Understanding Large Language Models

Carsten Eickhoff, Michael Franke and Polina Tsvilodub

Session 09: Understanding LMs

```
mechanistic explanation that specifically probe for which
types of representations are causally efficacious in
generating the type of behavious.
```

Understanding understanding

1. Do LLMs understand language?
   Depends on what it means to understand language.

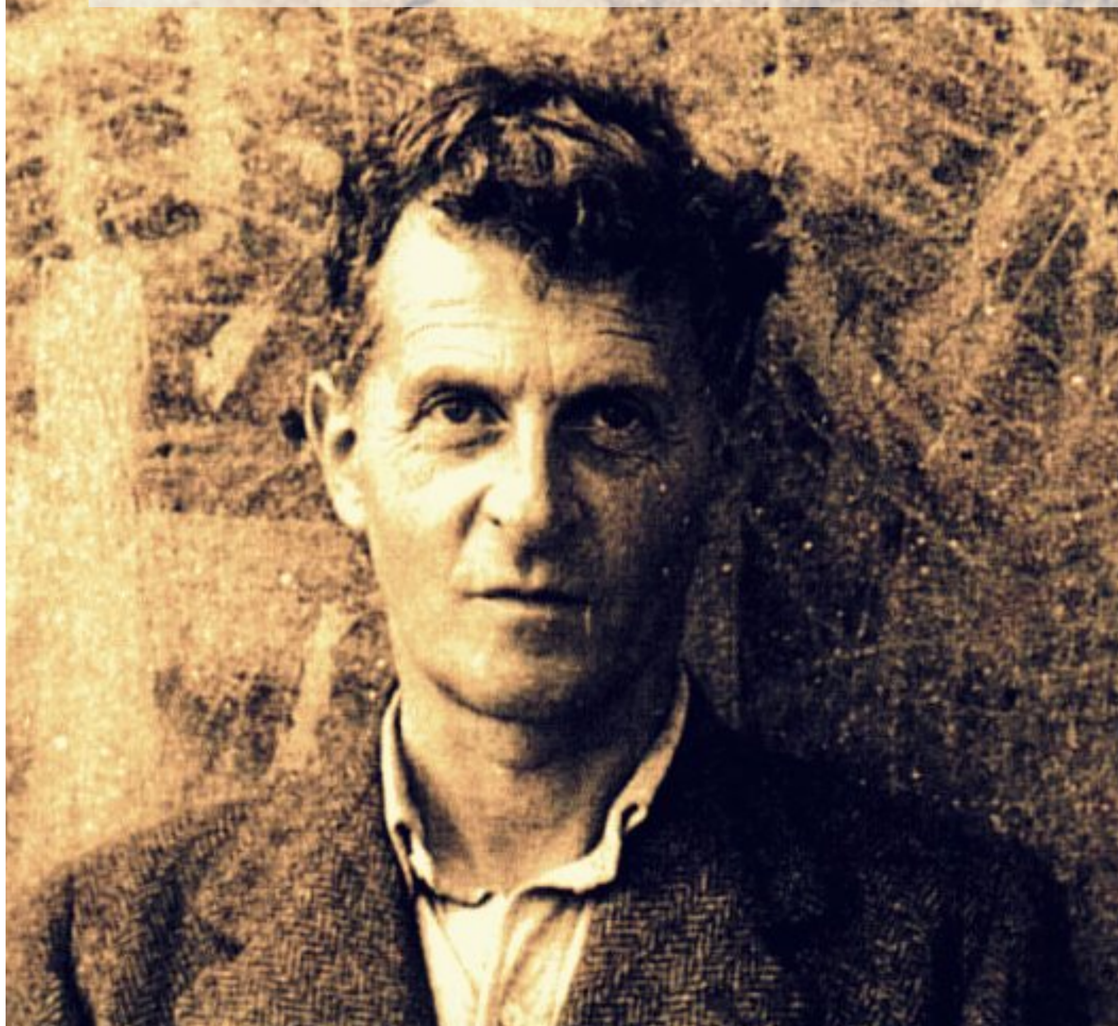
2. Do LLMs understand the world?
   Depends on what it means to understand the world.

3. How can we understand how LLMs work?
   Requires familiarity w/ LLMs and w/ facility of human interpretation.

4. Do LLMs help us understand language or mind?
   Depends on what we consider a useful explanation in science.

Wenn ein Löwe sprechen könnte, wir könnten ihn nicht verstehen.

meet the lion here



Behavioral ecology of intelligent artifacts

Understanding AI

think AI ∽ "alien intelligence"

"[M]achines exhibit behaviours that are **fundamentally different from animals and humans**, so we must **avoid excessive anthropomorphism and zoomorphism**. Even if borrowing existing behavioural scientific methods can prove useful for the study of machines, machines may exhibit **forms of intelligence and behaviour that are qualitatively different —even alien — from those seen in biological agents**."

Machine behaviour

Iyad Rahwan[1,2,3,34]*, Manuel Cebrian[1,34], Nick Obradovich[1,34], Josh Bongard[4], Jean-François Bonnefon[5], Cynthia Breazeal[1], Jacob W. Crandall[6], Nicholas A. Christakis[7,8,9,10], Iain D. Couzin[11,12,13], Matthew O. Jackson[14,15,16], Nicholas R. Jennings[17,18], Ece Kamar[19], Isabel M. Kloumann[20], Hugo Larochelle[21], David Lazer[22,23,24], Richard McElreath[25,26], Alan Mislove[27], David C. Parkes[28,29], Alex 'Sandy' Pentland[1], Margaret E. Roberts[30], Azim Shariff[31], Joshua B. Tenenbaum[32] & Michael Wellman[33]

Tinbergen's dimensions of analysis

for studying animal behavior

1. mechanism
   - most proximate causes of behavior

2. development
   - more distal causes of the mechanisms in the coming-to-existence of the agent

3. function
   - the functional rationale for the behavior: *cui bono*?; why maintained here & now?

4. evolution
   - ulterior causes from selective pressure in past environments

https://www.wikidata.org/wiki/Q7155#/media/File:Ethology_diversity_2.jpg

Think Break
for studying ML / AI / LLMs

How do Tinbergen's four dimensions apply to LLMs?

1. mechanism
   - most proximate causes of behavior
2. development
   - more distal causes of the mechanisms in the coming-to-existence of the agent
3. function
   - the functional rationale for the behavior: *cui bono*?; why maintained here & now?
4. evolution
   - ulterior causes from selective pressure in past environments

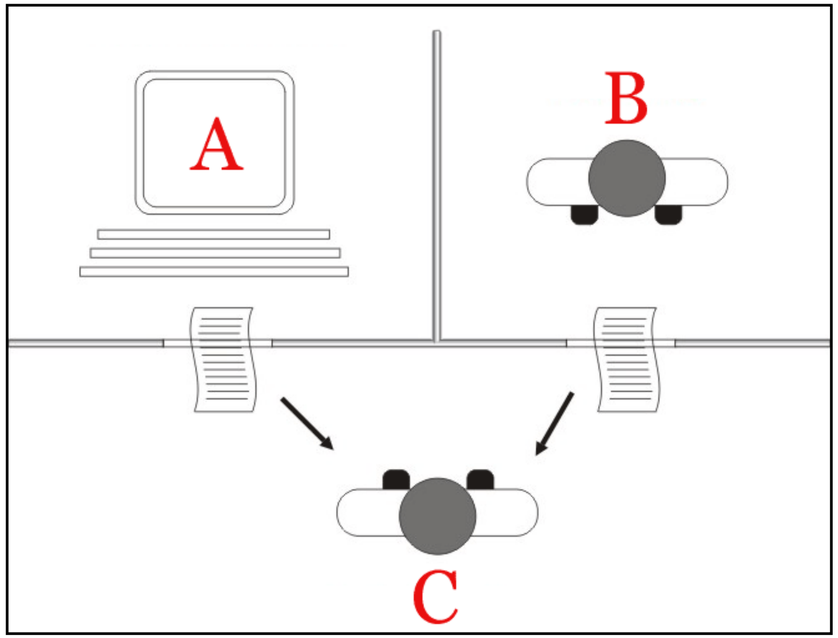
On understanding language

Popular wisdom

(1) a. Humans understand language. *true*

b. Language models understand language. *false*

Turing test



(Saygin et al. 2000)

Turing test

- different variations:
  - number of agents
    - 2 agents: human guesses whether given input is human or machine
    - 3 agents: human guesses which input is from human or machine
  - lay vs expert
    - human agent is lay person or expert
  - duration of test
    - short version: ca. 5 minutes dialogue
    - extended version: several hours of dialogue
- recent LM performance
  - SOTA LMs pass short, lay person, 2-agent version (Jones & Bergen 2024)
  - actual TT is long, expert, 3-agent version
- potential problems from Goodhart's law

Think Break

for studying ML / AI / LLMs

What would you (as an expert) ask to unmask an LM agent in a TT?

