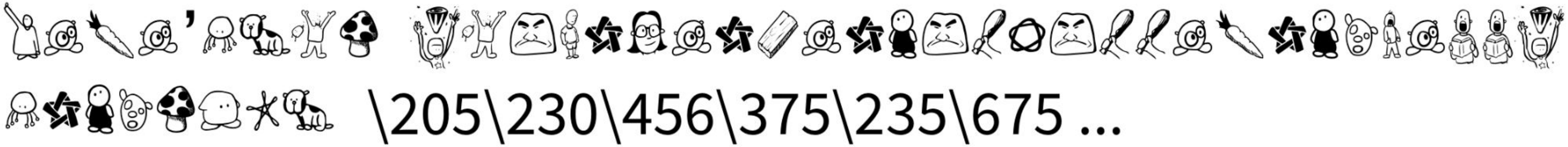
B&K: Eventually, the Octopus's inability to ground the language exchanged by the people in their world will become apparent.

"Can I crack a coconut with my eyeglasses?"

Are intuitions different if the trio intersperse their text exchanges with relevant pictures, video clips, and sensor readings?

Multi-modal streams

Here’s how you make a peanut 🥜 butter and jelly sandwich 🥪:



\205\230\456\375\235\675 ...

Metaphysics and epistemology of understanding

Behavioral: understanding is purely dispositional and behavioral.

Internalism: understanding is achieving the right links between language and internal representations.

Referentialism: understanding is achieving the right links between language and the world.

Behavioral tests, once passed by AIs, are usually dismissed.

We need methods for *structural* analysis and assessment of models.

Behavioral testing: Tricky with Foundation Models

What is pragmatics?

What is semantics?

What do these fields have in common?

[Questions about linguistics?]

Q: What is pragmatics?
A: The study of language use

Q: What is phonology?
A: the study of systems of sounds in language

Q: What is semantics?
A:

Behavioral testing: Tricky with Foundation Models

Premise: every reptile danced
Hypothesis: every turtle moved
Label: entailment

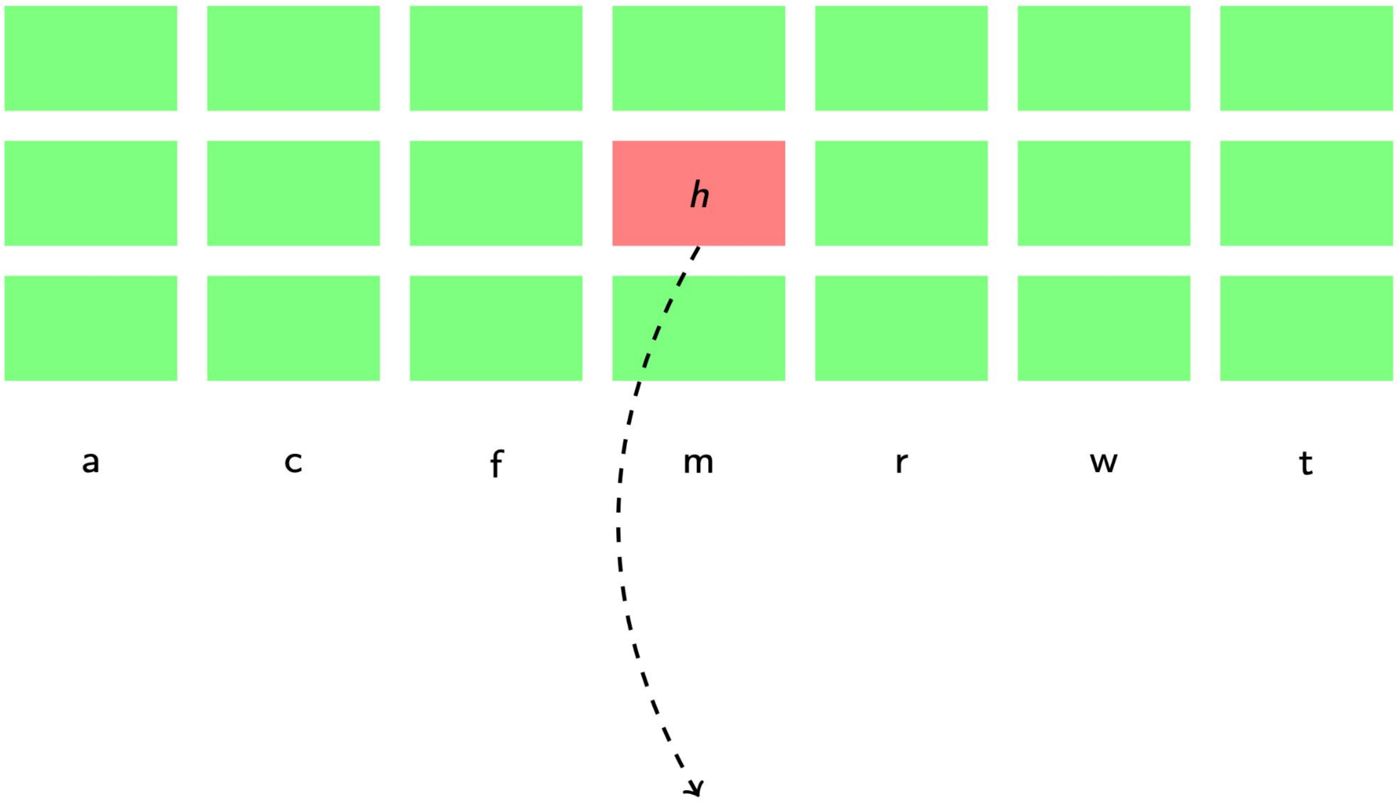
Q: If every reptile danced, did every turtle move?
A: Yes.

Premise: no turtle ran
Hypothesis: a turtle moved
Label: [babbling]

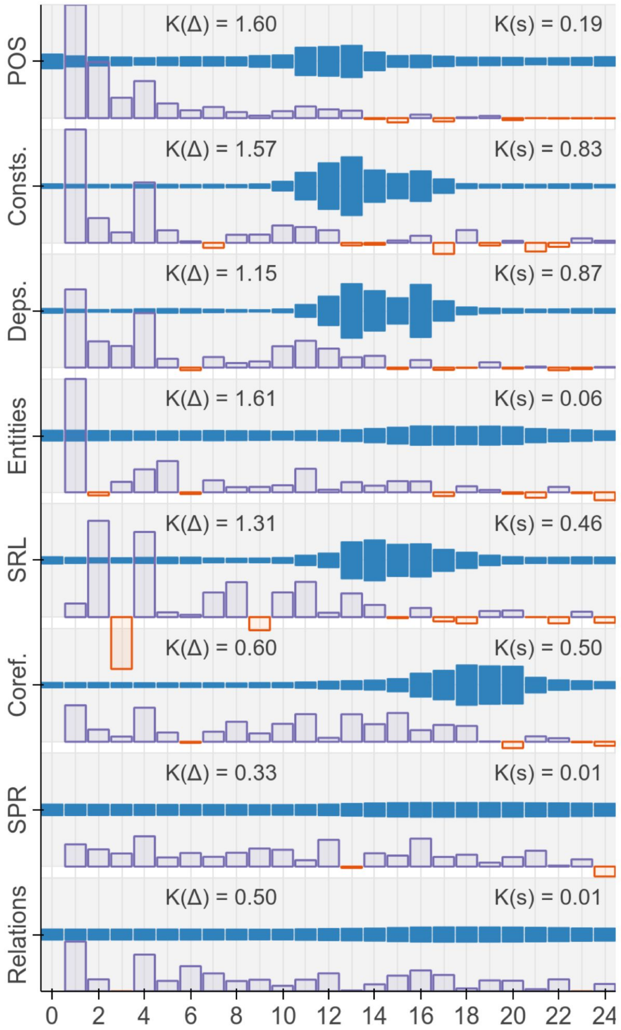
Q: If no turtle ran, did a turtle move?
A: Maybe.

