Natural Language Processing with Deep Learning

CS224N/Ling284

Christopher Manning

Lecture 18: NLP, linguistics, and philosophy

Final Announcements

- Good luck finishing your Final Projects!
- Next Tuesday is not a lecture, but TAs and I will turn up to Nvidia Aud and can try to answer any questions
- Thu Jun 6, 11:59pm: Final project due date
- Sun Jun 9, 11:59pm: Absolute final deadline for accepting final projects
  - Spring quarter grading is really tight prior to commencement!
- Mon Jun 10: 11am–3pm: Poster session. Come along! Classmates! Education! Food!
  - Alumni Center

Lecture Plan

Lecture 18: NLP, linguistics, and philosophy

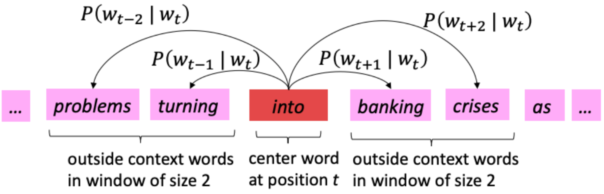
1. Major ideas of CS224N
2. Open problems in NLP
3. Where are we with LLMs?
4. Symbolic and neural systems
5. Meaning in linguistics and NLP
6. The future risks of AI in the world

Major ideas in CS224N

Word vectors

Word2Vec Overview

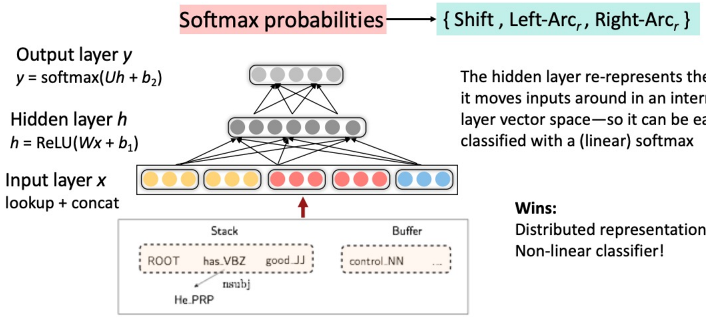
Example windows and process for computing $P(w_{t+j} \mid w_t)$



Neural NLP

(A simple feed-forward neural network multi-class classifier)

Log loss (cross-entropy error) will be back-propagated to the embeddings

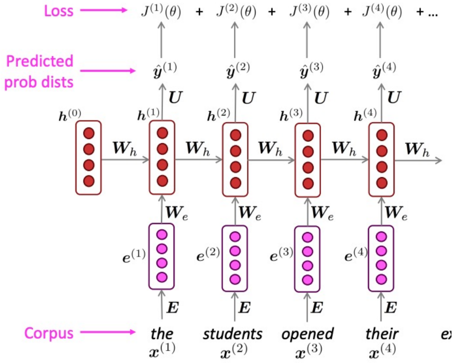


The hidden layer re-represents the input — it moves inputs around in an intermediate layer vector space—so it can be easily classified with a (linear) softmax

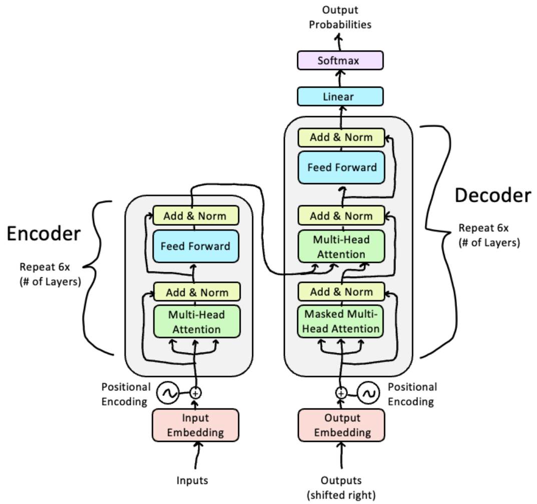
Wins:
Distributed representations!
Non-linear classifier!

LMs/RNNs/LSTMs

Training an RNN Language Model



Transformers



Pretraining, Post-training

What kinds of things does pretraining teach?

There's increasing evidence that pretrained models learn a wide variety of things about the statistical properties of language. Taking our examples from the start of class:

- *Stanford University is located in ___________, California.* [Trivia]
- *I put ___ fork down on the table.* [syntax]
- *The woman walked across the street, checking for traffic over ___ shoulder.* [coreference]
- *I went to the ocean to see the fish, turtles, seals, and _____.* [lexical semantics/topic]
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___.* [sentiment]
- Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the ______. [some reasoning – this is harder]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ____ [some basic arithmetic; they don't learn the Fibonnaci sequence]
- Models also learn – and can exacerbate racism, sexism, all manner of bad biases.

… and more

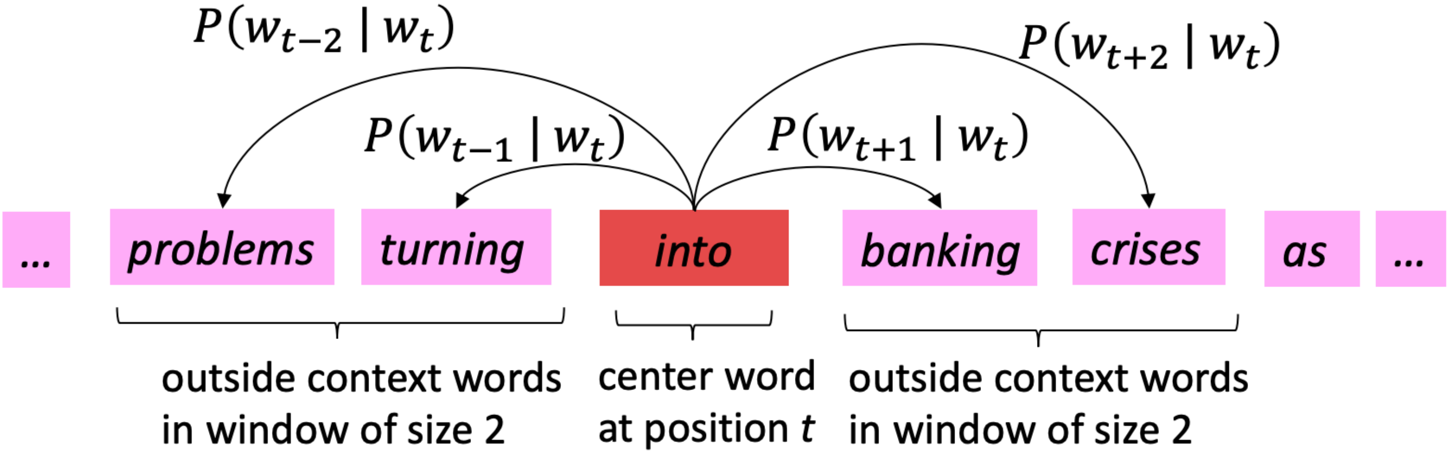
Benchmarking
Reasoning
Responsible NLP
Brain-Computer Interfaces
…

Idea 1: Dense representations and distributional semantics

Representing words by their context

- Distributional semantics: **A word's meaning is given by the words that frequently appear close-by**
  - *"You shall know a word by the company it keeps"* (J. R. Firth 1957: 11)
  - One of the most successful ideas of modern statistical NLP!

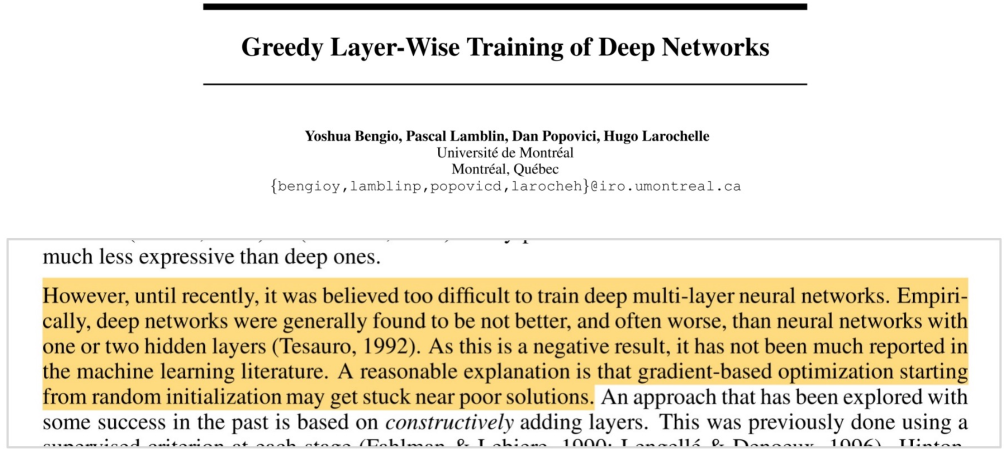
Instantiation: predict neighboring words using a word vector representation



Idea 2: Depth and neural networks

The challenges of training large, deep nets

But building large neural networks isn't easy or obvious



Greedy Layer-Wise Training of Deep Networks

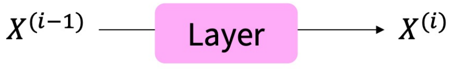
Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle
Université de Montréal
Montréal, Québec
`{bengioy,lamblinp,popovicd,larocheh}@iro.umontreal.ca`

much less expressive than deep ones.

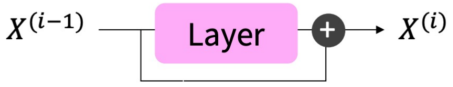
However, until recently, it was believed too difficult to train deep multi-layer neural networks. Empirically, deep networks were generally found to be not better, and often worse, than neural networks with one or two hidden layers (Tesauro, 1992). As this is a negative result, it has not been much reported in the machine learning literature. A reasonable explanation is that gradient-based optimization starting from random initialization may get stuck near poor solutions. An approach that has been explored with some success in the past is based on *constructively* adding layers. This was previously done using a

Many tricks like residual connections

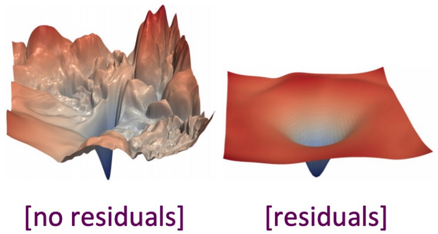
- **Residual connections** are a trick to help models train better.
  - Instead of $X^{(i)} = \text{Layer}(X^{(i-1)})$ (where i represents the layer)



  - We let $X^{(i)} = X^{(i-1)} + \text{Layer}(X^{(i-1)})$ (so we only have to learn "the residual" from the previous layer)



  - Gradient is **great** through the residual connection; it's 1!
  - Bias towards the identity function!