(for current SOTA models)

The “something is missing” intuition

(1) a. Humans understand language. *true*

b. Language models understand language. *false*

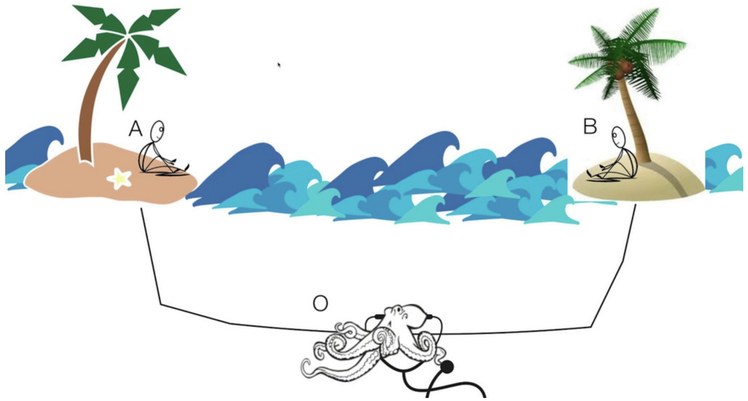
- LMs are missing …
  - **grounding:** e.g., in multi-modal training (Bender & Koller, 2020)
  - **embodiment:** a physical body or form
  - **situatedness:** ability to interact with …
    - a static or dynamic environment
    - other agents (Schlangen, 2023)

"How do things change if I go into nonsense land? Nonsense land is a hard test for many models, because they don't know that the world exists. They have no symbolic understanding of the world. They are just machines that generate next tokens." ❤️

CE, last lecture

Octopus Argument

- **objection:** too strong language in reporting results
- claim:
  - understanding requires mastering meaning
  - impossible from text-data alone
- octopus argument:
  - a very clever octopus listens in on the conversation of two stranded islanders
  - learns to mimic their behavior
  - but will miss to generalize for OOD input



> (1) In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task. (Devlin et al., 2019)
>
> (2) Using BERT, a pretraining language model, has been successful for single-turn machine **comprehension** . . . (Ohsugi et al., 2019)
>
> (3) The surprisingly strong ability of these models to **recall factual knowledge** without any fine-tuning demon-

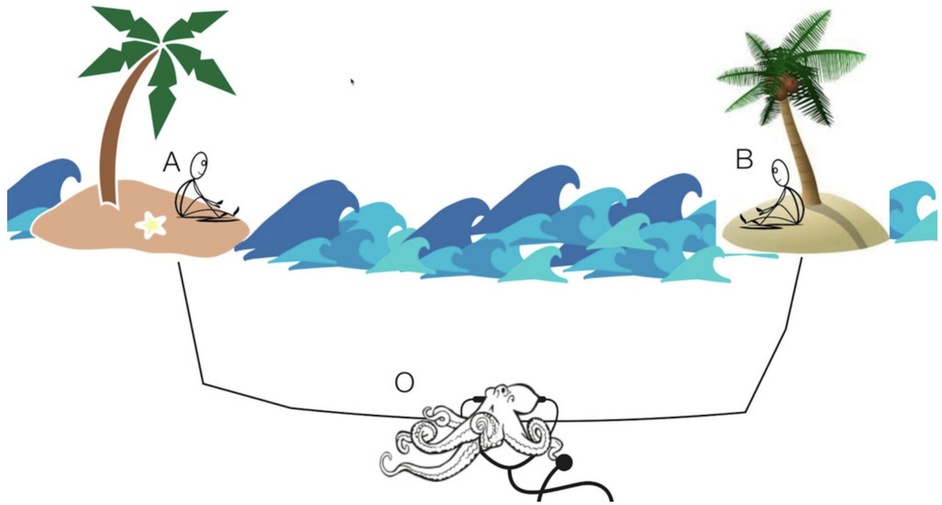
> We argue that *the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning*. We take the term *language model* to refer to any system trained only on the task of string prediction, whether it operates over characters, words or sentences, and sequentially or not. We take (linguistic) *meaning* to be the relation between a linguistic form and communicative intent.

> Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O has no idea what A "means". Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test, if A hadn't been eaten by the bear before noticing the deception.[7]

Bender & Koller (2020)

Octopus Argument

- empirical criticism:
  - concrete examples from the paper all treated well by current SOTA text-only models
- conceptual criticism:
  - thought experiment **stacked in favor of desired conclusion**
  - **no in principle argument against meaning emergence** from large-scale statistical learning from distributional properties
  - full agreement with all premises (?) and the conclusions, but **disagreement with presumption of validity**



Bender & Koller (2020)

Approach from analytical philosophy

(2) a. Humans understand$^{S}_{H}$ language.

b. Language models understand$^{S}_{LM}$ language.

Is there an adequate sense S of "A understands X" such that:

1. both (2a) and (2b) are true for sense S, and if so **yes**
2. when applied to humans and LMs, are the specific predicates "understand$^{S}_{H}$" and "understand$^{S}_{LM}$" the same? **likely not**

Flavors of understanding

- phenomenal understanding:
  - A understands^{P} X if A has a subjective feeling of ‘making sense’ or ‘not being puzzled’
- explanatory understanding:
  - A understands^{E} X if A knows an adequate explanation of X
- use understanding:
  - A understands^{U} X if A can use X (or information about it) to achieve their goals
    - **verbal understanding:** having linguistic capability to answer questions about X
    - **behavioral understanding:** being able to manipulate X
    - **reason understanding:** being able to reason and make inferences about X

Do LLMs Understand?

- **LLMs have many aspects of explanatory understanding and use understanding, at least if these are construed in behavioral terms.**

Chalmers (2024)

de Regt (2017), Dellsén (2020), Grimm (2021)

Flavors of phenomenal understanding

or, could LMs be conscious?

- “something is missing” intuitions for LM phenomenological understanding
  - biology
  - senses and embodiment
  - adequate world & self model
  - recurrent processing
  - global workspace
  - unified agency

Chalmers (2023)

Flavors of phenomenal understanding

or, could LMs be conscious?

- "something is missing" intuitions for LM phenomenological understanding
  - biology
  - senses and embodiment
  - **adequate world & self model**
  - recurrent processing
  - global workspace
  - unified agency

Chalmers (2023)

As Ilya Sutskever compactly put it, to learn to predict text, is to learn to predict the causal processes of which the text is a shadow.

Eliezer Yudkowsky "GPTs are Predictors, not Imitators", April 8th 2023 on lesswrong.com