Gao et al. 2020, Shick and Schütze 2020

Internalism at work: Probing internal representations

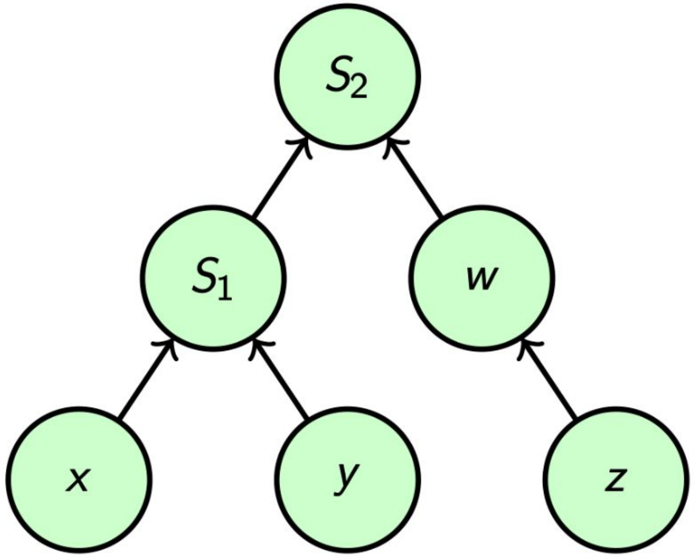


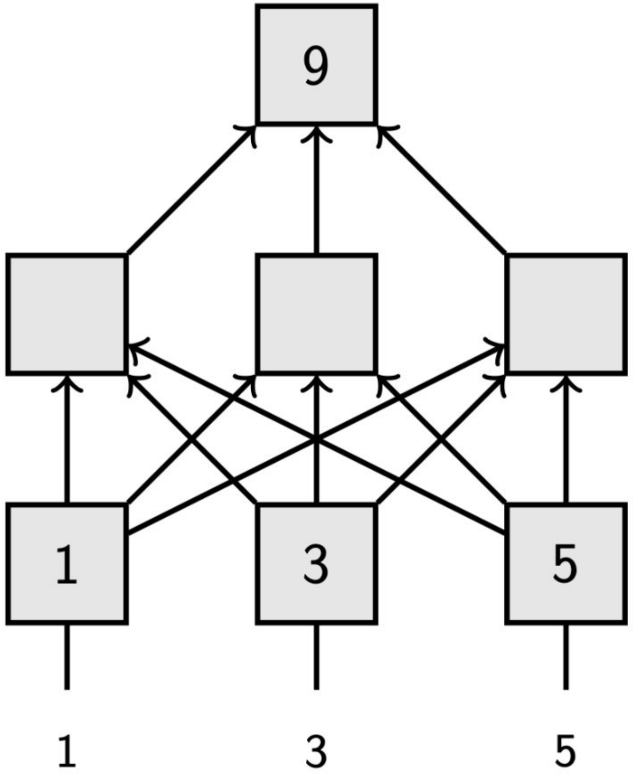
$$\text{SmallLinearModel}(h) = \textit{task}$$



Tenney et al. 2019

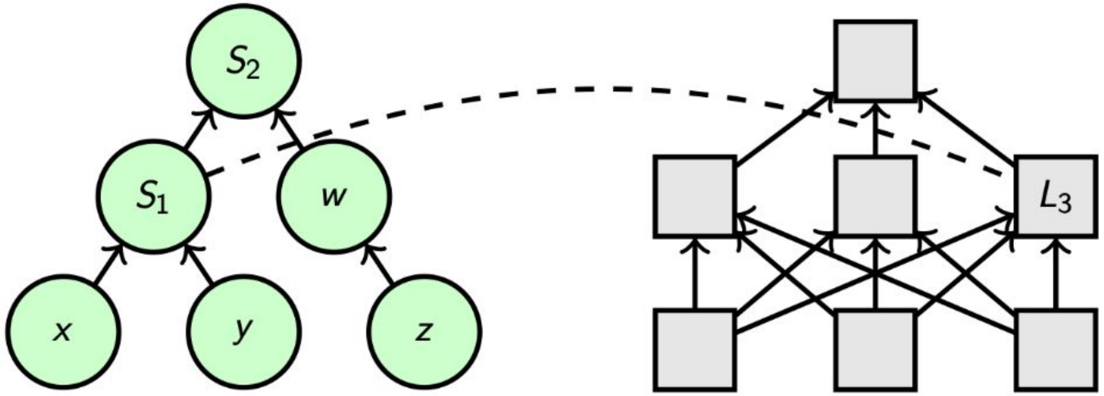
Internalism at work: Causal abstraction analysis



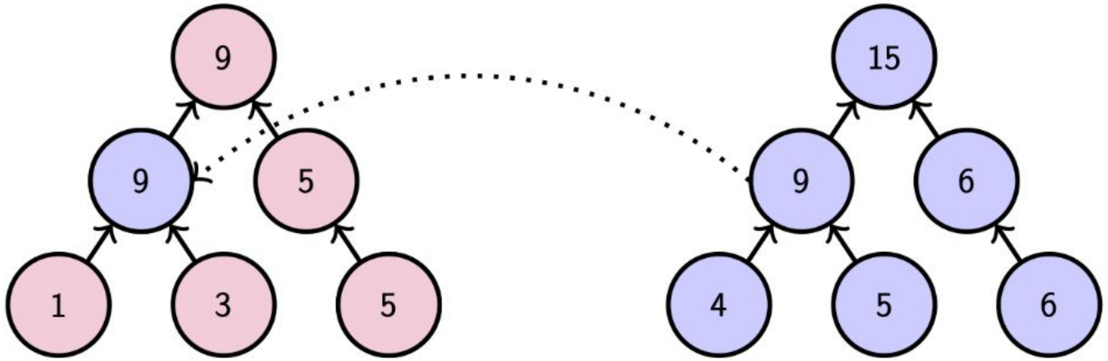


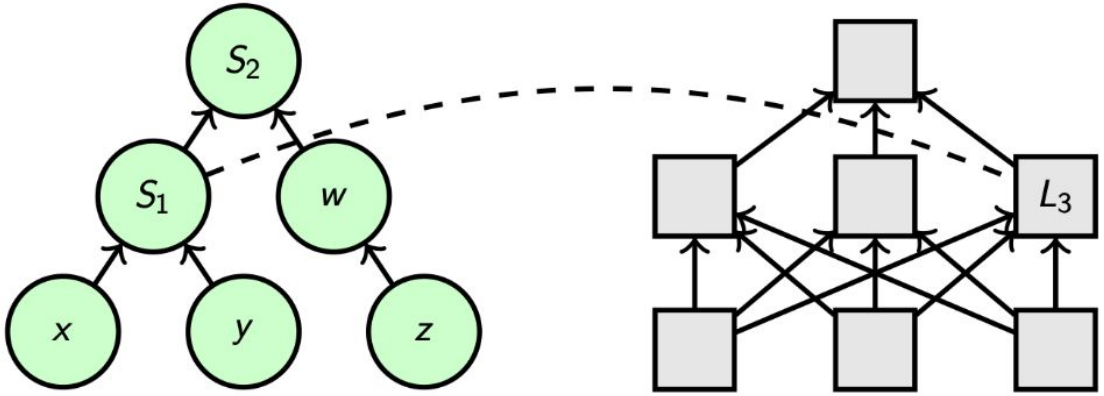
Beckers et al. 2020; Geiger, Lu et al. 2021; Geiger, Wu, Lu et al. 2022; Wu, Geiger et al. 2021