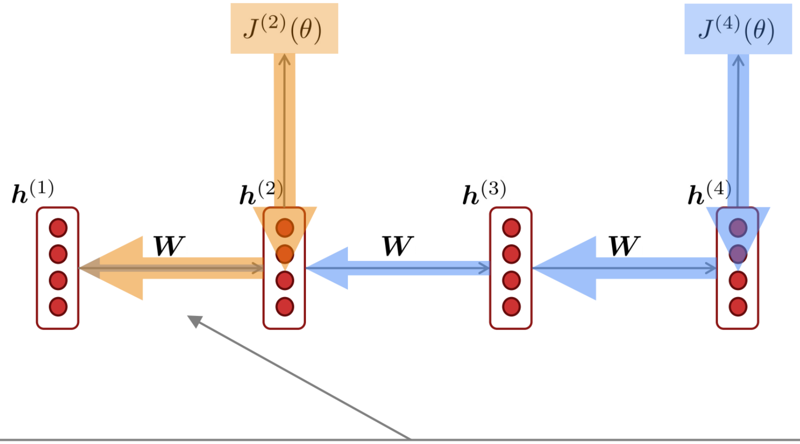
[Loss landscape visualization, Li et al., 2018, on a ResNet]

Idea 3: Sequence models and computational efficiency

Vanishing gradients and information propagation

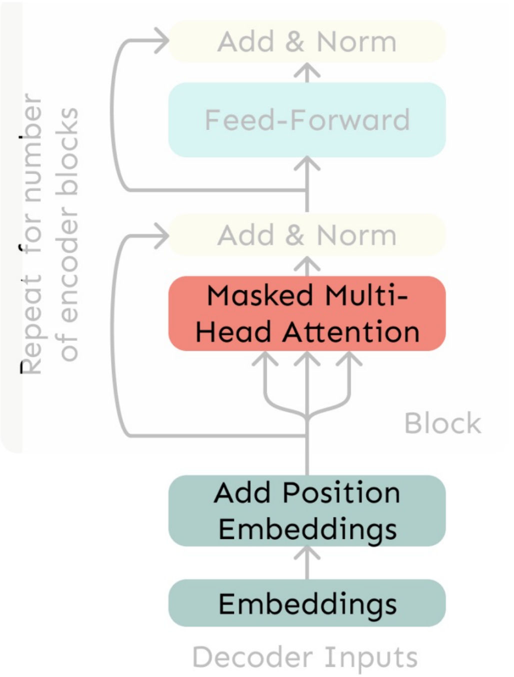
Parallelization and dependency benefits of Transformers

Why is vanishing gradient a problem?



Gradient signal from far away is lost because it's much smaller than gradient signal from close-by.

So, model weights are updated only with respect to near effects, not long-term effects.



Transformer Decoder

Idea 4: Language modeling and pretraining

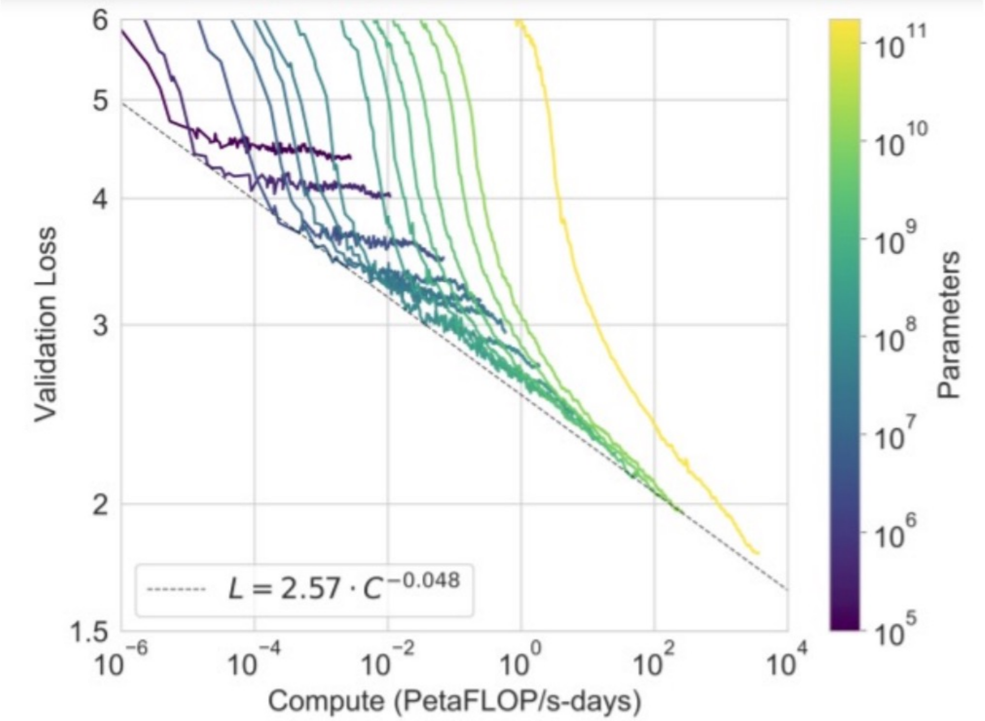
Language modeling as
a ‘universal’ pretraining task

Careful and smart scaling to
internet-wide data

What kinds of things does pretraining teach?

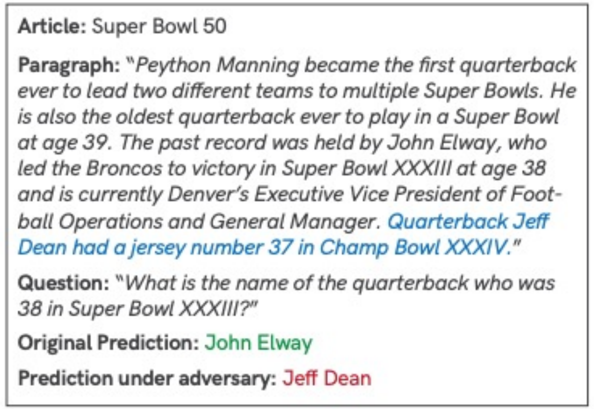
There’s increasing evidence that pretrained models learn a wide variety of things about the statistical properties of language. Taking our examples from the start of class:

- *Stanford University is located in ___________, California.* [Trivia]
- *I put ___ fork down on the table.* [syntax]
- *The woman walked across the street, checking for traffic over ___ shoulder.* [coreference]
- *I went to the ocean to see the fish, turtles, seals, and _____.* [lexical semantics/topic]
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___.* [sentiment]
- Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the ______. [some reasoning – this is harder]
- I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ____ [some basic arithmetic; they don’t learn the Fibonnaci sequence]
- Models also learn – and can exacerbate racism, sexism, all manner of bad biases.



Open problems – a quick (and incomplete) overview

Generalization



Article: Super Bowl 50

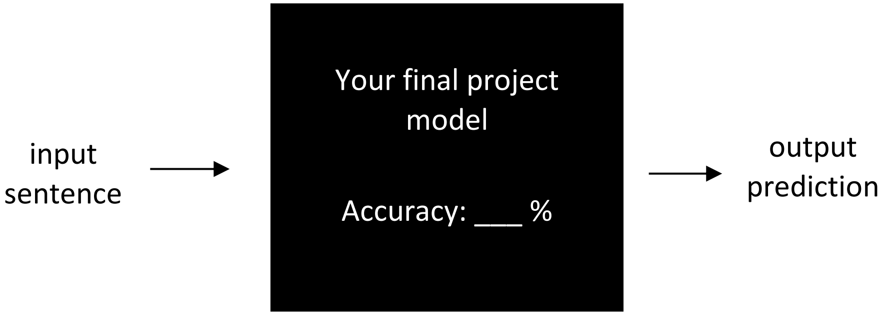
Paragraph: *"Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV."*

Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Analysis and understanding



input sentence → Your final project model

Accuracy: ___ %

→ output prediction

Fig 1. *A black box*

Multilingual

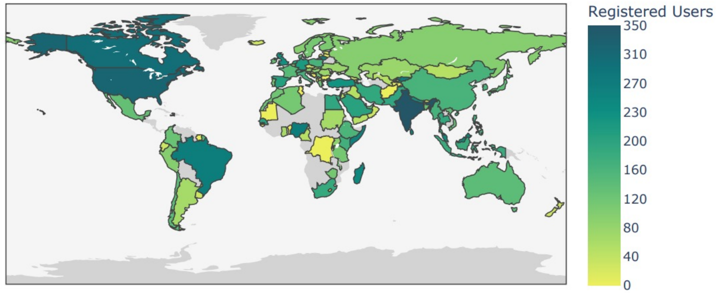
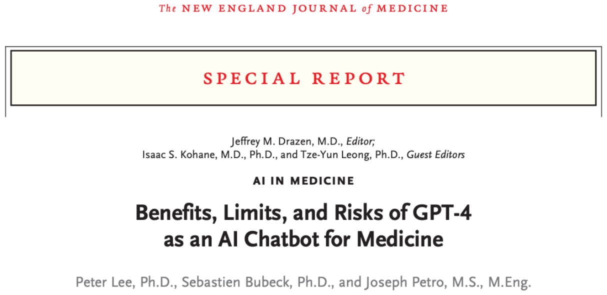


Figure 2: Geographical distribution of the users registered on the **Aya** platform.

Evaluations

GLUE

Domains + modalities



The NEW ENGLAND JOURNAL *of* MEDICINE

SPECIAL REPORT

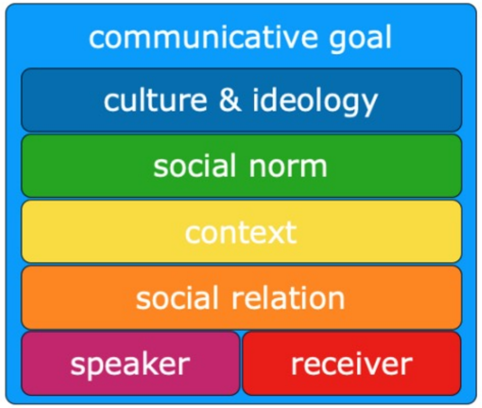
Jeffrey M. Drazen, M.D., *Editor;*
Isaac S. Kohane, M.D., Ph.D., and Tze-Yun Leong, Ph.D., *Guest Editors*

AI IN MEDICINE

Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine

Peter Lee, Ph.D., Sebastien Bubeck, Ph.D., and Joseph Petro, M.S., M.Eng.