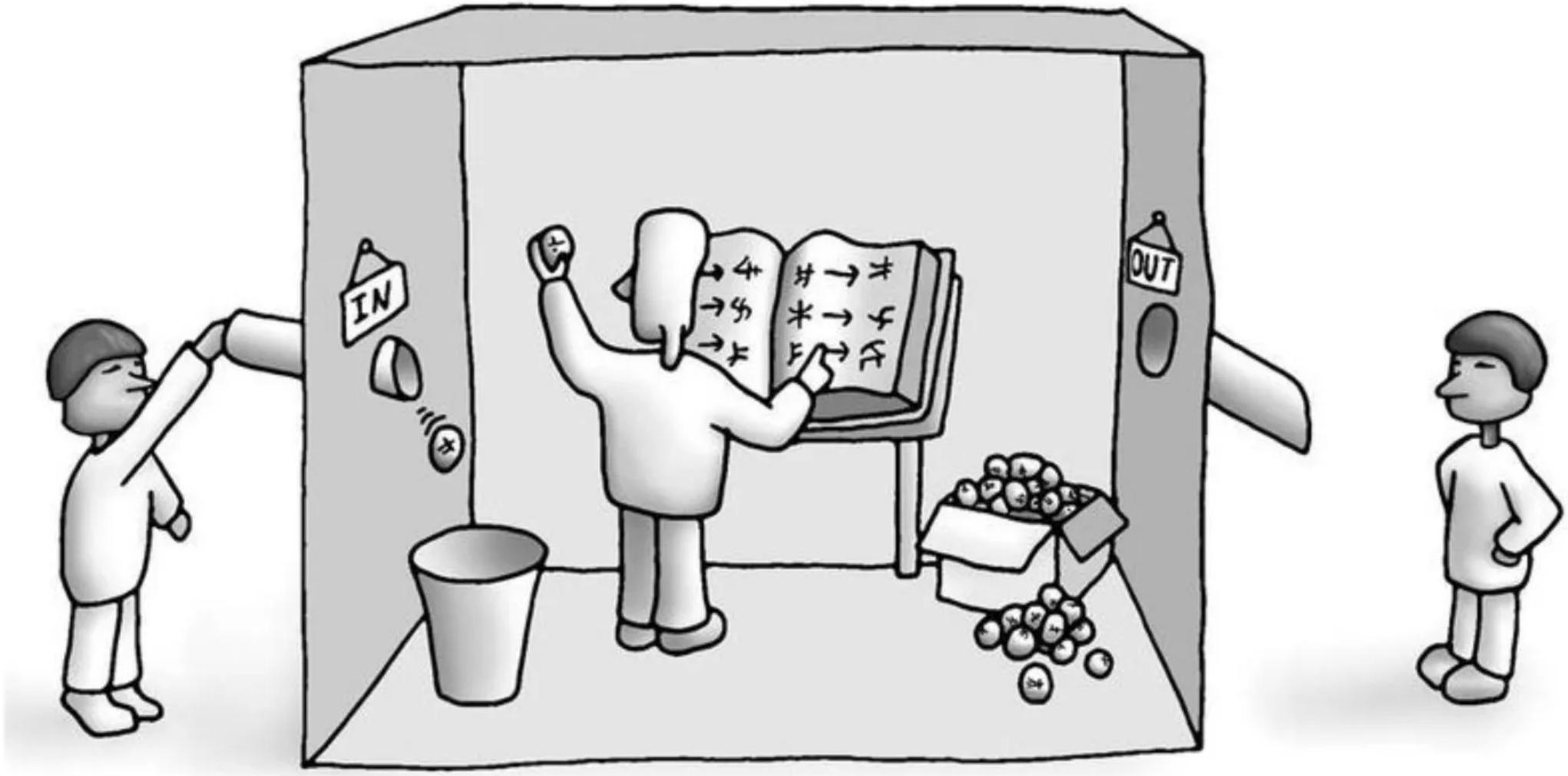
My contribution

1. Japanese Room Argument
   - intuitive argument against “subjective experience” in LM understanding

2. Minimal Models Argument
   - formal argument against “world-model” sense of LM understanding

The Japanese Room Argument

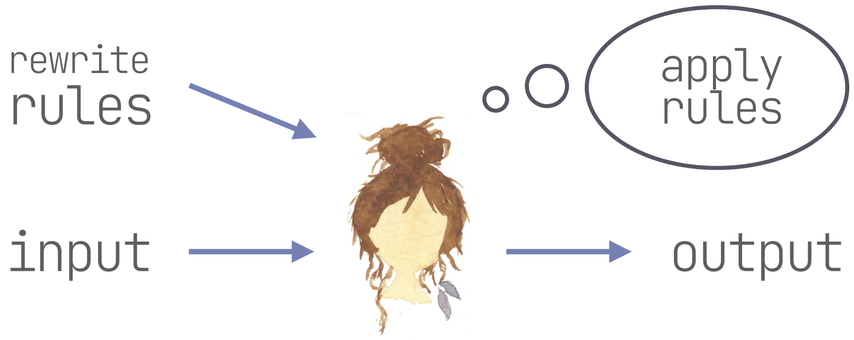
Chinese Room Argument



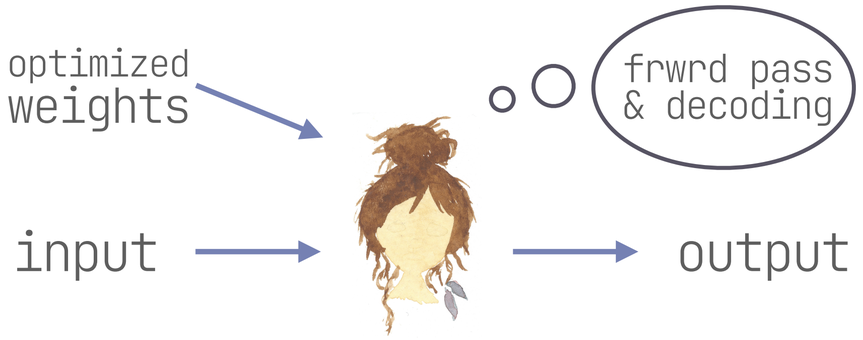
symbol manipulation

statistical learning

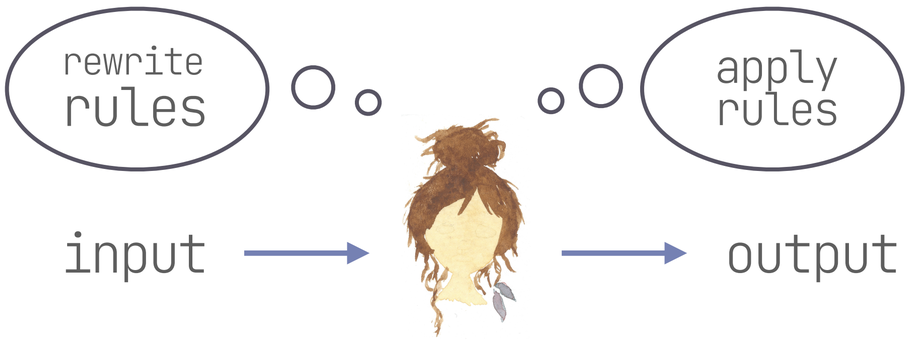
narrow experiencer

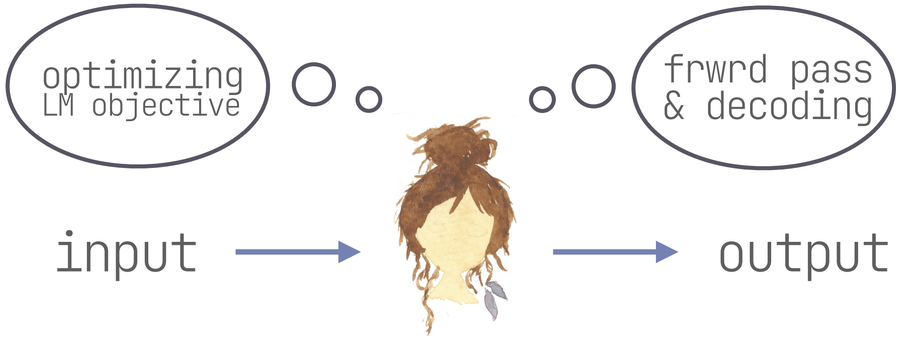


CRA



wide experiencer





JRA

Main question

Would a human experiencer **necessarily** develop understanding^{S} of language if they undergo **the same learning and generation protocols as LMs** with training **data of perfect quality and unlimited quantity** consisting exclusively of natural language text?

focus on phenomenological understanding

Main question

Extraordinary claims require extraordinary evidence

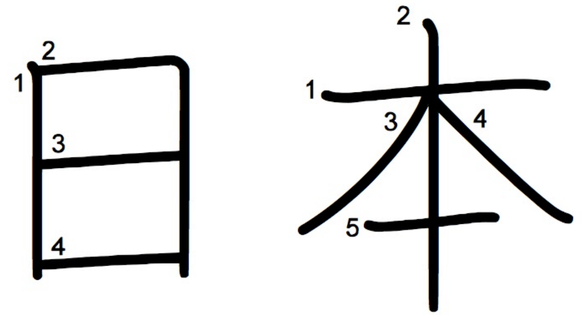
Would a human experiencer **necessarily** develop understanding^{S} of language if they undergo **the same learning and generation protocols as LMs** with training **data of perfect quality and unlimited quantity** consisting exclusively of natural language text?