Internalism at work: Causal abstraction analysis

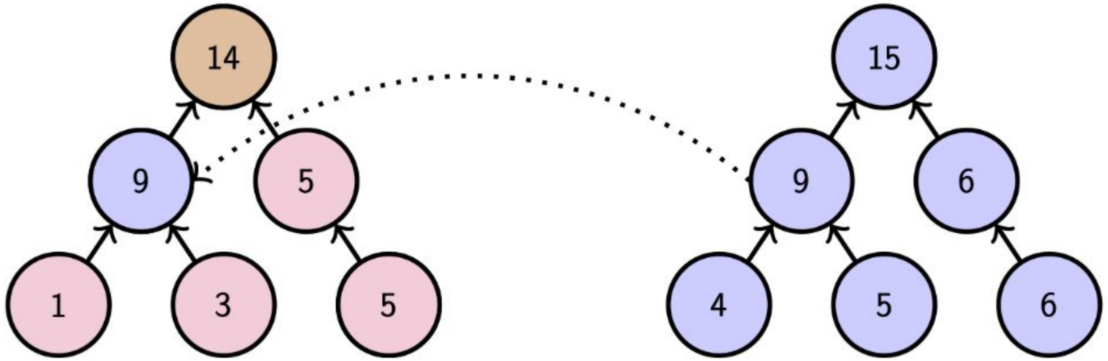


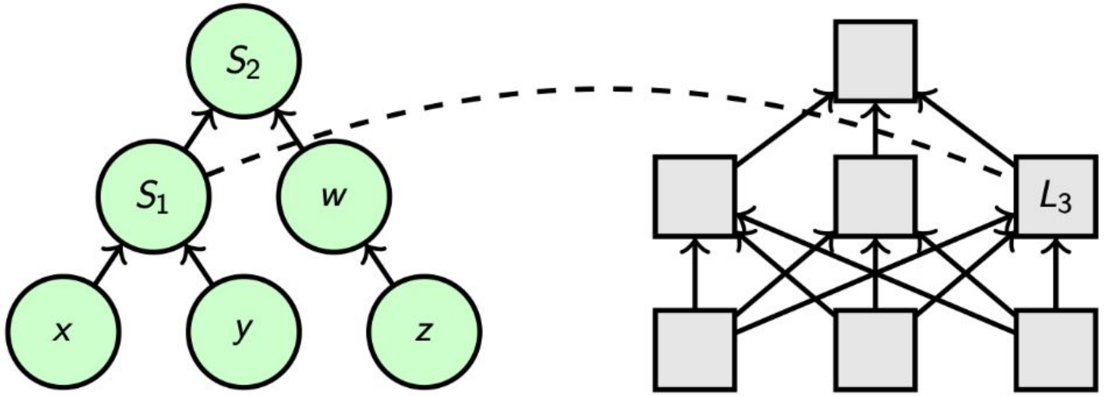
Internalism at work: Causal abstraction analysis



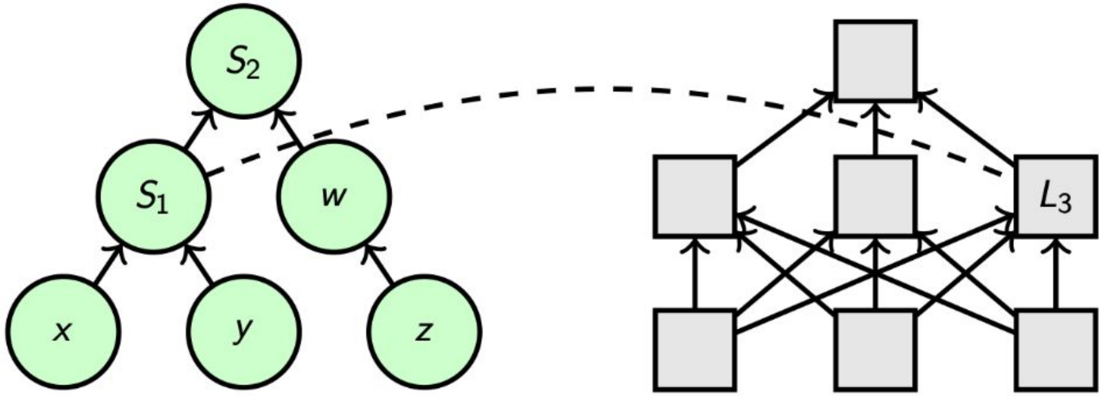


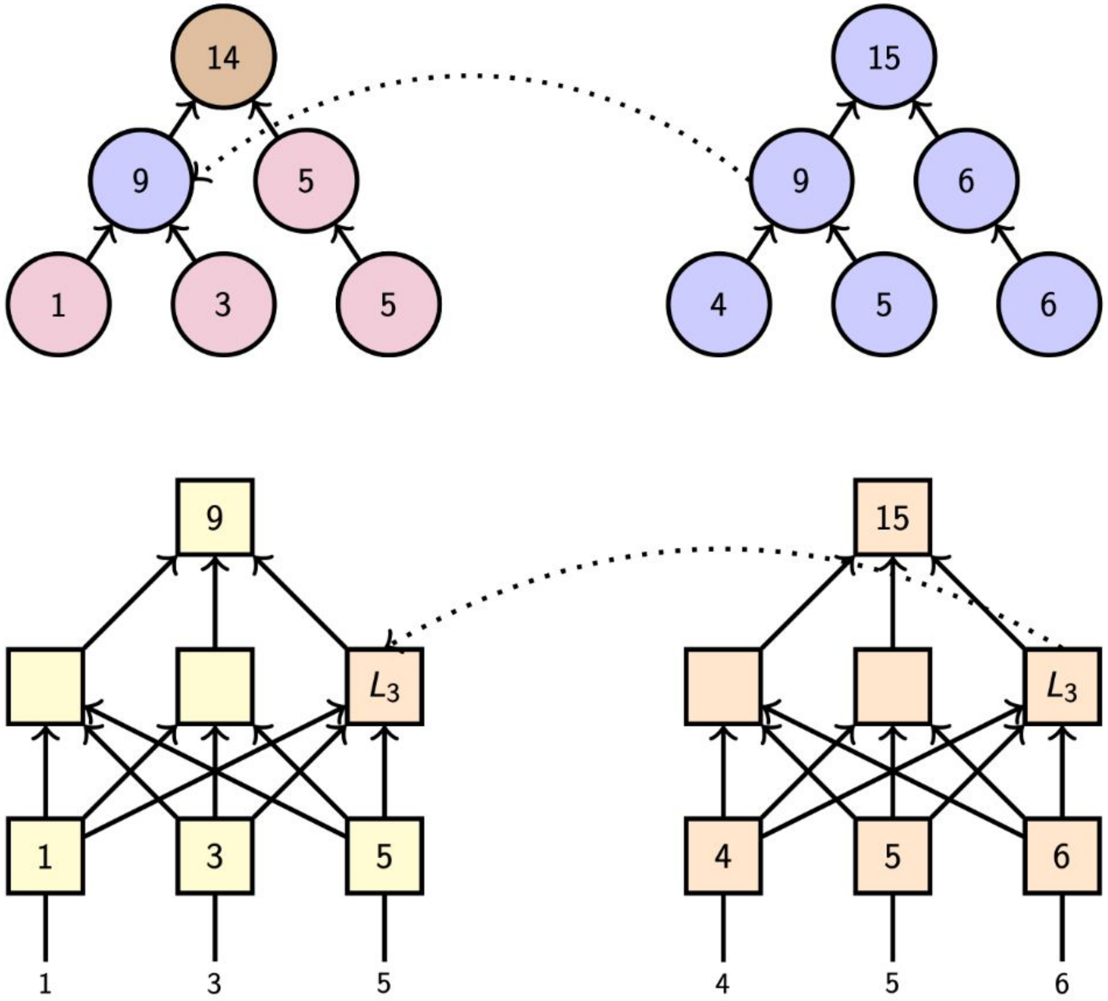
Internalism at work: Causal abstraction analysis