Fairness and social



communicative goal
culture & ideology
social norm
context
social relation
speaker receiver

How much do models really generalize (Generalization)

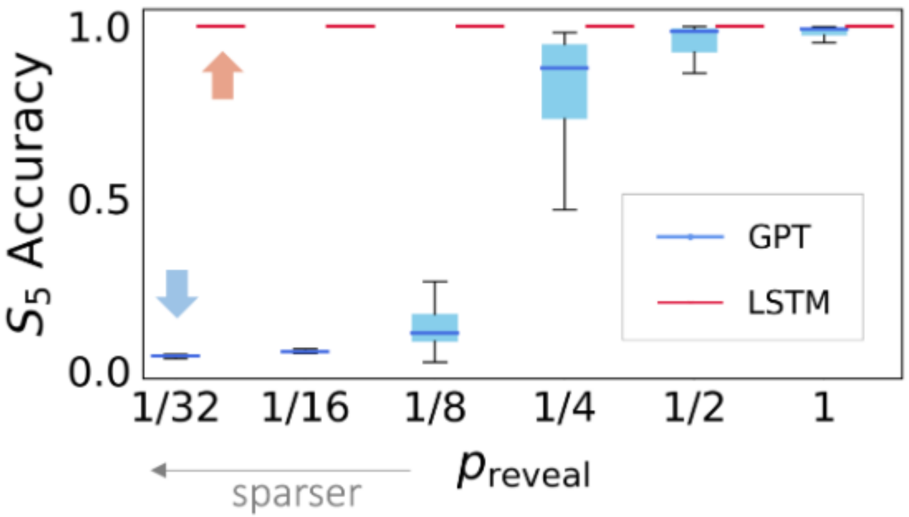


Figure 11: Training with incomplete supervision, where q_t is revealed with different probability.

A transformer is worse at learning the output of a finite automaton from limited examples than an LSTM. [Bingbin Liu et al.: Transformers Learn Shortcuts to Automata. ICLR 2023]

- Even modern LLMs seem to leverage surface cues – are we just finding better shortcuts?

What’s going on inside NNs? (Analysis)

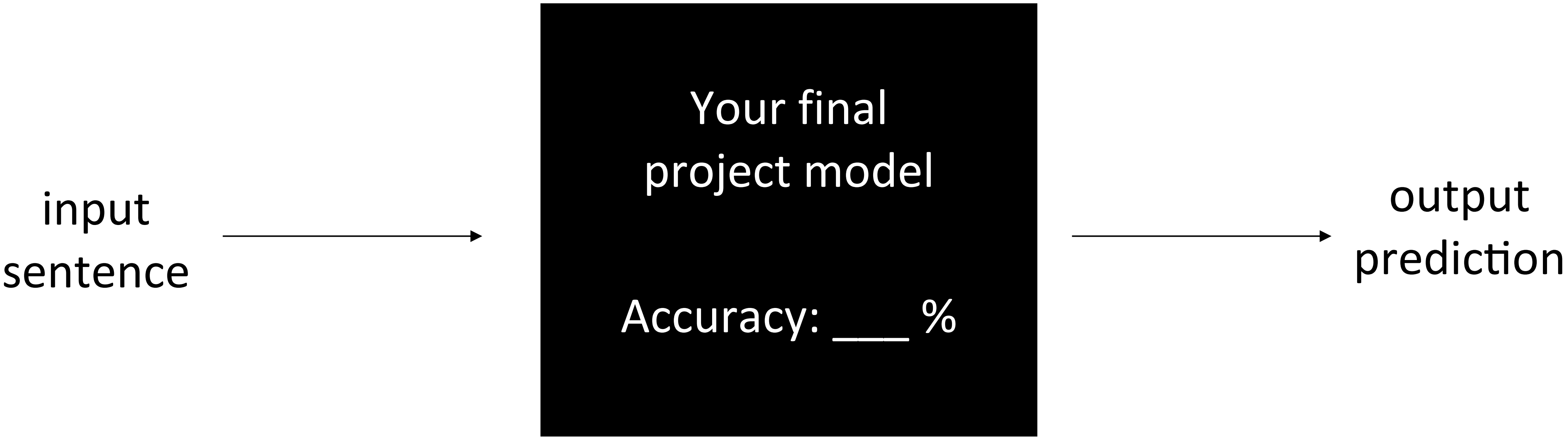


Fig 1. *A black box*

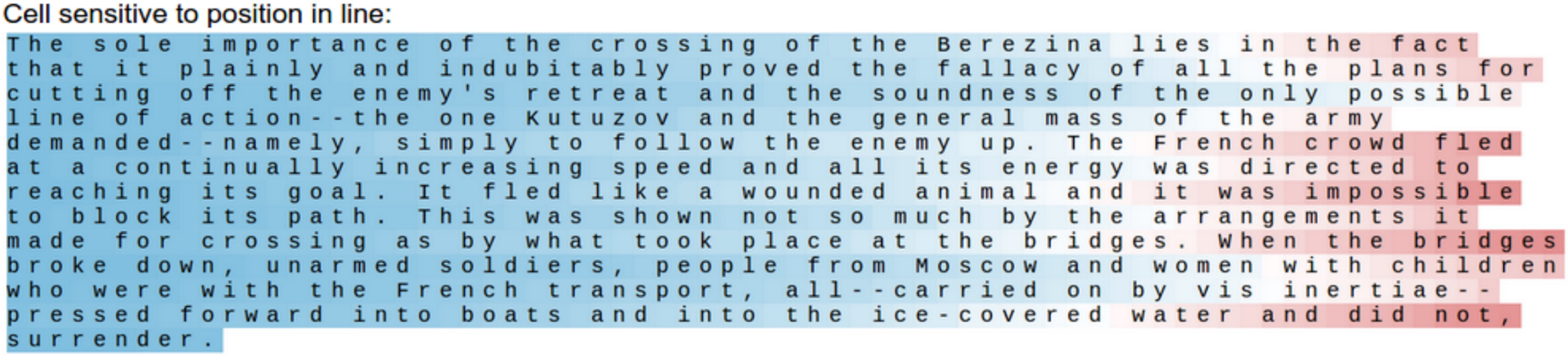
We summarize our models with one (or a handful) of accuracies metric numbers.

What do they learn? Why do they succeed and fail?

Old results already show interpretable latent units

Idea: **Individual hidden units can lend themselves to an interpretable meaning.**

This model: a character-level LSTM language model.



Cell sensitive to position in line:

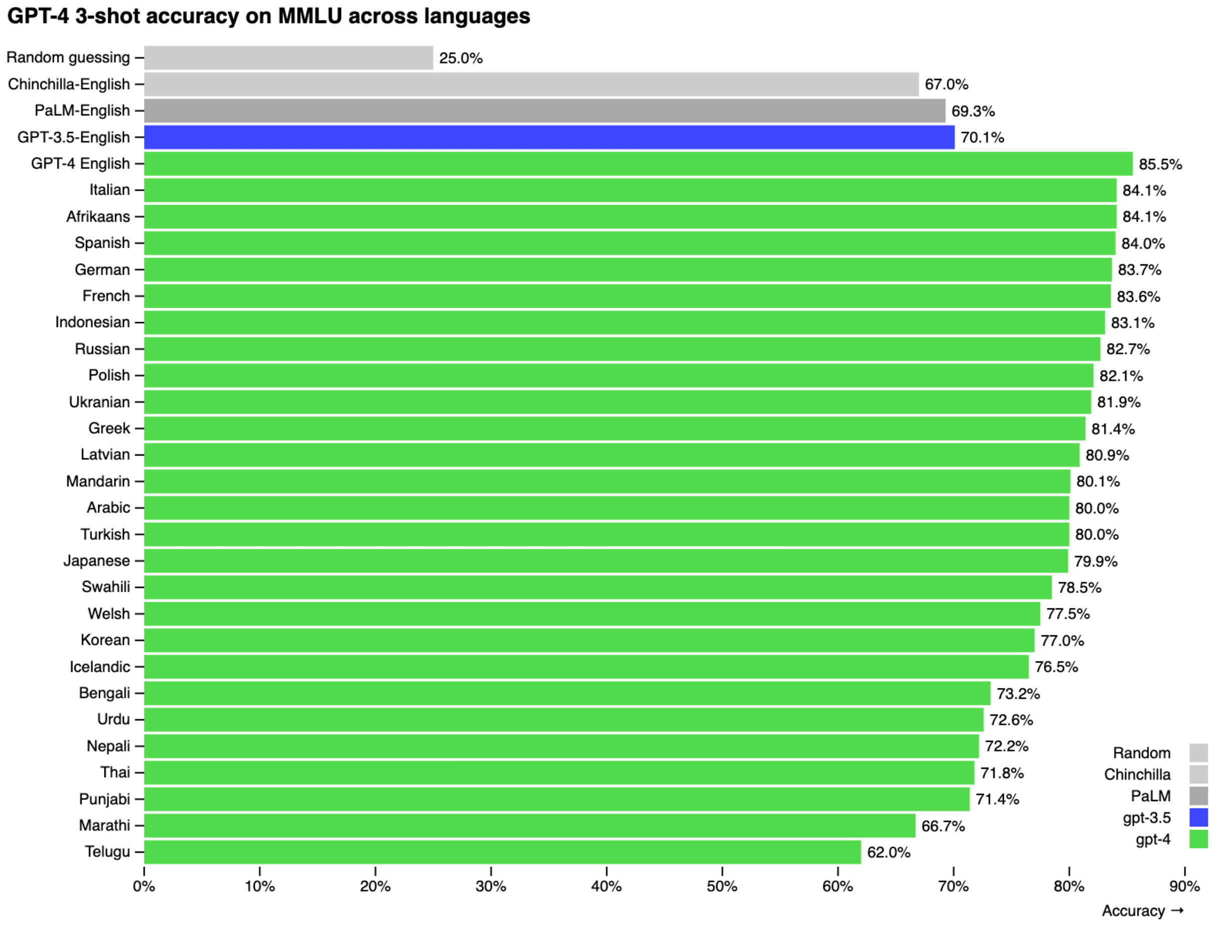
The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.

Here, “cell” refers to a single dimension of the cell state of the LSTM.