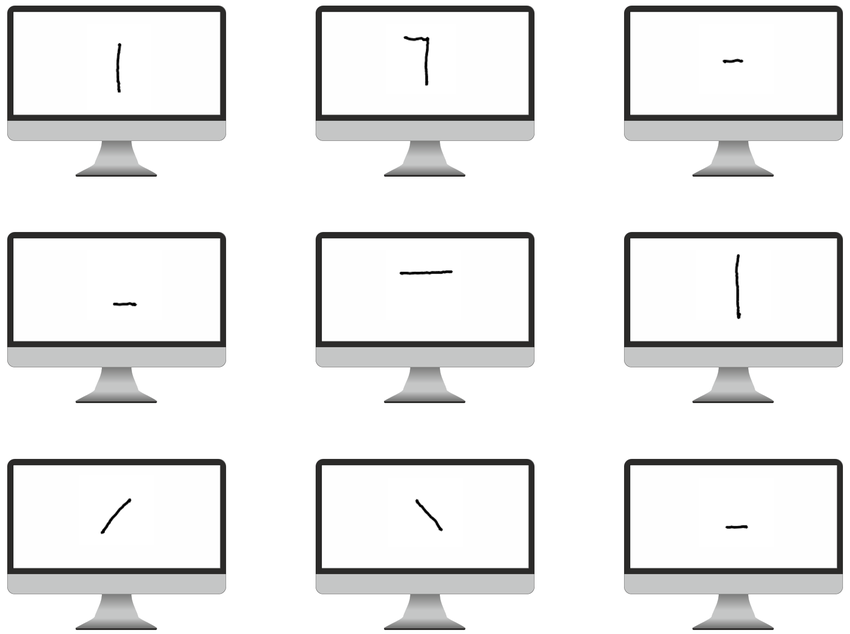
focus on phenomenological understanding

Japanese Room Argument

version 1: sequential & predictive (incorrect)

- you do not know Japanese
- you are strapped in front of a video screen
- Japanese **text is presented stroke-by-stroke**
  - individual tokens may be meaningless (!)
- after each stroke, **you predict the next stroke**
- feedback on whether your prediction was correct
- over time you get better at this
  - Fermi calculation:
    - LLAMA one pre-trained on ~1.4e12 tokens,
    - make that 1.2e12
    - JRA agent takes 10 seconds per trial
    - ~3e7 seconds per year
    - **~400k years of training** (Homo heidelbergensis, early Neanderthals)

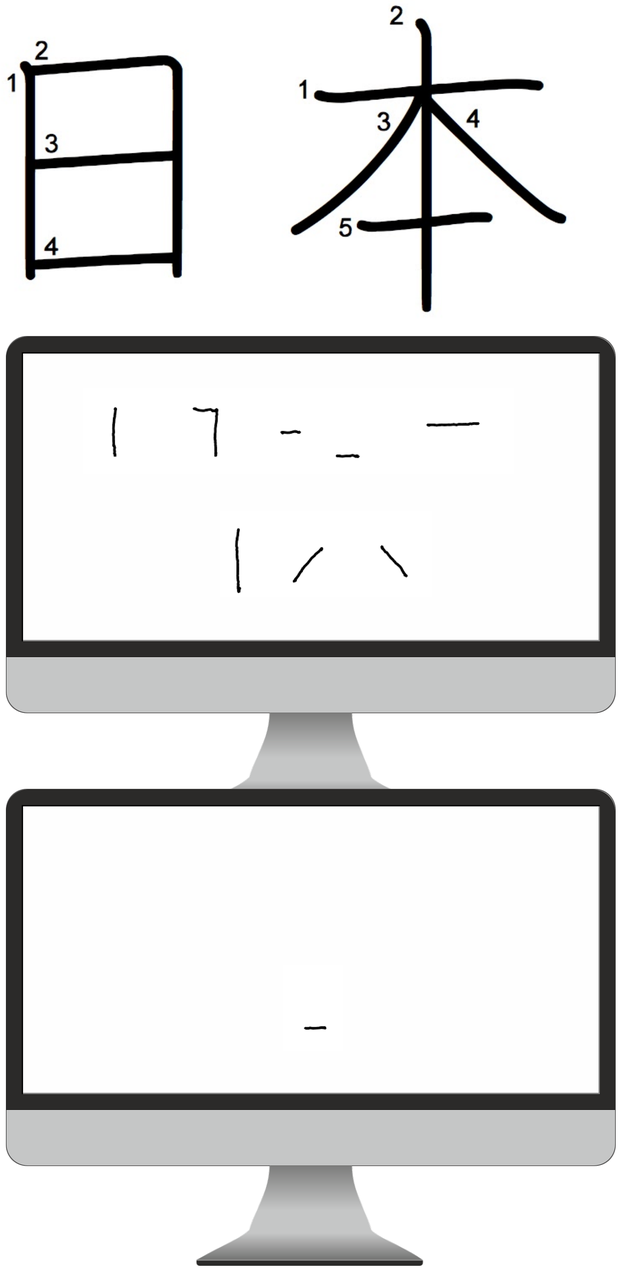




Japanese Room Argument

version 2: non-sequential & anticipatory (better)

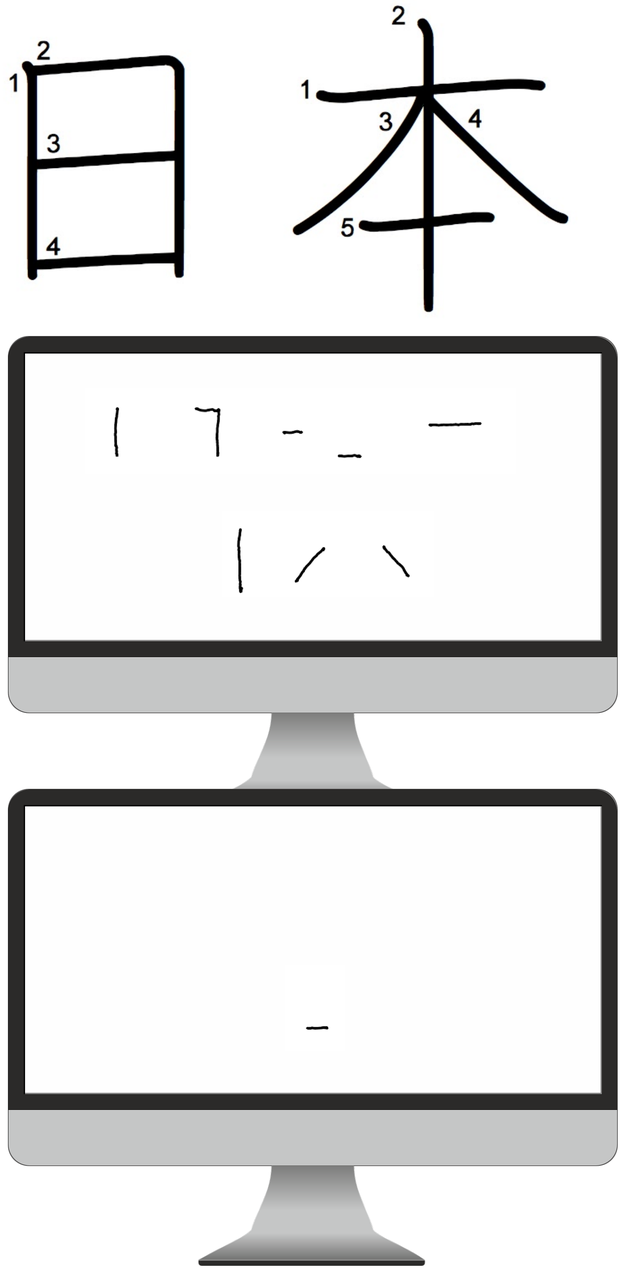
- you do not know Japanese
- you are strapped in front of a video screen
- Japanese **text is presented as a large grid of strokes**
- after the grid disappears **a single stroke appears**
- a machine measures how surprised you are
- you get an electric shock inversely proportional to your **ability to anticipate the stroke**
- over time you get better at this



Japanese Room Argument

key properties

- memoryless
- non-sequential
- non-predictive
  - anticipatory
  - non-agentive
- non-reflective / non-agentive
  - habitual reaction, possibly below awareness
  - System 1, not System 2
- non-normative