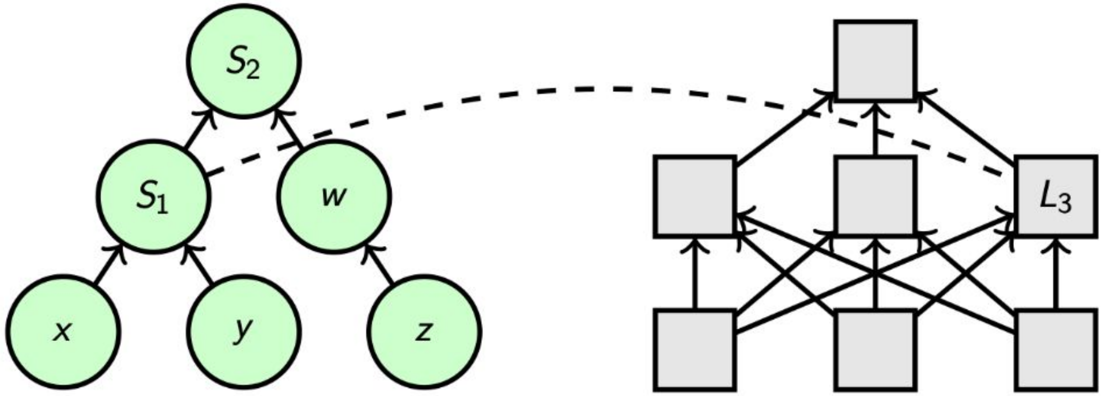


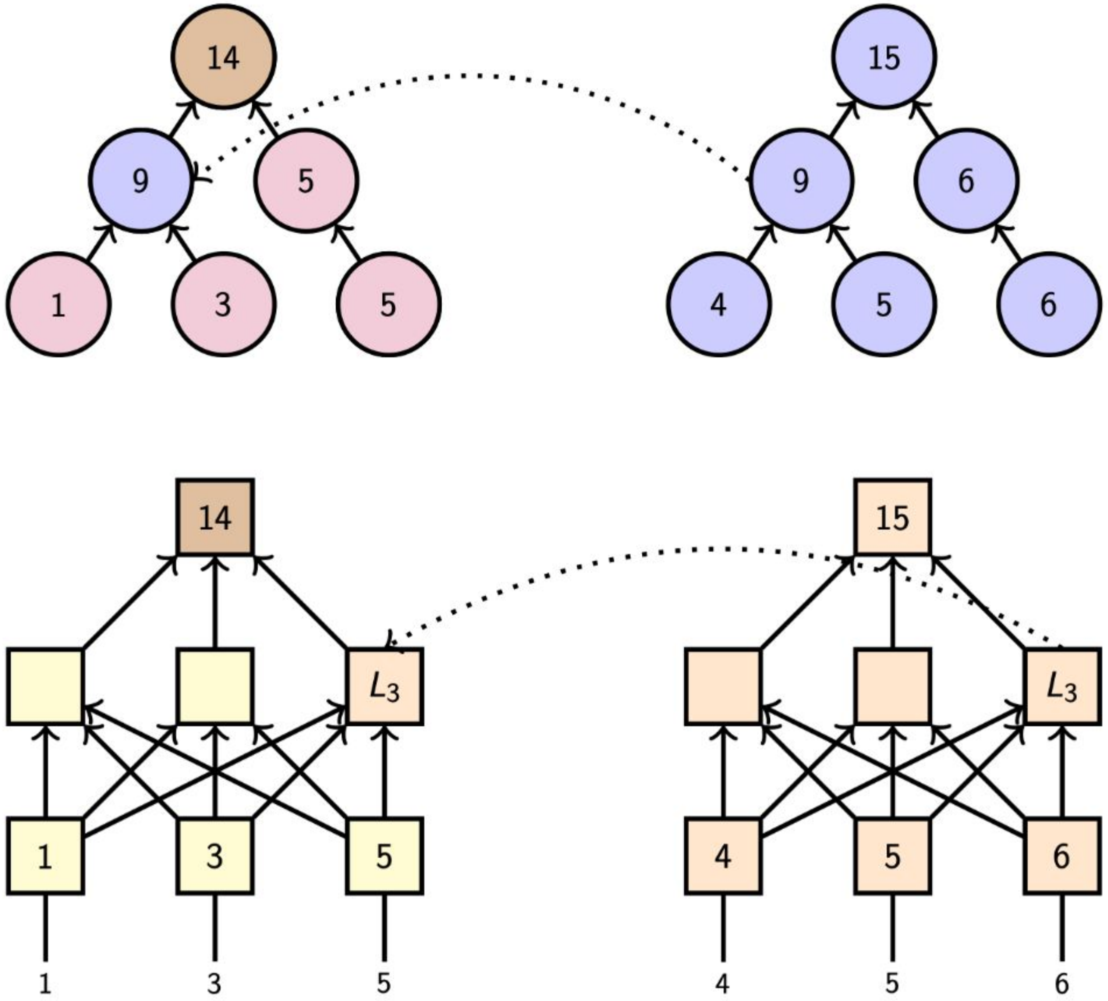
Internalism at work: Causal abstraction analysis