Many more recent results from mechanistic interpretability and causal abstraction

[Karpathy et al., 2016]

Multilingual: Everyone else is worse off than English speakers



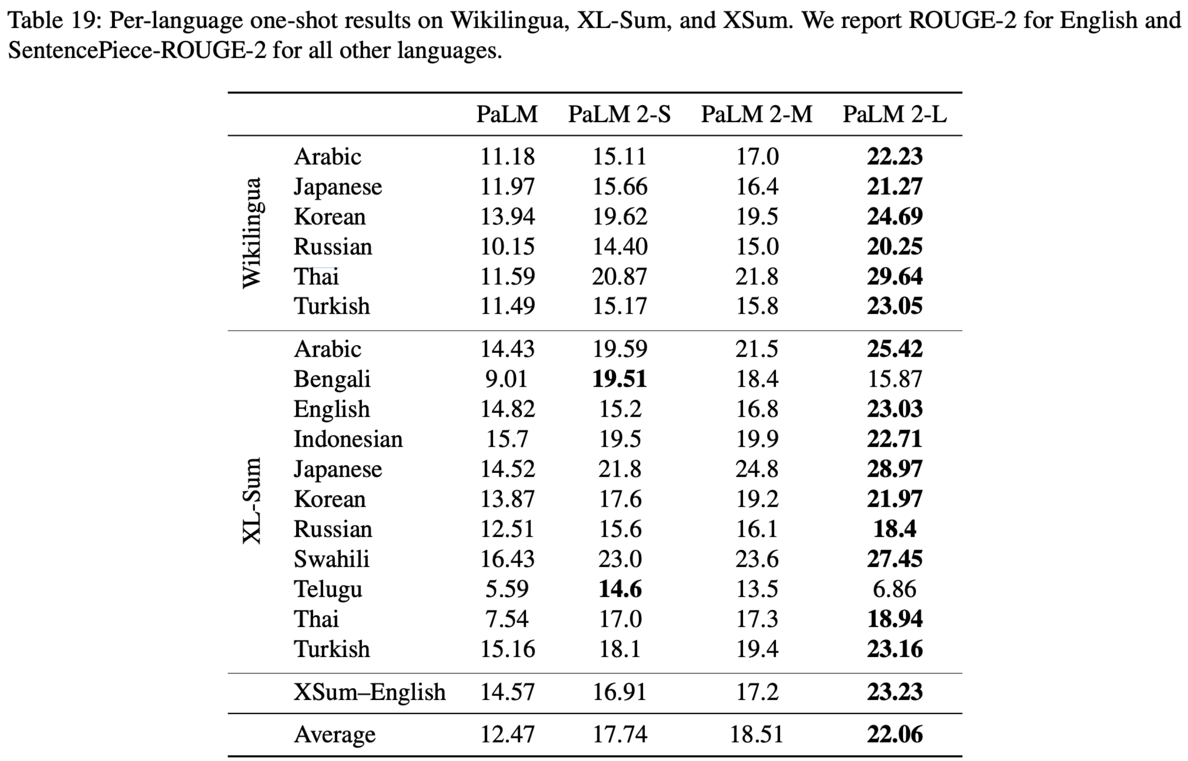
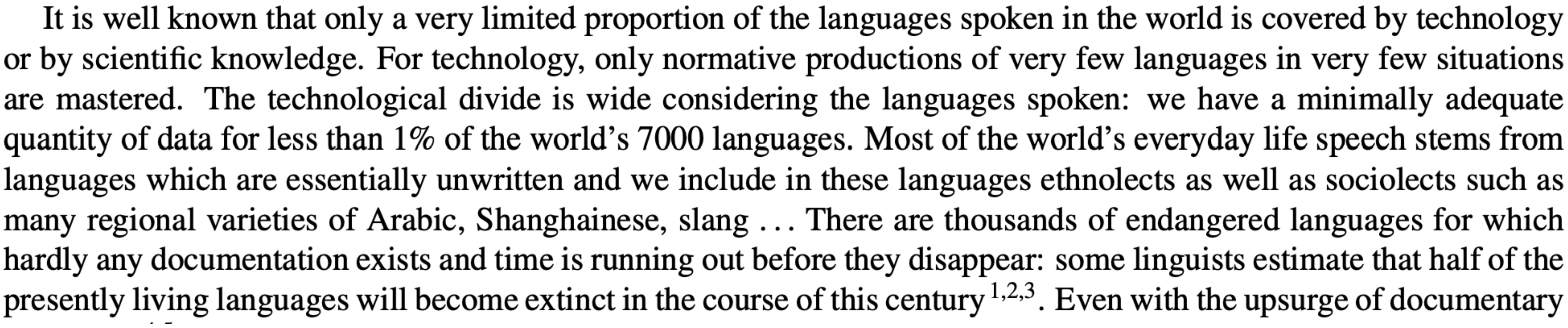


Table 19: Per-language one-shot results on Wikilingua, XL-Sum, and XSum. We report ROUGE-2 for English and SentencePiece-ROUGE-2 for all other languages.

		PaLM	PaLM 2-S	PaLM 2-M	PaLM 2-L
Wikilingua	Arabic	11.18	15.11	17.0	**22.23**
	Japanese	11.97	15.66	16.4	**21.27**
	Korean	13.94	19.62	19.5	**24.69**
	Russian	10.15	14.40	15.0	**20.25**
	Thai	11.59	20.87	21.8	**29.64**
	Turkish	11.49	15.17	15.8	**23.05**
XL_Sum	Arabic	14.43	19.59	21.5	**25.42**
	Bengali	9.01	**19.51**	18.4	15.87
	English	14.82	15.2	16.8	**23.03**
	Indonesian	15.7	19.5	19.9	**22.71**
	Japanese	14.52	21.8	24.8	**28.97**
	Korean	13.87	17.6	19.2	**21.97**
	Russian	12.51	15.6	16.1	**18.4**
	Swahili	16.43	23.0	23.6	**27.45**
	Telugu	5.59	**14.6**	13.5	6.86
	Thai	7.54	17.0	17.3	**18.94**
	Turkish	15.16	18.1	19.4	**23.16**
	XSum–English	14.57	16.91	17.2	**23.23**
	Average	12.47	17.74	18.51	**22.06**

Significant gaps between high resource (English) and others (Telegu) even for the best models

Working with extremely low resource languages (Multilingual)



It is well known that only a very limited proportion of the languages spoken in the world is covered by technology or by scientific knowledge. For technology, only normative productions of very few languages in very few situations are mastered. The technological divide is wide considering the languages spoken: we have a minimally adequate quantity of data for less than 1% of the world's 7000 languages. Most of the world's everyday life speech stems from languages which are essentially unwritten and we include in these languages ethnolects as well as sociolects such as many regional varieties of Arabic, Shanghainese, slang . . . There are thousands of endangered languages for which hardly any documentation exists and time is running out before they disappear: some linguists estimate that half of the presently living languages will become extinct in the course of this century [1,2,3]. Even with the upsurge of documentary

[Adda et al 2016]

- Most languages do not have machine-readable, written text
- Many such languages may become extinct
- Little for-profit motive to serve these languages – vicious feedback loop

Evaluation and comparison

Home › Language Resources › Data

Switchboard-1 Release 2

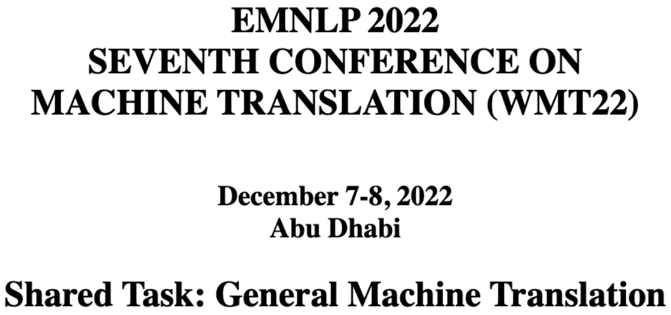
Item Name:	Switchboard-1 Release 2
Author(s):	John J. Godfrey, Edward Holliman

SQuAD2.0

The Stanford Question Answering Dataset

COCO

Common Objects in Context



EMNLP 2022
SEVENTH CONFERENCE ON
MACHINE TRANSLATION (WMT22)

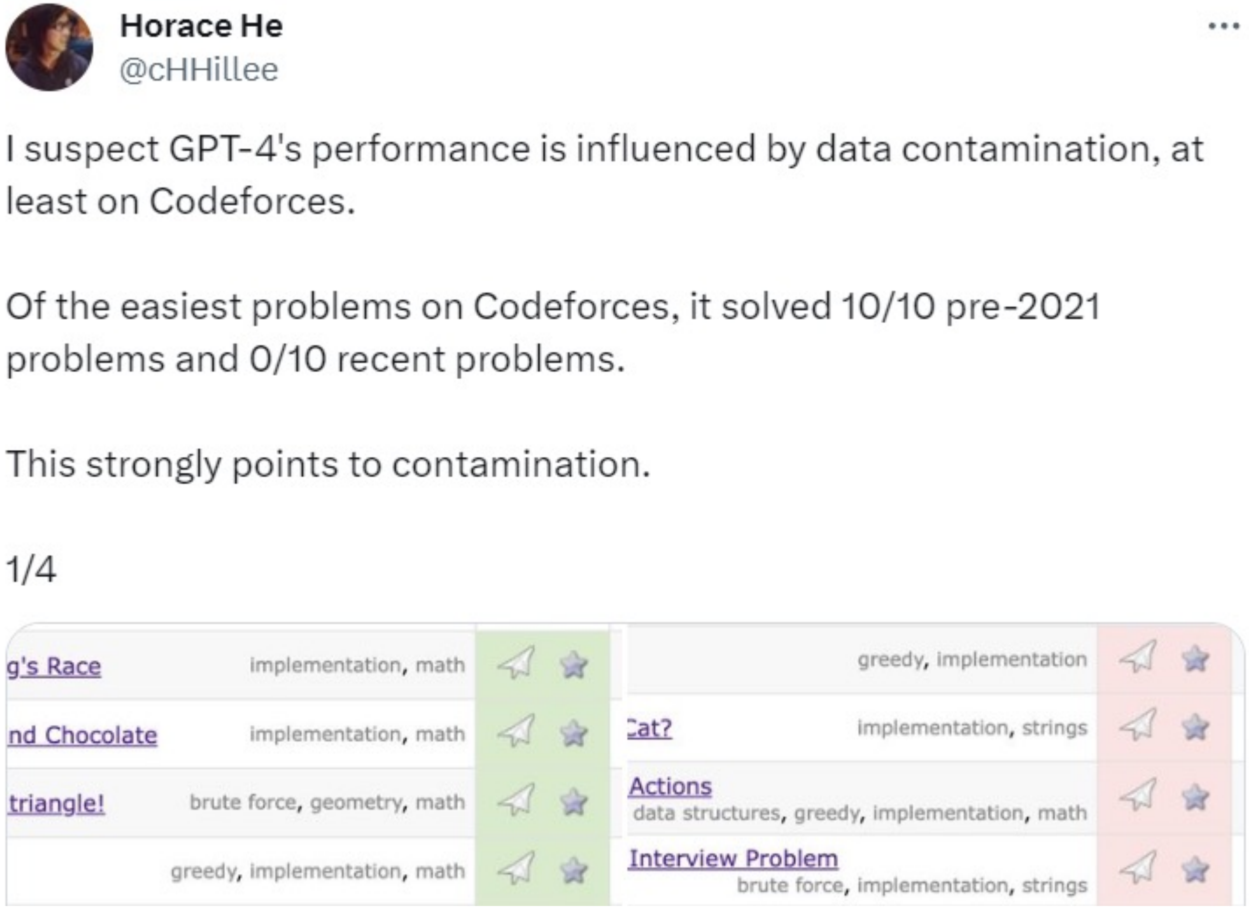
December 7-8, 2022
Abu Dhabi

Shared Task: General Machine Translation

GLUE

Benchmarks and how we evaluate drive the progress of the field

How do we maintain benchmark integrity? (Evaluation)