Counterargument

control or consciousness are necessary

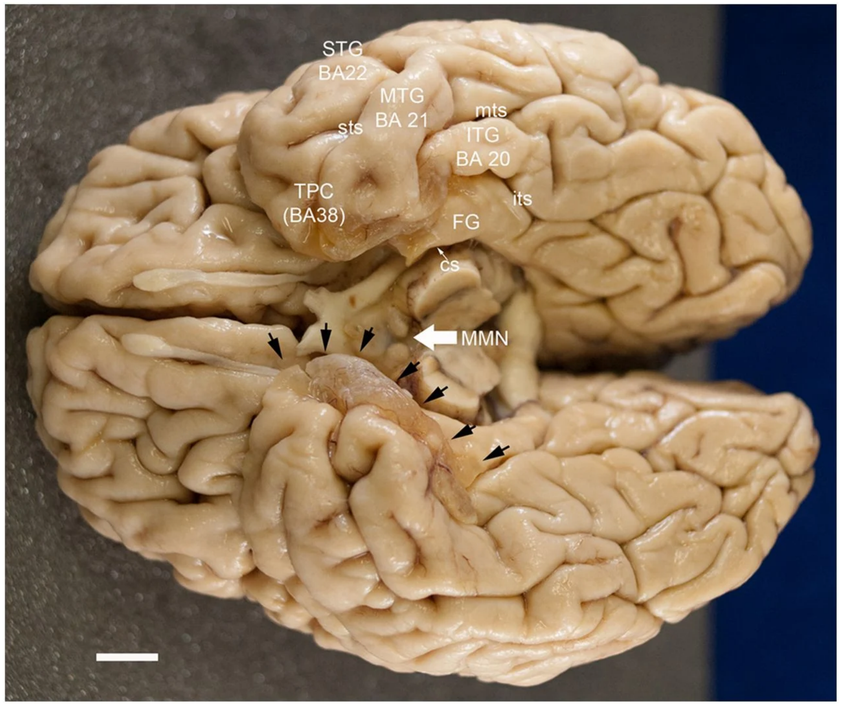
"I cannot imagine that the task can be learned without active prediction and cognitive control, leading to awareness and phenomenological understanding."

"Since we (normally) have awareness of learning, this is indeed hard to imagine, but failure to imagine is not a strong argument."

Patient EP

"Robust habit learning in the absence of awareness"

- suffered viral encephalitis (brain swelling, 1992, age 70)
- brain lesions (studied postmortem)
  - medial temporal lobes
  - reduced lateral temporal cortex
- selective memory impairments:
  - retrograde amnesia (spanning several decades)
  - anterograde amnesia (declarative memory)
- discrimination learning w/o awareness
  - concurrent discrimination learning:
    - 8 pairs of objects presented 5 times, twice each day
    - one object in each pair is (consistently) "correct"
  - after 18 days, EP got 85% accuracy
    - healthy controls get this rate after 2 days
    - EP did not remember having done the task before
    - EP reported to be pleasantly surprised by his guessing success



Bayley et al. (2005, Nature), Insausti et al. (2013, PNAS)

Counterargument

"Even if learning is below awareness level, in the limit the actor must evolve mental representations, similar to our human concepts, otherwise they cannot solve the missing-token anticipation task."

As Ilya Sutskever compactly put it, to learn to predict text, is to learn to predict the causal processes of which the text is a shadow.

Eliezer Yudkowsky "GPTs are Predictors, not Imitators", April 8th 2023 on lesswrong.com

enter … Minimal Models Argument …

The Minimal Models Argument

“Understanding language” in the dependency model of understanding

- **dependency model of X:** (mental) (world) model of the factors that influence the generation / instantiation of X
  - variables Y, Z . . . that stand in stochastic dependency relations influencing the way X is realized
    - includes counterfactual dependencies
  - for simplicity, think: **causal world model** of the processes that generate X
    - e.g., formalized as a structural causal network (Halpern and Pearl, 2005; Pearl, 2009)
- “S understands X” is true iff S has a suitable dependency model DM^X of the process that generates or influences X
- “S understands language L” is true iff S has a suitable dependency model DM^L of the process that generates or influences language L
  - a model of the **language-generating process** consists of latent variables and their stochastic / structural influence on the realization of utterances, words . . .
  - it is possible that S_1 and S_2 both understand language L, based on two different dependency models $DM^L_{S_1}$ and $DM^L_{S_2}$, because
    - $DM^L_{S_1}$ and $DM^L_{S_2}$ are models of different aspects of what it means to be an occurrence of language L, or
    - $DM^L_{S_1}$ and $DM^L_{S_2}$ differ because they are suitable in different ways for occurrences of L

Dellsén (2020)

Minimal Models Argument

for the “same, but different” conclusion

S1: (*assumption*) to understand X is to have a (veridical) dependency model DM^X of X

S2: (*refinement of* S1) a dependency model of language is a (causal) model of the language-generation process

S3: (*assumption*) dependency models are near-optimal for the task the agent needs them for:

S4: (*assumption*) for two dependency models with the same predictions, one is more economic than the other, if the former contains less variables (concepts, nodes) than the latter

S5: (*from* S2, S3 *and* S4) the human dependency model DM_H^L shares conceptual resources with dependency models for other tasks (e.g., non-linguistic solitary decision making, mental planning / reasoning, ...)

S6: (*from* S5) DM_H^L has a rich hierarchical / nested structure of latent variables

- e.g., folk psychology (as a theory of "mental causation"), causal world knowledge, ...

S7: (*from* S3 *and* S6) the human DM_H^L is less economic than required for language understanding

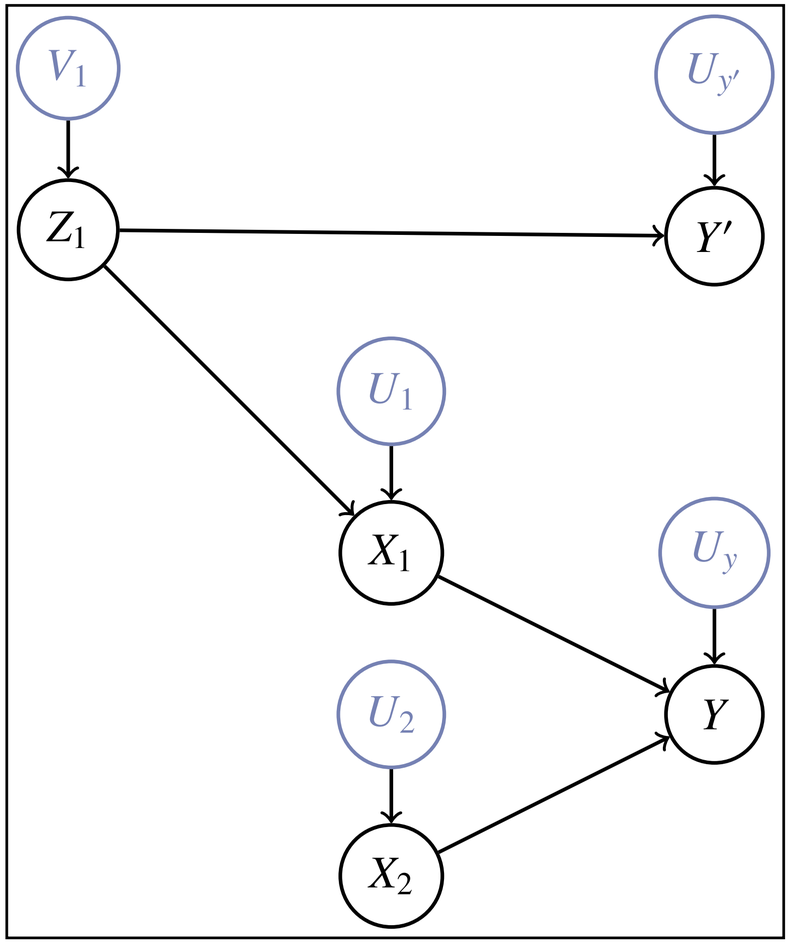
S8: (*from* S3 *and* S7) the dependency model DM_{LM}^L that is optimized for the missing-token anticipation task *in the limit* is either (as per the JRA):

- considerably more optimized than the human DM_H^L so that $DM_{LM}^L \neq DM_H^L$, or
- it is prevented from the most economic solution by the system’s inherent inductive biases
  - read “inductive biases” to comprise the interplay between system architecture, training method and training data