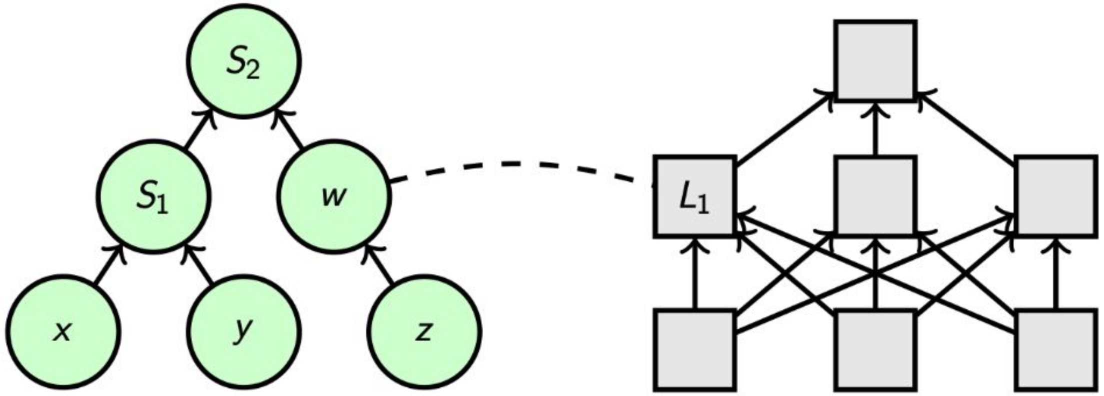


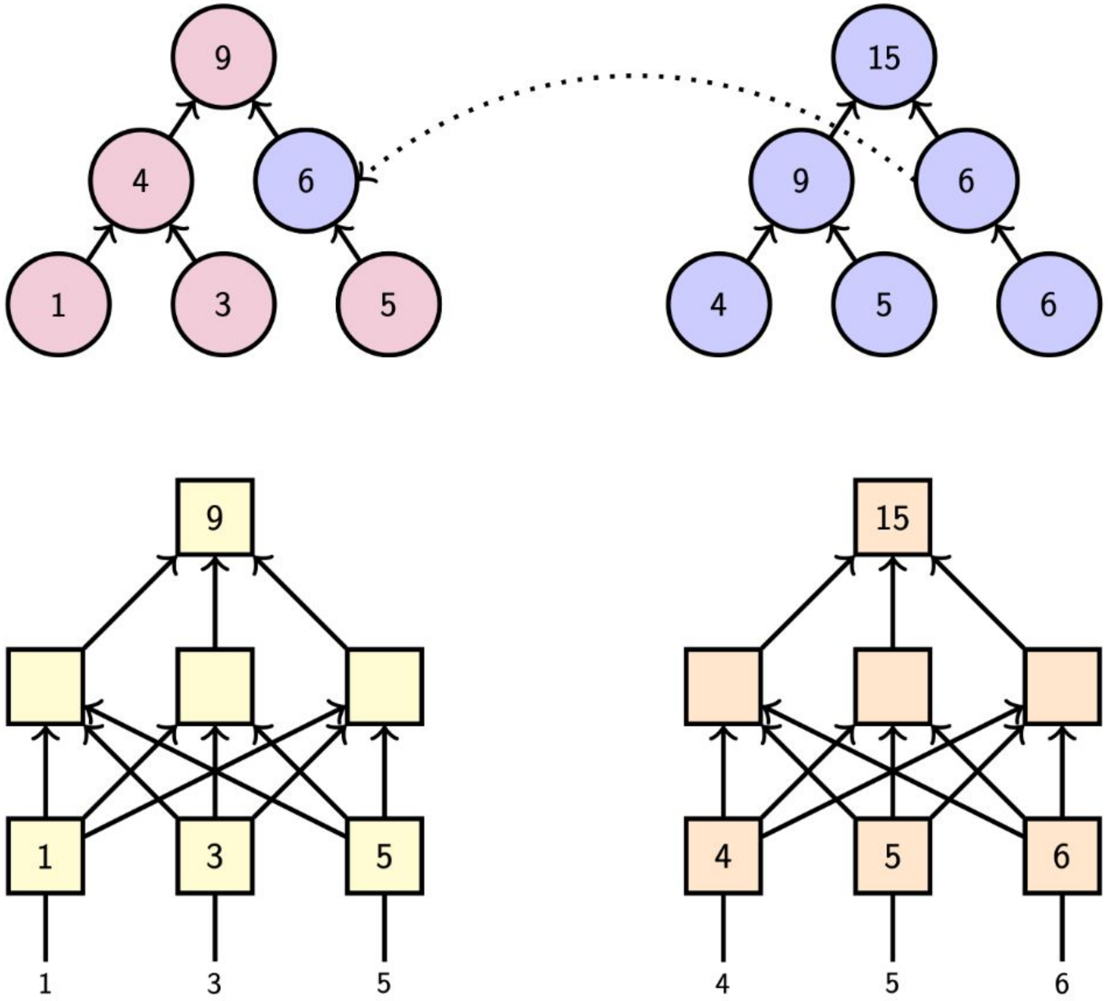
Internalism at work: Causal abstraction analysis



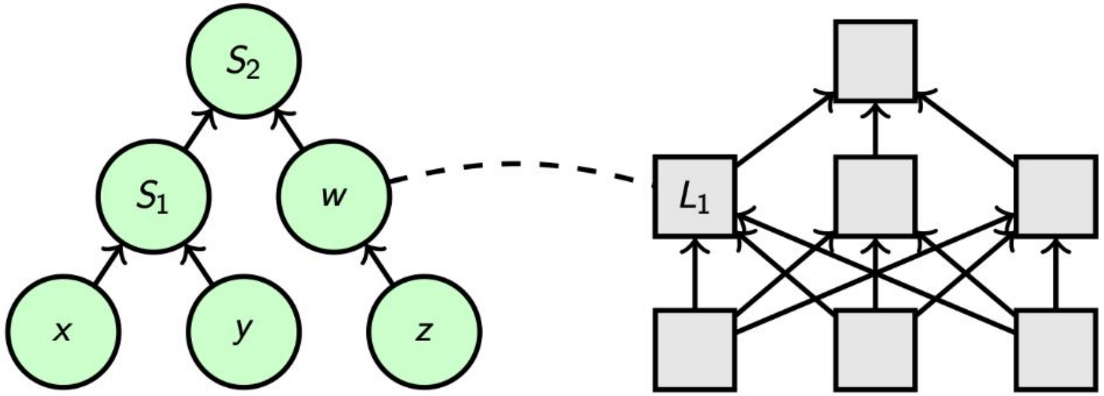


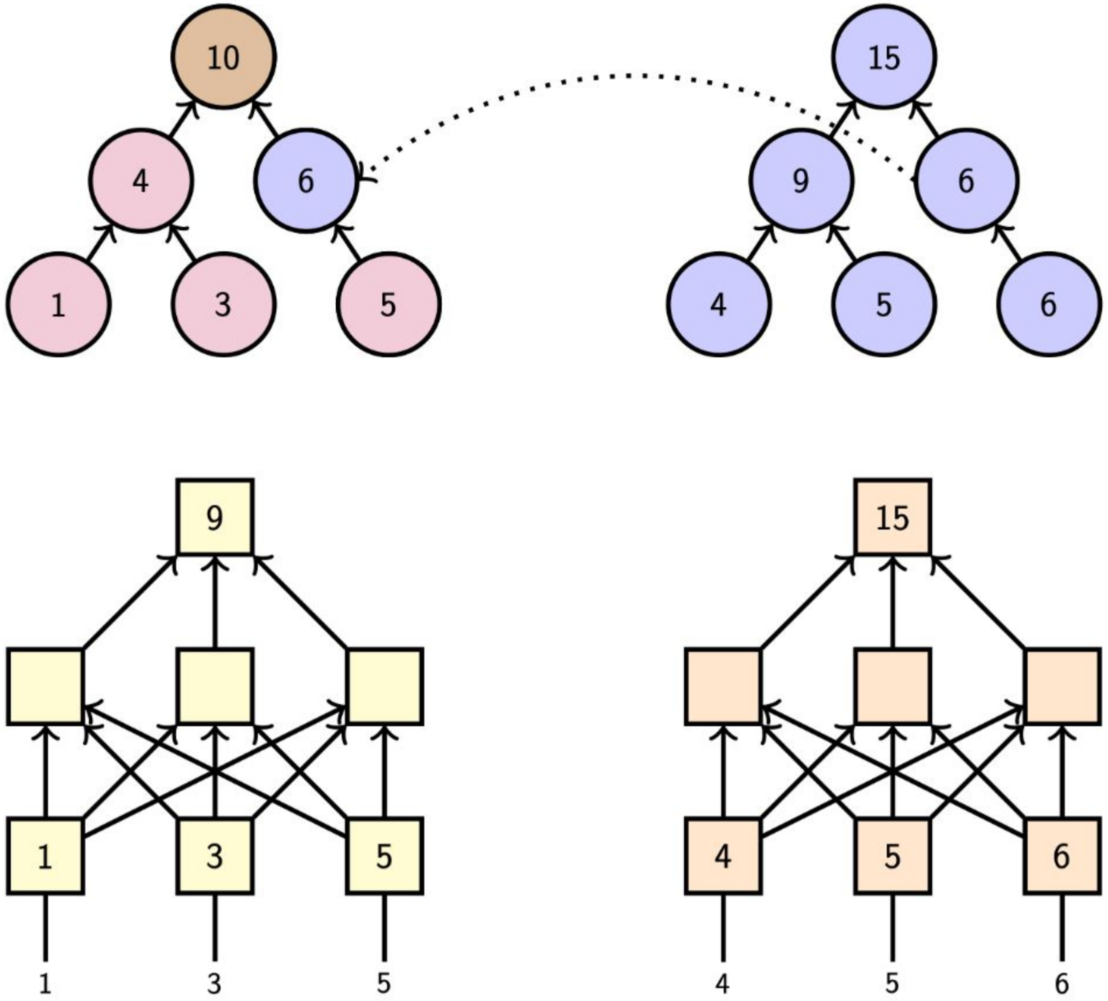
Internalism at work: Causal abstraction analysis