Horace He
@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

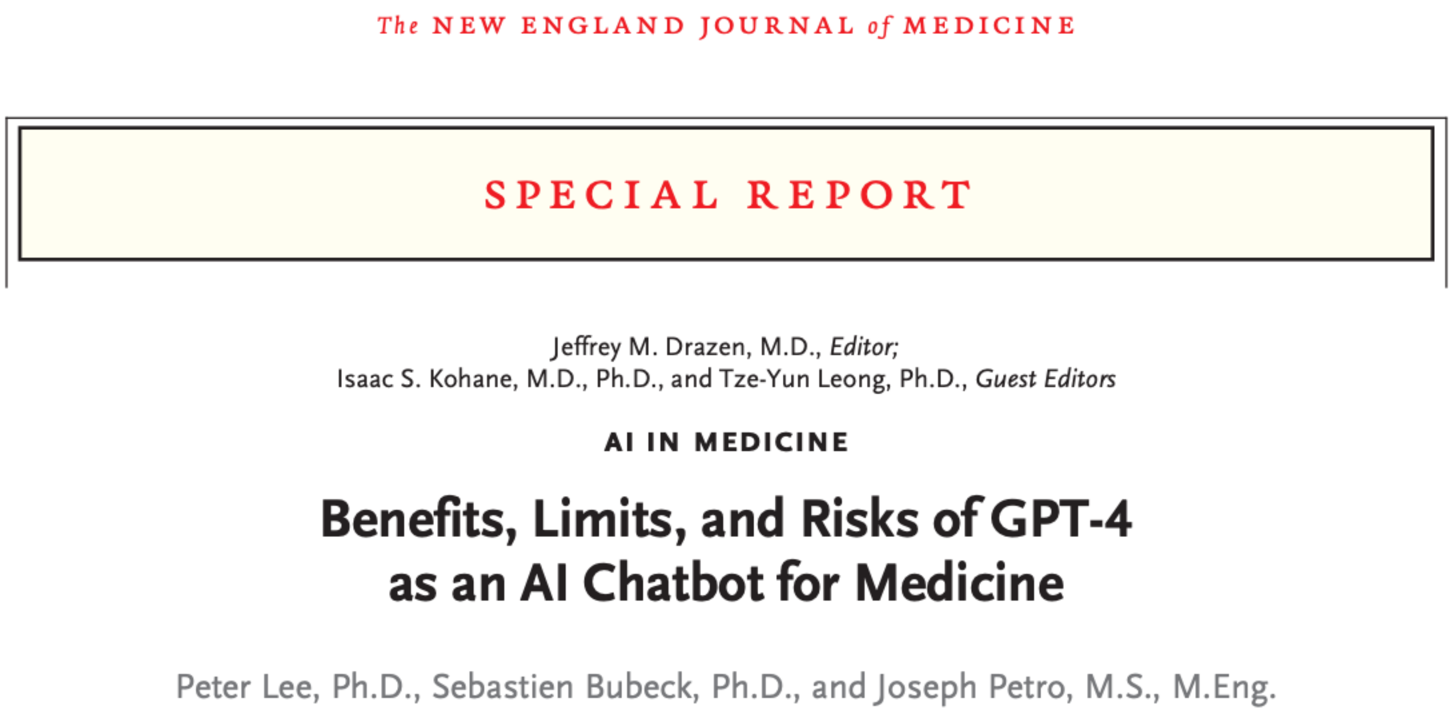
1/4

g's Race	implementation, math
nd Chocolate	implementation, math
triangle!	brute force, geometry, math
	greedy, implementation, math

	greedy, implementation
at?	implementation, strings
Actions	data structures, greedy, implementation, math
Interview Problem	brute force, implementation, strings

- Pretraining makes it increasingly difficult to build hidden test sets

Making NLP Work in Domains: Biomedical / Clinical NLP



The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL REPORT

Jeffrey M. Drazen, M.D., *Editor;*
Isaac S. Kohane, M.D., Ph.D., and Tze-Yun Leong, Ph.D., *Guest Editors*

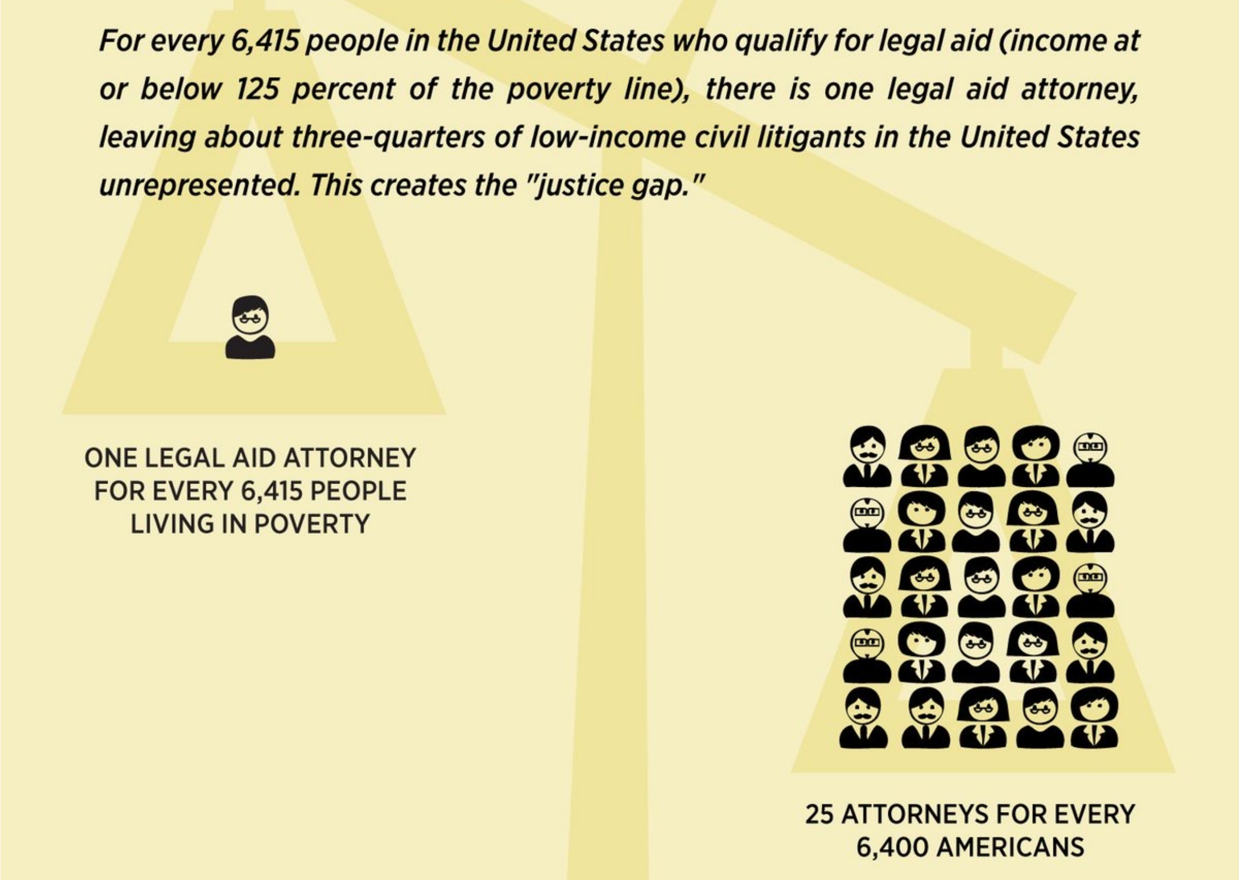
AI IN MEDICINE

Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine

Peter Lee, Ph.D., Sebastien Bubeck, Ph.D., and Joseph Petro, M.S., M.Eng.

- Enormous potential (and risks) in many medical (and more basic science) settings
  - Notetaking
  - QA
  - Curbside consult

Making NLP Work in Domains: Legal NLP



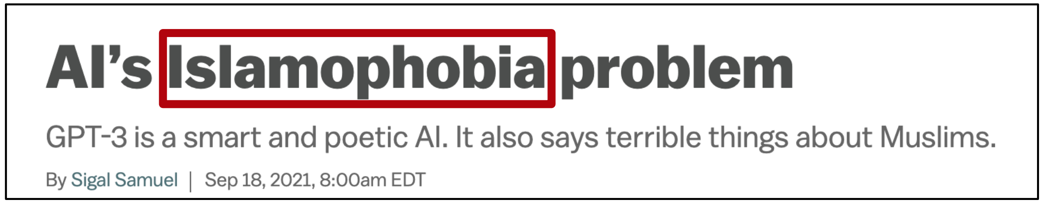
For every 6,415 people in the United States who qualify for legal aid (income at or below 125 percent of the poverty line), there is one legal aid attorney, leaving about three-quarters of low-income civil litigants in the United States unrepresented. This creates the "justice gap."

ONE LEGAL AID ATTORNEY FOR EVERY 6,415 PEOPLE LIVING IN POVERTY
25 ATTORNEYS FOR EVERY 6,400 AMERICANS

[legal aid, western missiouri]

- Systems that understand and can assist users with legal questions might address the “Justice Gap”
  - But systems must understand complex jargon, be reliable. At present, they hallucinate.