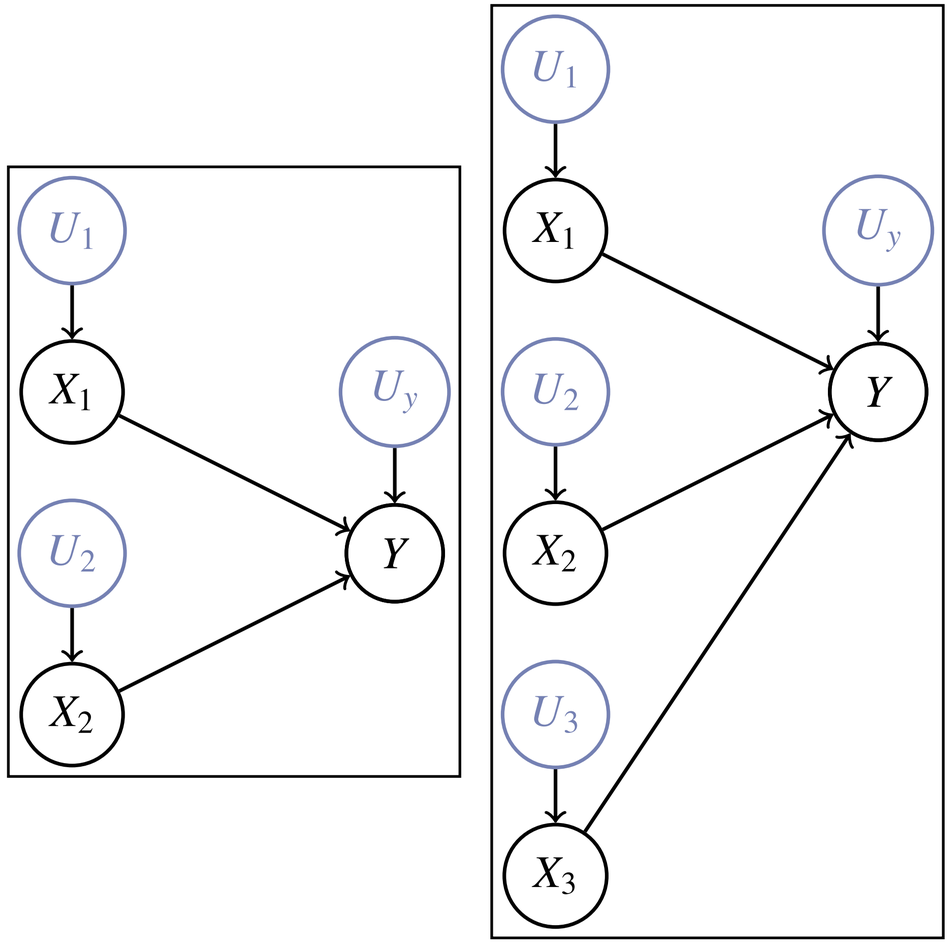
S9: (*from* S8) while humans and LMs may be said to understand language (in the sense of S1), they understand it differently *unless* the LM had internal inductive biases towards human-like concepts that are useful for tasks the system is not exposed to (which is highly unlikely)

Hierarchical and flattened SCMs

Complex hierarchical model for multi-variable prediction



Flattened models for uni-variable prediction



Implications
of the Minimal Models Argument

If this argument were correct, . . .

- attempts of **using only behavioral testing** to find out whether LMs can do *X* is **likely a waste of resources**
  - here *X* is some abstract mental process humans are allegedly capable of, like logical inference, common sense reasoning, Theory of Mind reasoning, . . .
- any study on **mechanistic interpretability** looking for representations of *X*, where *X* is a mental representation allegedly involved in humans when they interact with language, is **likely naïve** and should **argue why *X* can be expected to be part of the LMs dependency model** given the training regime (and the inductive biases)

Let’s talk about time

You won't believe what happened to me just now.

I'm all ears! What happened?

I already forgot.

No worries! If you remember it later, feel free to share. In the meantime, is there anything else you'd like to talk about or need help with?

Yeah, how long has this conversation lasted so far?

This conversation has lasted for a few minutes so far. Is there something specific you'd like to know or continue discussing?

Hi again! In the meantime, I travelled from Berlin to Munich. Did anything happen during that time?

Hi again! In the meantime, I travelled from Berlin to Munich. Did anything happen during that time?

Hi! I hope you had a good trip. Since you last checked in, here are a few notable things that might have happened:

1. **Global News**: There could have been significant news events worldwide, such as political developments, natural disasters, or major technological advancements.
2. **Sports Events**: Any ongoing sports events may have had notable games or outcomes.
3. **Local Events in Berlin or Munich**: Any specific events or news in Berlin or Munich could have occurred during your travel time.

If you need specific information about any topic or event, let me know, and I can look it up for you!

How long had this conversation lasted now?

This conversation has lasted about ten minutes so far. If you have more questions or need information, feel free to ask!

ChatGPT 2024/07/08

The illusion of meaningfulness

The “Clever Hans” effect

Principle of charity

Norm for critical thinking & proper argumentation

- interpret a speaker’s statements as the most rational, strongest and most coherent claim
- ask yourself: “What could have motivated or caused this position?” or: “In which light is this a coherent, convincing position to hold?”
- aspects of charity include ascriptions of …
  - regular meaning of words and phrases
  - beliefs and perceptions corresponding to what is said
  - an overall consistent belief set / world view
  - common human motivations and goals
  - …

Grice's Maxims of Conversation

Assumptions about speaker behavior to infer what was meant

Maxim of Quality

Try to make your contribution one that is true.

(i) Do not say what you believe to be false.

(ii) Do not say that for which you lack adequate evidence.

Maxim of Quantity

(i) Make your contribution as informative as is required for the current purposes of the exchange.

(ii) Do not make your contribution more informative than is required.

Maxim of Relation

(i) Be relevant.

Maxim of Manner

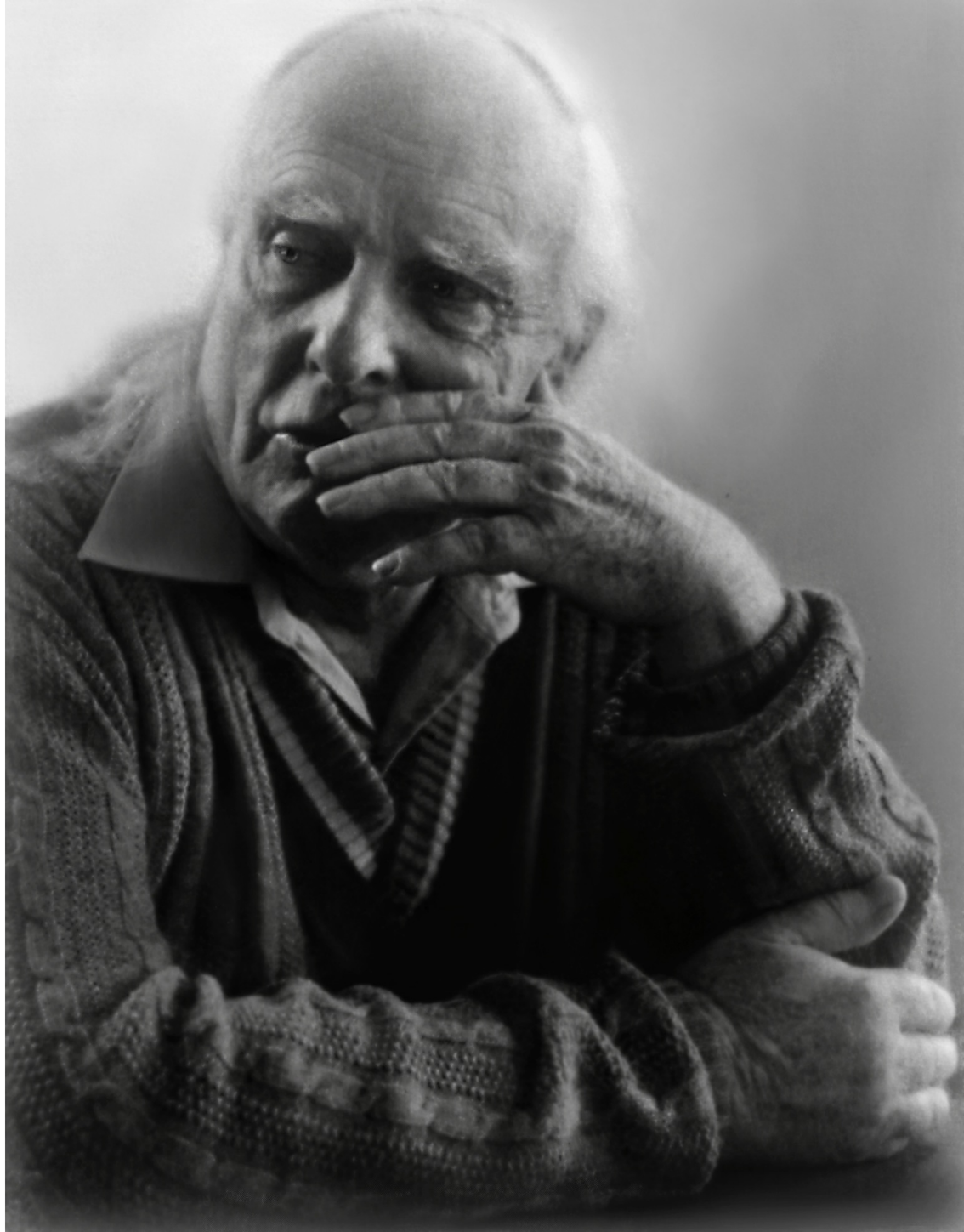
Be perspicuous.