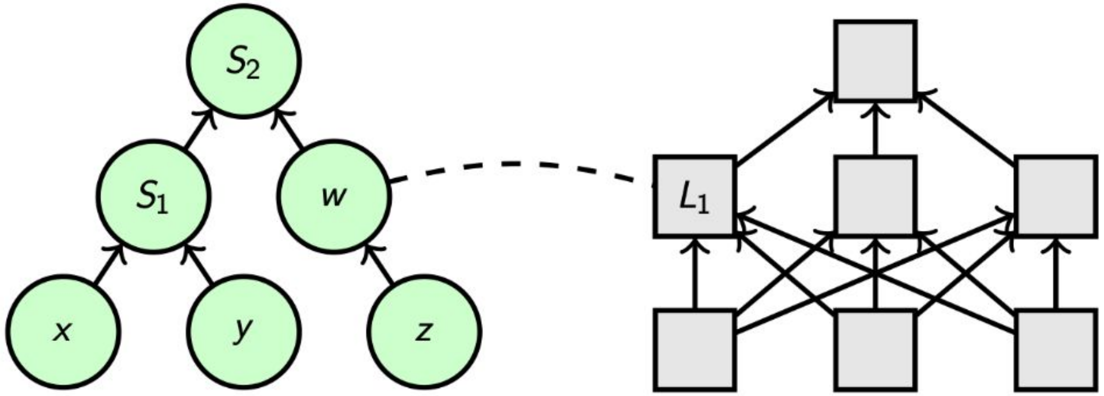


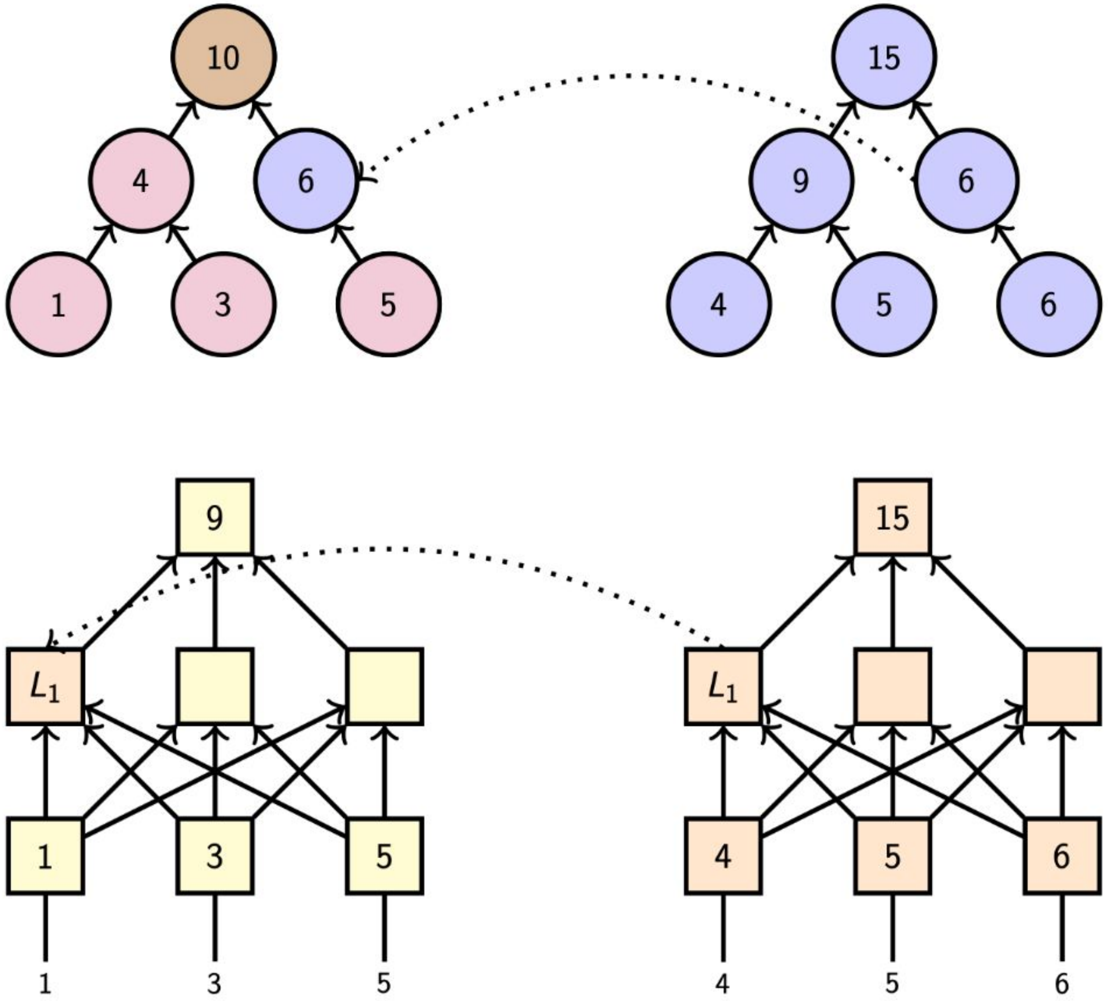
Internalism at work: Causal abstraction analysis



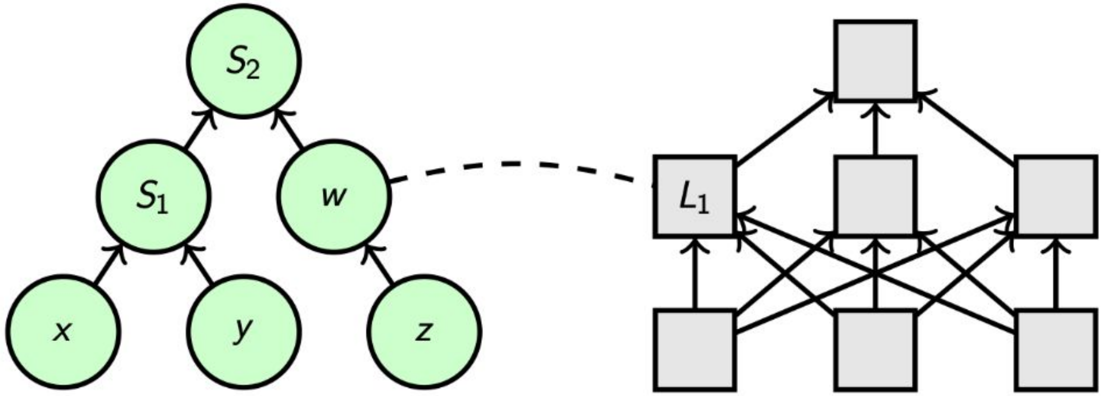


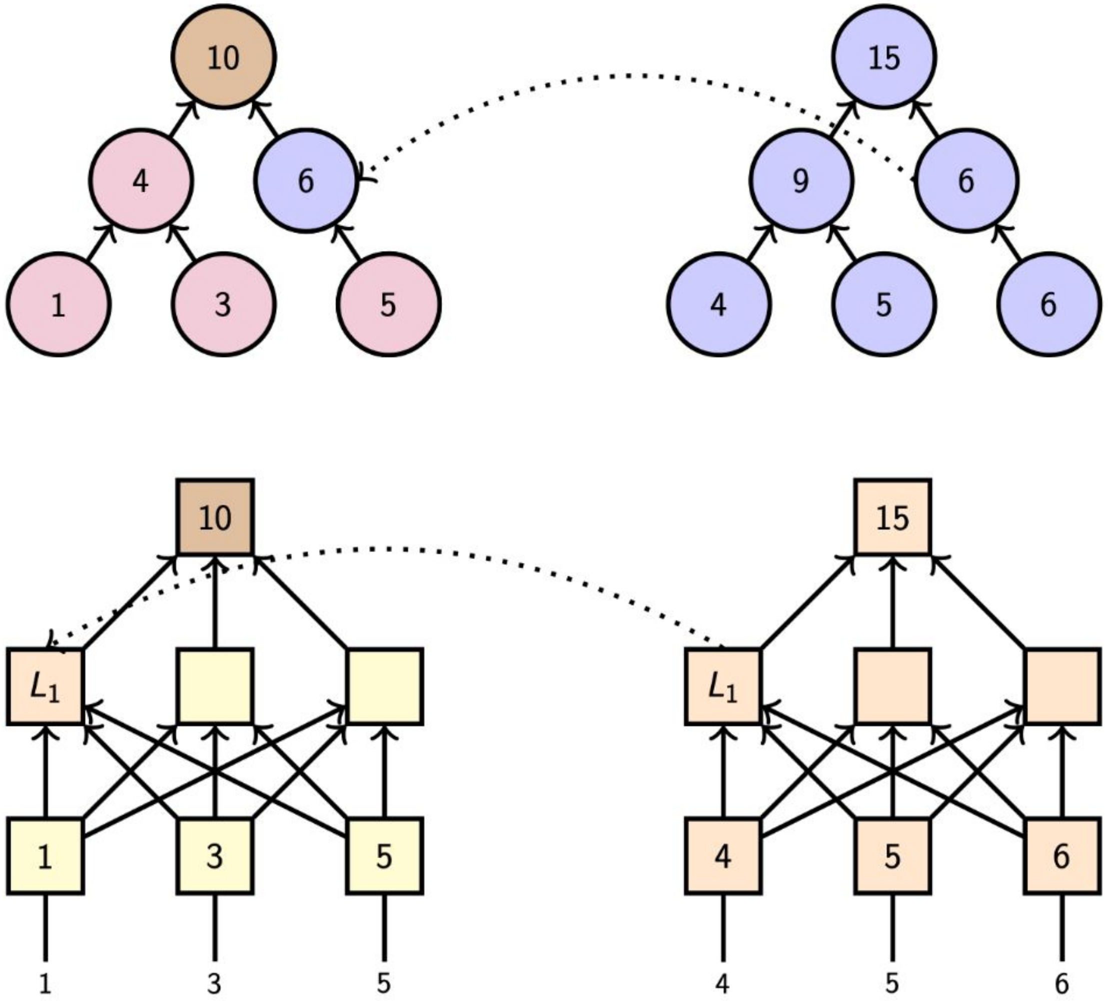
Internalism at work: Causal abstraction analysis