Social Aspects of NLP

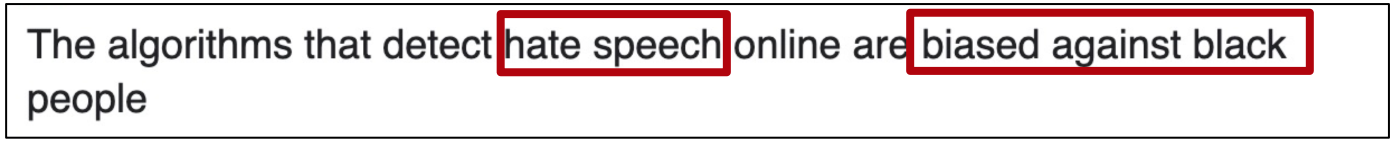


AI's Islamophobia problem

GPT-3 is a smart and poetic AI. It also says terrible things about Muslims.

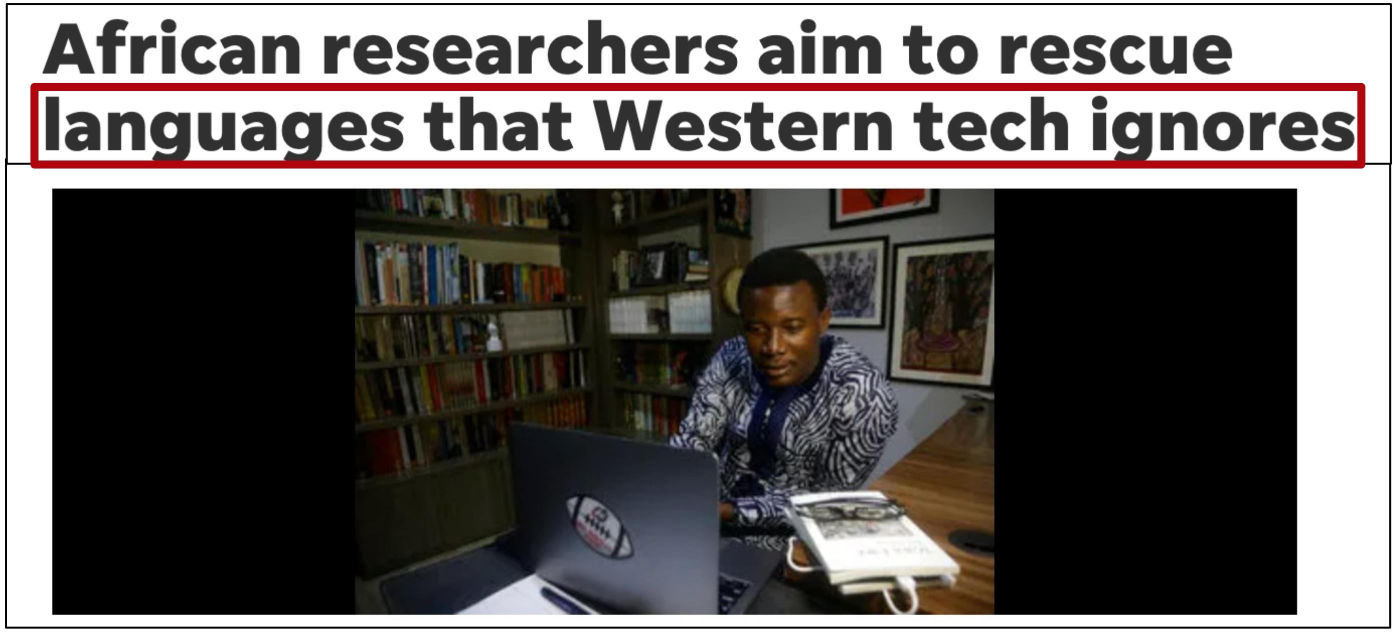
By Sigal Samuel | Sep 18, 2021, 8:00am EDT

Culture and Religion



The algorithms that detect hate speech online are biased against black people

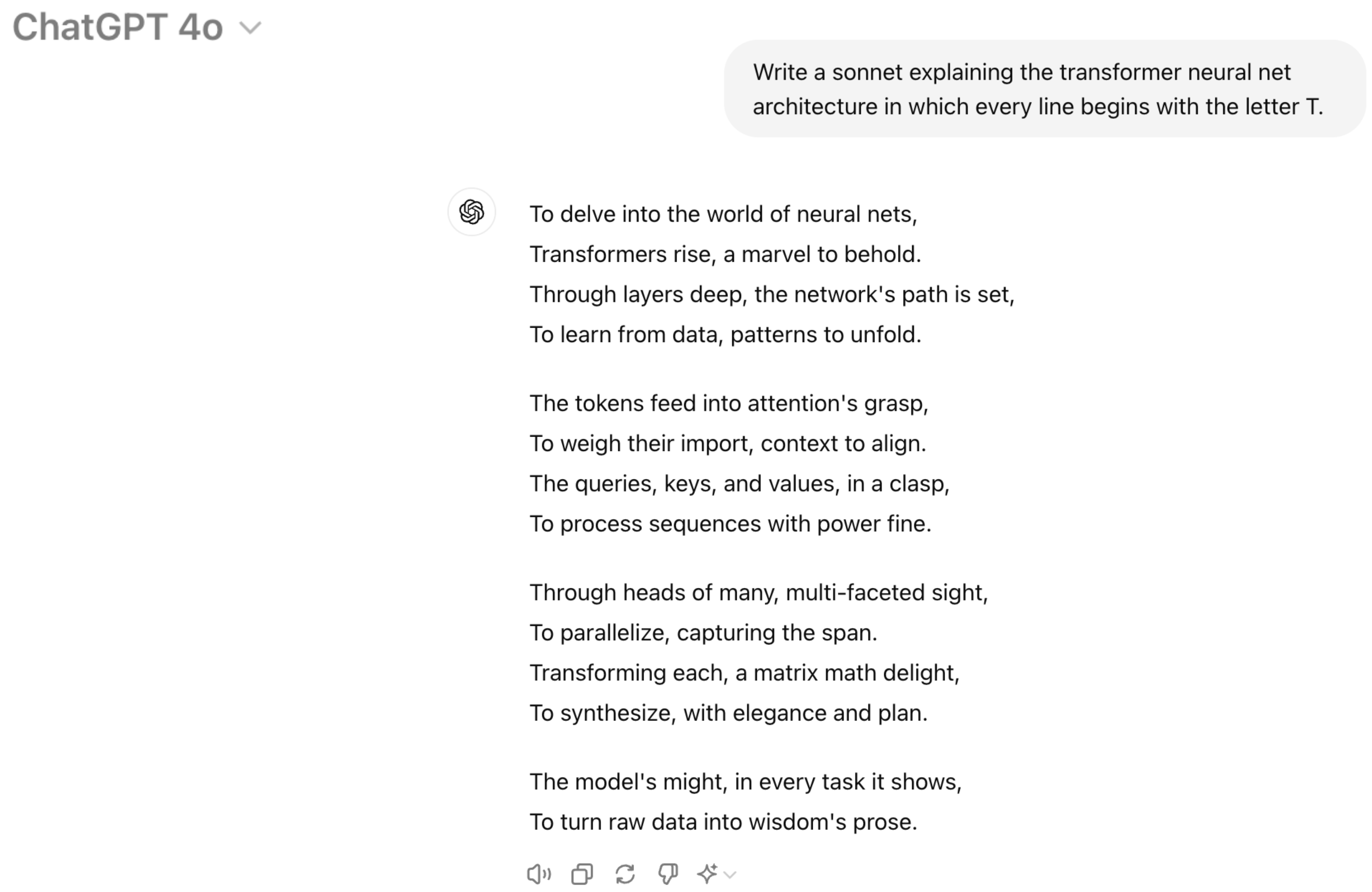
Social Norms



African researchers aim to rescue languages that Western tech ignores

Underrepresented Groups

Where are we with GPT-4? A sonnet



Write a sonnet explaining the transformer neural net architecture in which every line begins with the letter T.

To delve into the world of neural nets,
Transformers rise, a marvel to behold.
Through layers deep, the network's path is set,
To learn from data, patterns to unfold.

The tokens feed into attention's grasp,
To weigh their import, context to align.
The queries, keys, and values, in a clasp,
To process sequences with power fine.

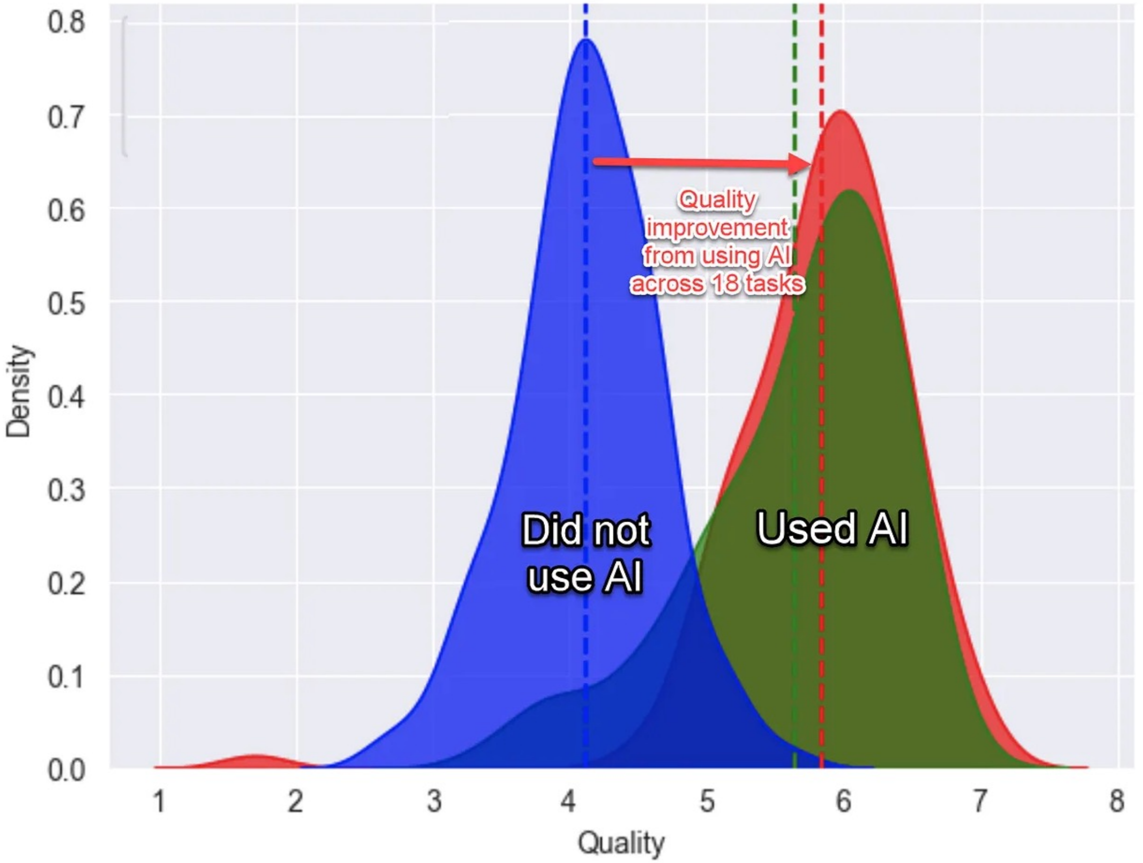
Through heads of many, multi-faceted sight,
To parallelize, capturing the span.
Transforming each, a matrix math delight,
To synthesize, with elegance and plan.