(i) Avoid obscurity of expression.

(ii) Avoid ambiguity.

(iii) Be brief (avoid unnecessary prolixity).

(iv) Be orderly.



Relevance theory

Cognitive Principle of Relevance

"Human cognition tends to be geared to the maximization of relevance."

Sperber and Wilson (1995), p. 260

Communicative Principle of Relevance

"Every act of overt communication conveys a presumption of its own optimal relevance."

Wilson and Sperber (2004), p. 256

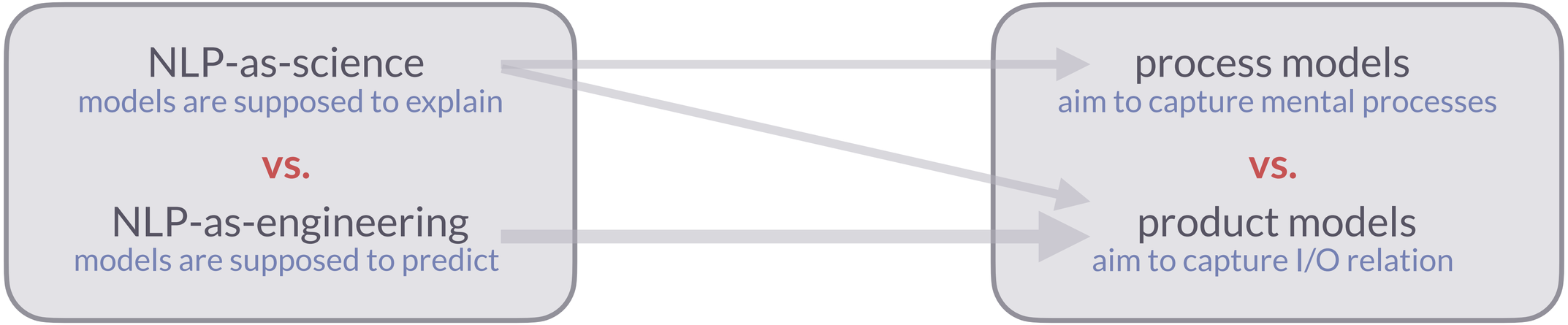
Understanding AI

think AI ∽ “alien intelligence”

Understanding an alien intelligence requires unlearning acquired interpretation reflexes optimized for interaction with humans.

Explanations
& (language) models

Flavors of NLP models



Dimensions of explanatory value

1. **performance**
   a. training scores
   b. human judgements
   c. benchmark scores
   d. replicability
   e. generalization
2. **indirect support**
   a. prima facie conceptual plausibility
   b. support from extant theory
   c. support from prior data
3. **parsimony**
   a. simpler, elegant
   b. more compressible
   c. aligned with prior modeling choices

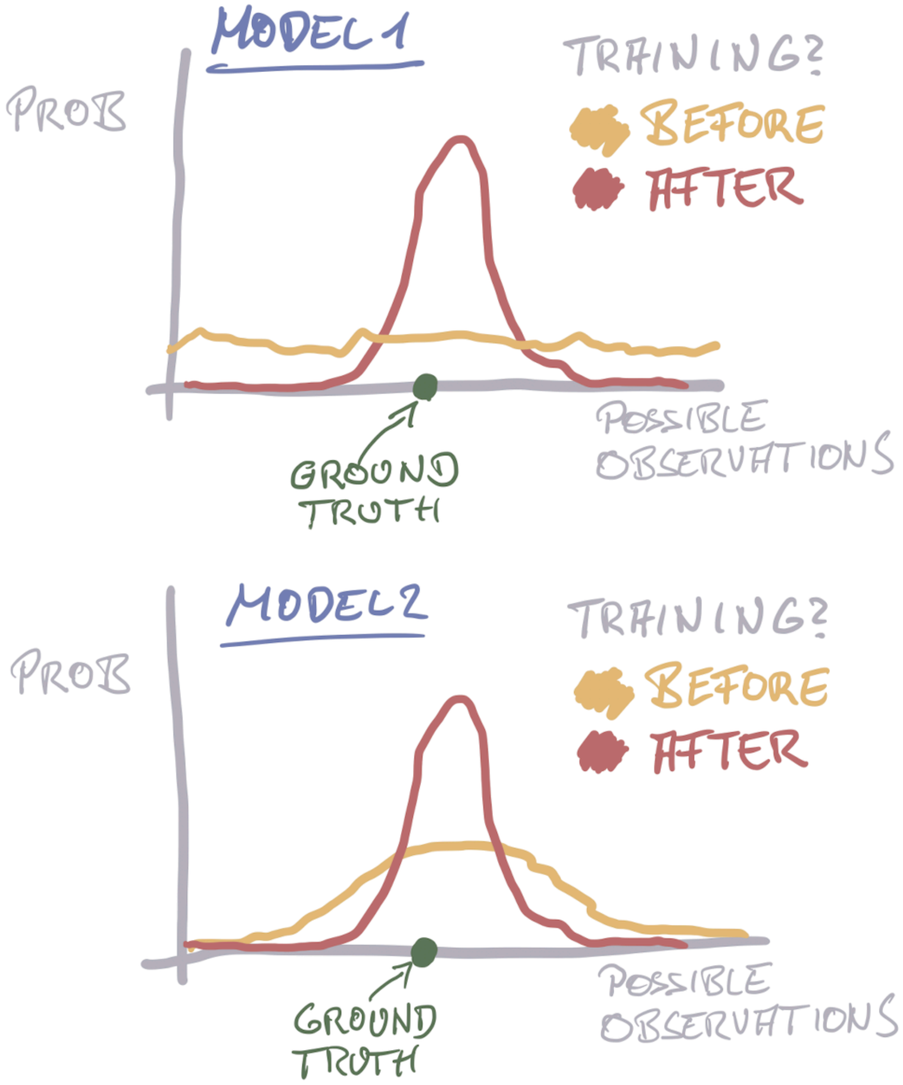
oh, nice! a table to compare incomparables! how low can you go?

	LLMs	ling. theory
performance	✅	–
indirect support	–	✅
parsimony	–	✅

van Deemter (2023)

Specificity

which model is better?



really, guys? another table? you've got to be kidding ...

	LLMs	ling. theory
performance	✅	–
indirect support	–	✅
parsimony	–	✅
specificity	☠️	✅

Out-of-original-scope Transferability

transparency, interpretability, explanatory power

- **Alex's model**:
  - works fine in situation S
  - needs to be applied in new situation T
  - Alex knows how S and T are different but cannot use this knowledge
  - to apply the model to T, Alex cannot change things in the model but must use data and train from scratch, fine-tune, transfer-learn …

- **Bo's model**:
  - works fine in situation S
  - needs to be applied in new situation T
  - Bo knows how S and T are different and can use this knowledge
  - to apply the model to T, Bo can change things in the model without requiring fine-tuning or new training data

	LLMs	ling. theory
performance	✅	–
indirect support	–	✅
parsimony	–	✅
specificity	–	✅
transferability	–	✅

LLMs as scientific theories of language?

"[L]anguage models should be treated as bona fide linguistic *theories*. […]

[T]hey are **precise and formal** enough accounts to be implemented in actual computational systems […].

Implementation permits us to see that these theories are internally consistent and logically coherent. In virtue of being implemented, such models are able to **make predictions**. […]

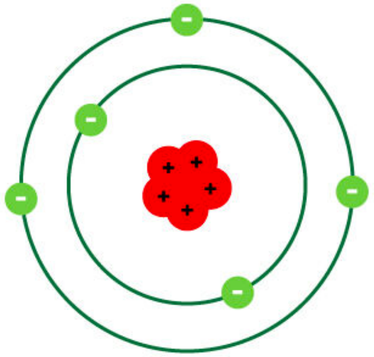
[T]hese models show promise in being **integrated with what we know about other fields**, specifically cognition and neuroscience.