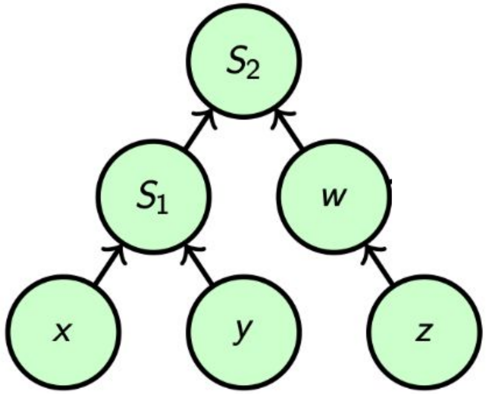


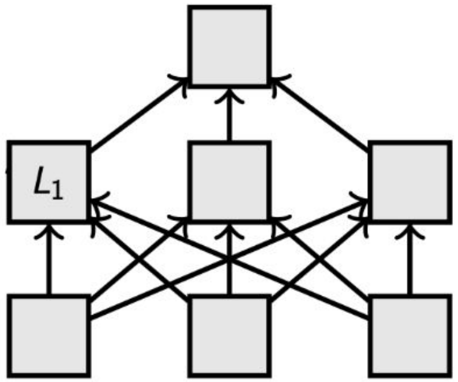
Internalism at work: Causal abstraction analysis

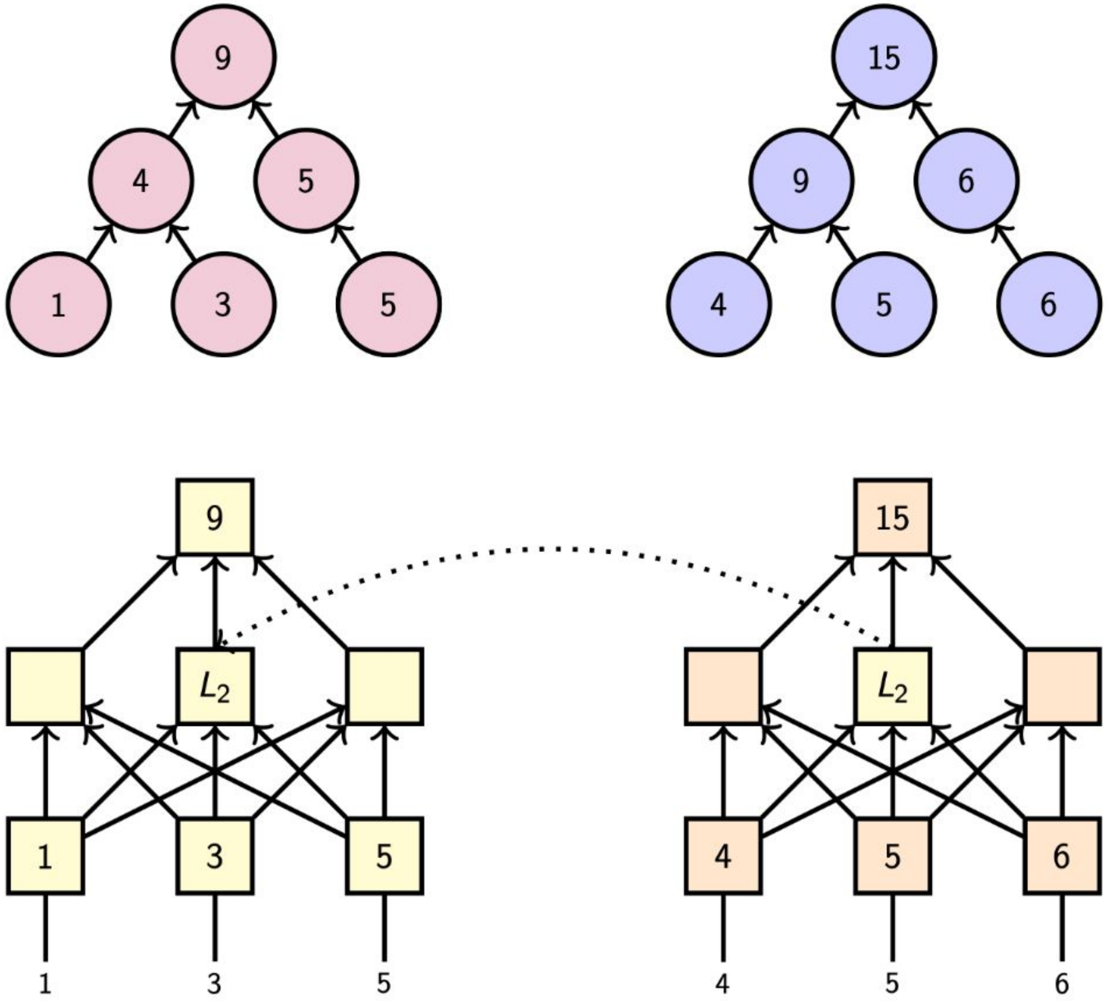




Internalism at work: Causal abstraction analysis







Findings of causal abstraction in large networks

- Fine-tuned BERT models succeed at hard, out-of-domain examples involving lexical entailment and negation *because* they are abstracted by simple monotonicity programs.
- Models succeed at the MNIST Pointer Value computer vision task *because* they are abstracted by simple programs like “if the digit is 6, then the label is in the lower left”.
- Models can be trained through interchange intervention training to better conform to high-level causal models/programs.

If a Foundation Model

- succeeds at hard language generalization tasks in a domain; and
- simulates a high-level causal model of that domain and the language used to describe it

then surely it has achieved grounded language understanding in that domain.

Could a purely self-supervised Foundation Model achieve grounded language understanding?