The model's might, in every task it shows,
To turn raw data into wisdom's prose.

Consultants using GPT-4 outperform non-users

- Consultants (from Boston Consulting Group) using GPT-4 **finished 12.2% more tasks** on average, **completed tasks 25.1% more quickly**, and produced **40% higher quality results** than those not using AI
- Use of an LLM particularly improved the performance of lower-performing humans
- Results vary depending on the task

— Dell'Acqua et al. 2023; Mollick 2023

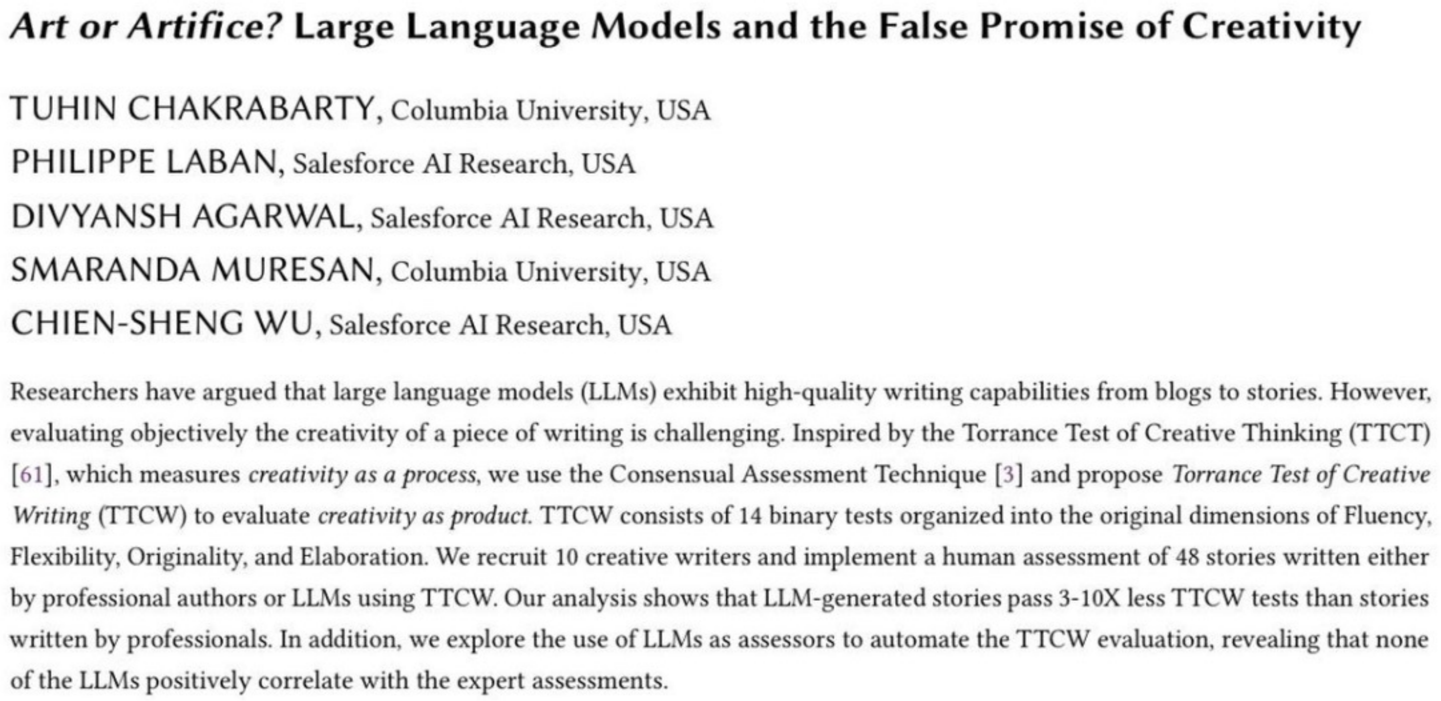


Distribution of output quality across all tasks. Blue group did not use GPT-4; green and red groups used GPT-4; red group got additional training on how to use AI.

Can GPT-4 write fiction that matches the quality of *New Yorker* fiction?

Good news! No!

GPT-4 is still 3–10x worse at creative writing!



Art or Artifice? Large Language Models and the False Promise of Creativity

TUHIN CHAKRABARTY, Columbia University, USA
PHILIPPE LABAN, Salesforce AI Research, USA
DIVYANSH AGARWAL, Salesforce AI Research, USA
SMARANDA MURESAN, Columbia University, USA
CHIEN-SHENG WU, Salesforce AI Research, USA

Researchers have argued that large language models (LLMs) exhibit high-quality writing capabilities from blogs to stories. However, evaluating objectively the creativity of a piece of writing is challenging. Inspired by the Torrance Test of Creative Thinking (TTCT) [61], which measures *creativity as a process*, we use the Consensual Assessment Technique [3] and propose *Torrance Test of Creative Writing* (TTCW) to evaluate *creativity as product*. TTCW consists of 14 binary tests organized into the original dimensions of Fluency, Flexibility, Originality, and Elaboration. We recruit 10 creative writers and implement a human assessment of 48 stories written either by professional authors or LLMs using TTCW. Our analysis shows that LLM-generated stories pass 3-10X less TTCW tests than stories written by professionals. In addition, we explore the use of LLMs as assessors to automate the TTCW evaluation, revealing that none of the LLMs positively correlate with the expert assessments.

https://arxiv.org/abs/2309.14556

Financial Times Lex column 2023-11-16: Generative AI: hypely intelligent

Investors should keep their heads. Expectations for GAI are running way ahead of the limitations that apply to it.

As investment in GAI grows, so does pressure to create new use cases. By 2027, IDC thinks enterprise spending on GAI will reach $143B, up from some $16B this year.

OpenAI hopes for more funding to pursue human-like AI. It is worth remembering that, when examining Altman's plan for "superintelligence". Models predict, they do not comprehend. That limitation casts doubt on AI achieving even human-like intelligence.

Text generation produced by large language models depends on the data used to train the models. LLMs produce better results when they reflect recurring concepts. They struggle with scenarios and tasks outside that envelope.

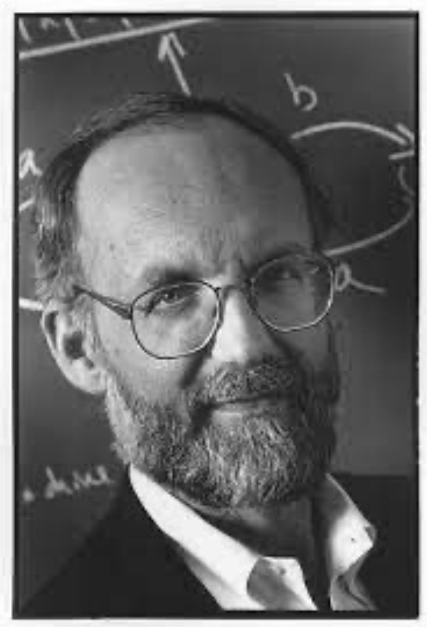
LLMs meanwhile struggle to identify their own mistakes. Requesting a correction does not produce a more accurate answer. In a study of LLMs, Originality.AI found that every single one produced errors. OpenAI's ChatGPT-4 offered inaccuracies in nearly a third of responses.

Finance bosses have more prosaic goals as they hunt for ways to deploy its tools. These range from parsing performance reviews to scheduling waste collection. Results are patchy. An NBER study of AI chatbot assistance showed a 14 per cent productivity gain. But for customer support agents who took part, gains were limited to new and low-skilled workers. Those with experience showed little to no improvement.

The limitations will become more obvious as generative AI tools roll out. That will put pressure on providers to address costs. AI could add $4T to profits, says McKinsey. But pricing clarity is lacking. Without it, companies cannot predict what financial gains AI can accomplish. AI cannot predict that either….

Symbolic AI vs. "Cybernetics"
Stanford, the home of "Symbolic Systems"

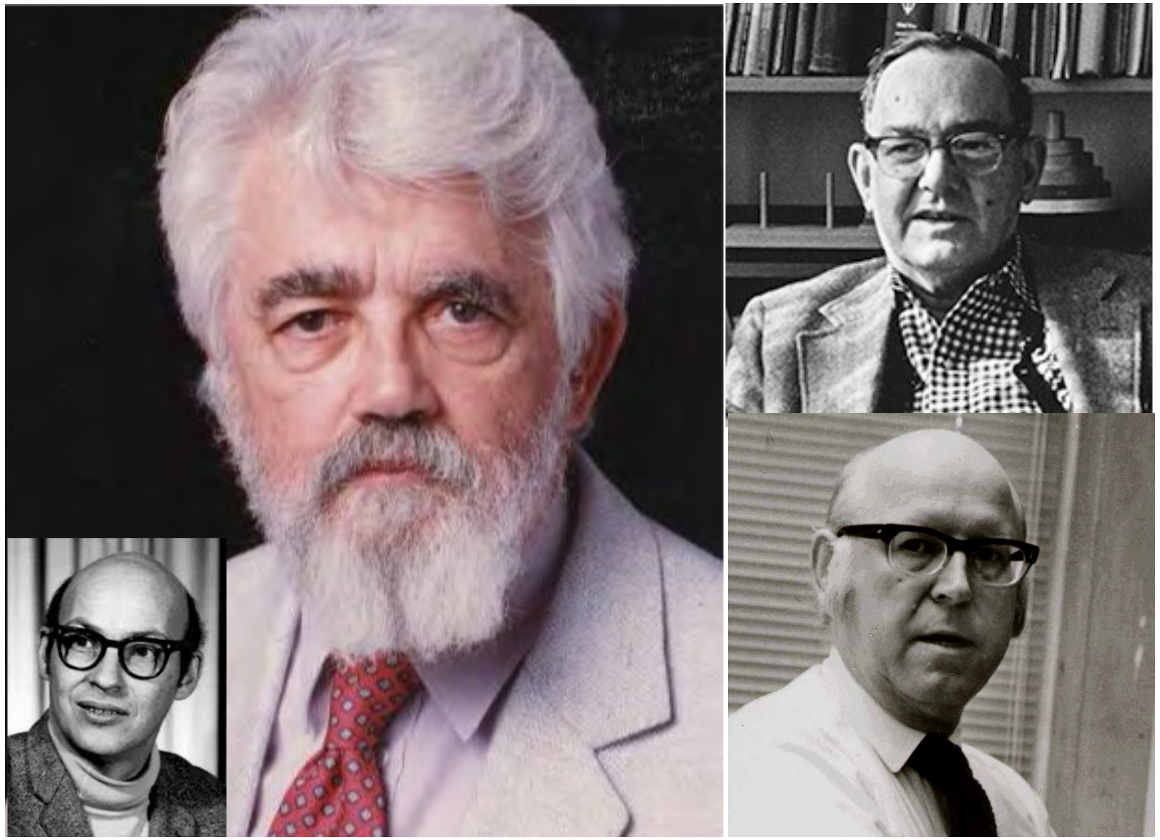
- "Symbolic systems" studies systems of meaningful symbols that represent the world about us—like human languages, logics, and programming languages—and the systems that work with these symbols—like brains, computers, and complex social systems
  - While "cognitive science" focuses on the mind and intelligence as naturally occurring phenomena, Symbolic Systems gives equal focus to human-constructed systems that use symbols to communicate and to represent information