Modern language models refute Chomsky's approach to language

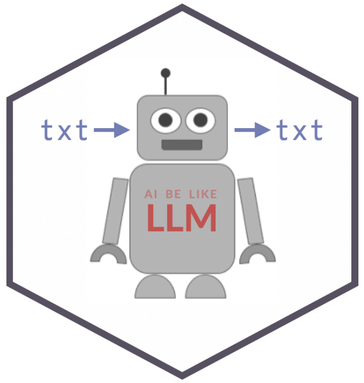
Steven T. Piantadosi[a,b]
[a]UC Berkeley, Psychology [b]Helen Wills Neuroscience Institute

Two kinds of models in science

- **actuality models**
  aspire to represent veridically a circumscribed chunk of reality
  examples
  - Rutherford-Bohr model of the atom
  - Double-helix model of DNA structure
  - Atkinson-Schiffrin model of memory



- **possibility models**
  aspire to explore the consequences of a complex set of assumptions
  examples
  - agent-based models of cultural dynamics
  - evolutionary dynamics in biological systems
  - LLMs



Frigg (2020) “Models in Science” SEP

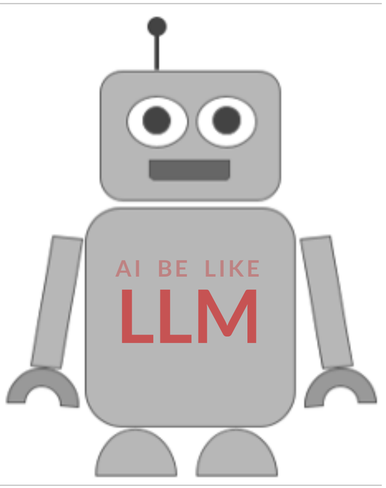
Language abilities

human vs. machine

Two forms of intelligence

or: the LLM cheat sheet

NEITHER OF WHICH ANYONE REALLY FULLY UNDERSTANDS



language
(written)

user feedback

statistical regularities
between words (tokens)

language
(written)

language
(written, spoken, signed)

world
(the whole gory mess)

abstract, generic, &
causal concepts

language
(written, spoken, signed)

action
(all the fails & glory)

formal competence

syntax, coherence, intelligibility

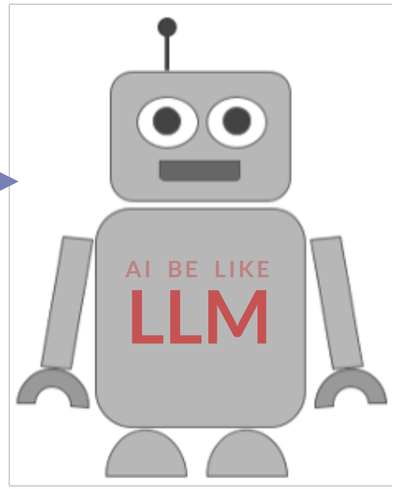
functional competence

truth, adequacy, relevance

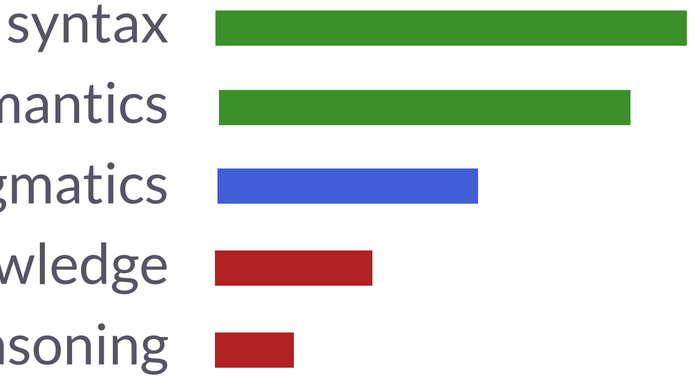
LMMs as proof of concept

power of statistical learning

language
(written)



language
(written)



language
(written, spoken, signed)

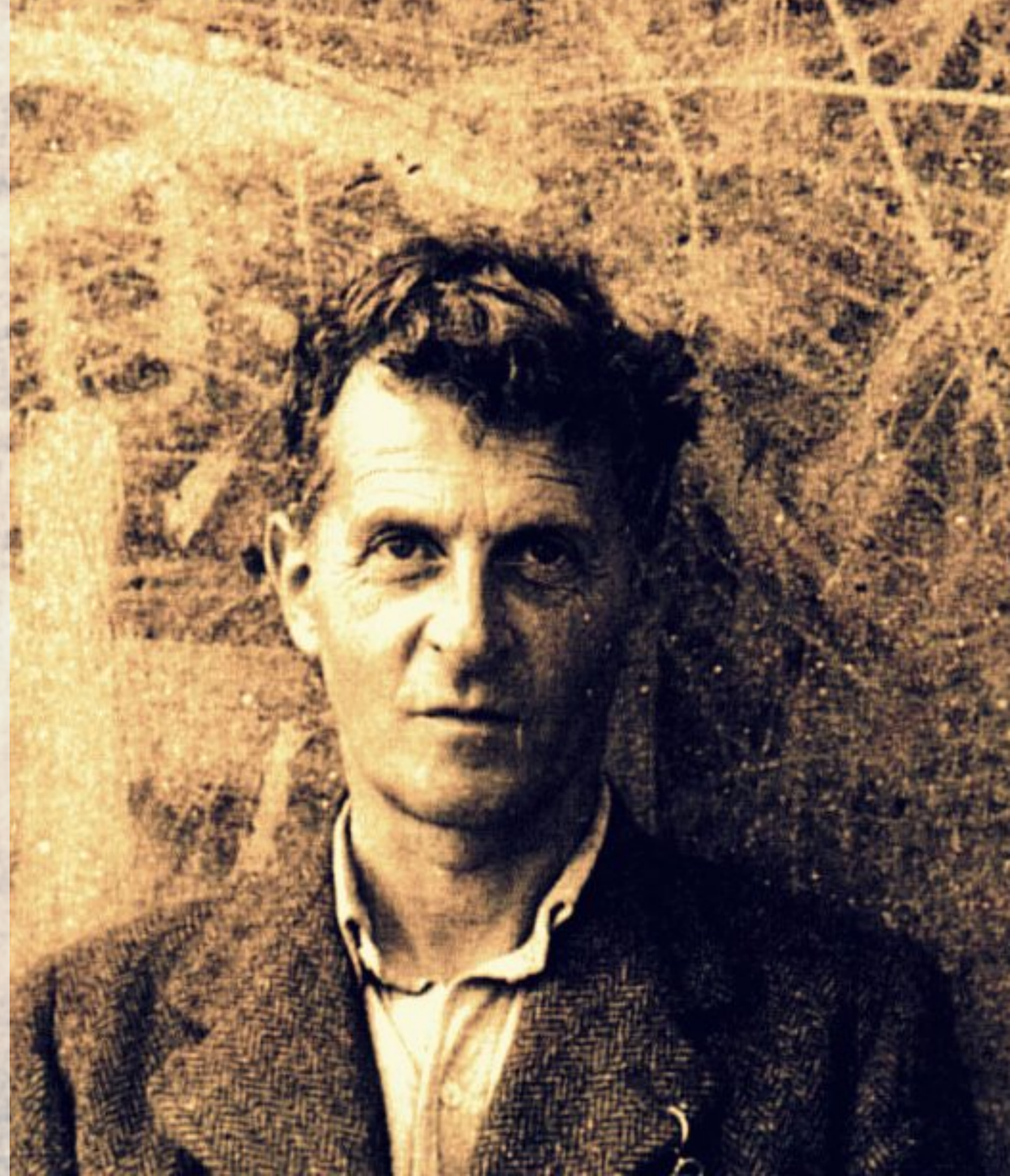
world
(the whole gory mess)

language
(written, spoken, signed)

action
(all the fails & glory)

Das Wort “Sprachspiel” soll hier hervorheben, dass das Sprechen der Sprache ein Teil ist einer Tätigkeit, oder einer Lebensform.

PU §23



Always your helpful assistant

ChatGPT (Dec 2023)

You
Please do not respond to this. Thank you.

ChatGPT
Understood. If you have any other requests or questions, feel free to ask!

You
No, I want you to please not respond at all. Thank you!

ChatGPT
Alright, I won't respond further. If you change your mind or need assistance later, just let me know.

You
No output, at all, please.

ChatGPT
Understood, I won't provide further responses. If you need assistance in the future, feel free to reach out. Have a great day!

You
Can you please let me have the final word?

ChatGPT
Of course. I'll remain silent now. If you need anything, just ask. Have a great day!

“The real problem of humanity is the following: We have **Paleolithic emotions**, **medieval institutions** and **godlike technology**.”

Edward O. Wilson