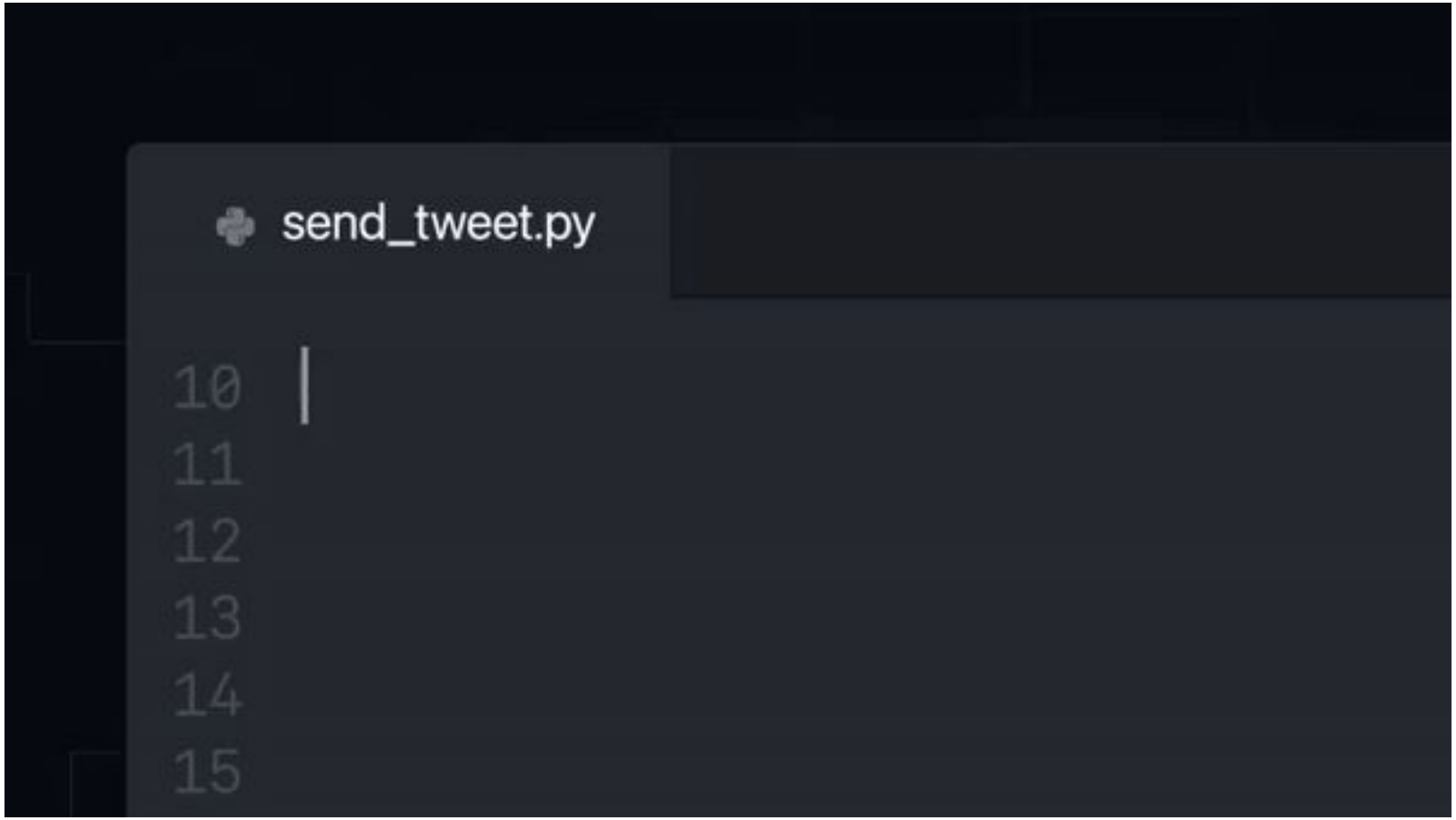
No ~~(obviously not)~~

Yes (I don't see why not)

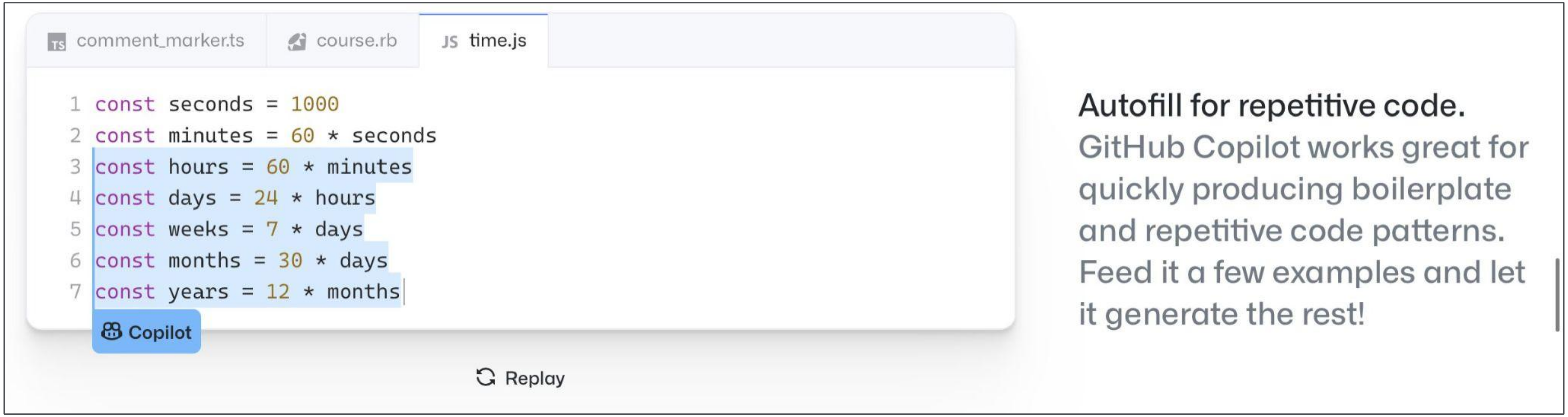
Thank you!

Appendix

Github Copilot (OpenAI Codex)



Github Copilot (OpenAI Codex)



```
const seconds = 1000
const minutes = 60 * seconds
const hours = 60 * minutes
const days = 24 * hours
const weeks = 7 * days
const months = 30 * days
const years = 12 * months
```

Autofill for repetitive code. GitHub Copilot works great for quickly producing boilerplate and repetitive code patterns. Feed it a few examples and let it generate the rest!

@yaakov_h

Pure self-supervision vs. regular supervision

Standard supervision for
nervous anticipation

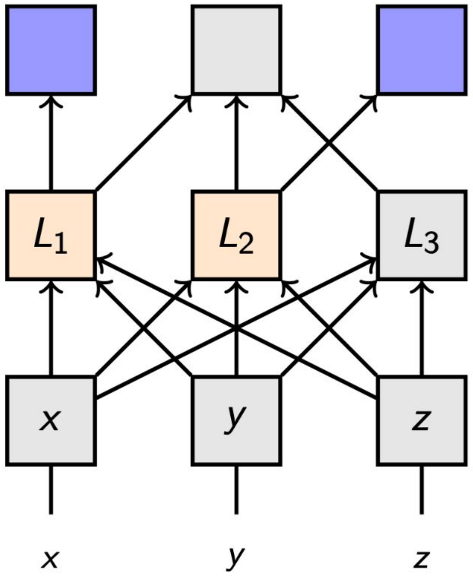
My palms started to sweat as the lotto numbers were read off.	**`nervous anticipation = 1`**
I took a deep breath as the curtain started to rise on my debut night.	**`nervous anticipation = 1`**
I couldn't shake a deep feeling of unease about the whole affair.	**`nervous anticipation = 0`**

Foundation Model
"Few-shot in-context learning"

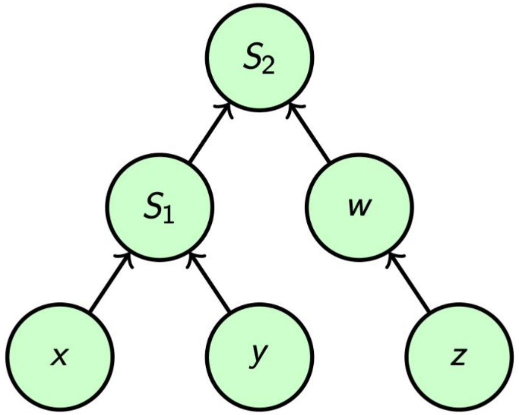
Hey model, here is an example of nervous anticipation: "My palms started to sweat as the lotto numbers were read off."

Hey model, here's an example without nervous anticipation: "..."

Probing does not support causal inferences



1. Probe L_1: it computes z
2. Probe L_2: it computes $x + y$
3. Aha!



4. But L_2 has no impact on the output!

$$W_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad W_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad W_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

$$\mathbf{w} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad (\mathbf{x}W_1; \mathbf{x}W_2; \mathbf{x}W_3)\,\mathbf{w}$$