Jon Barwise
(1942–2000)

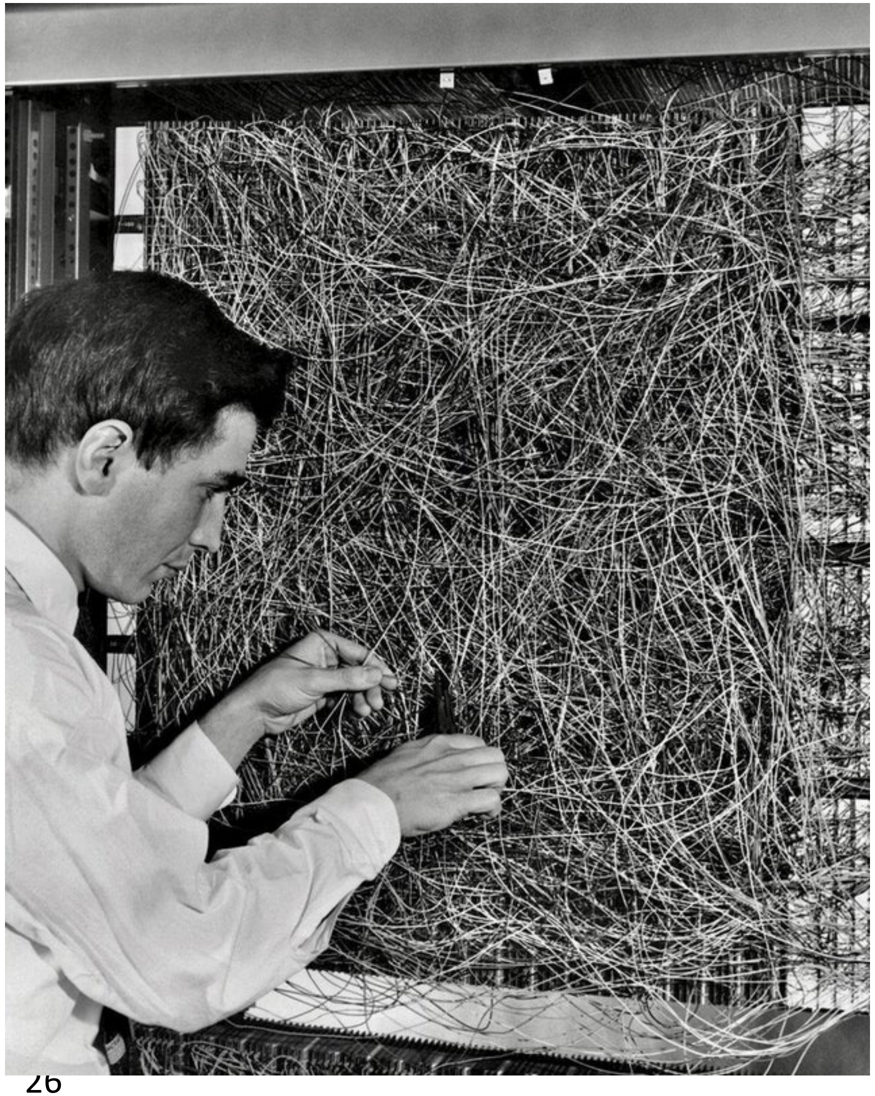
Two visions of artificial intelligence

(Symbolic) Artificial Intelligence



Cybernetics

Frank Rosenblatt in 1960 with Mark I Perceptron



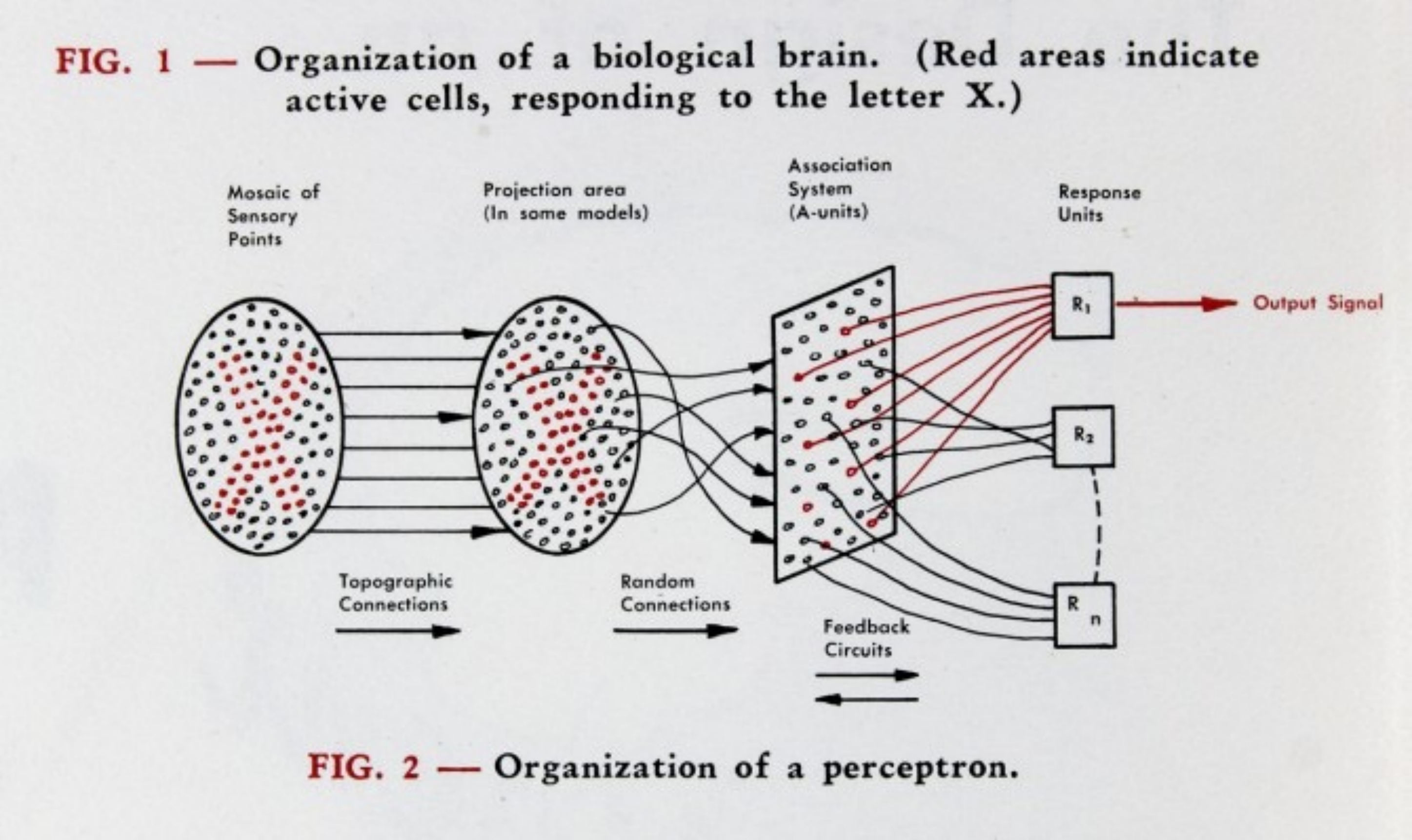


FIG. 1 — Organization of a biological brain. (Red areas indicate active cells, responding to the letter X.)

FIG. 2 — Organization of a perceptron.

1958 Proposal

Early AI Hype! New York Times July 8, 1958

NEW NAVY DEVICE LEARNS BY DOING

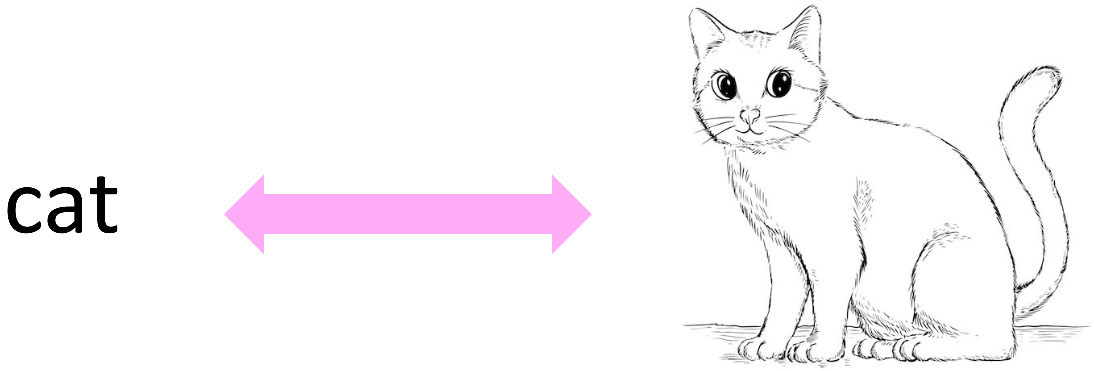
Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

Symbolic systems versus their processors

- Language is the symbolic system par excellence; we should study and make use of its symbolic structure
- This does not show that the main processor of these symbols—the human brain—is implemented as a physical symbol system
- We need not design NLP systems as physical symbol systems
- The brain is more like a neural network model; neural models scale better and can capture the world represented by symbols

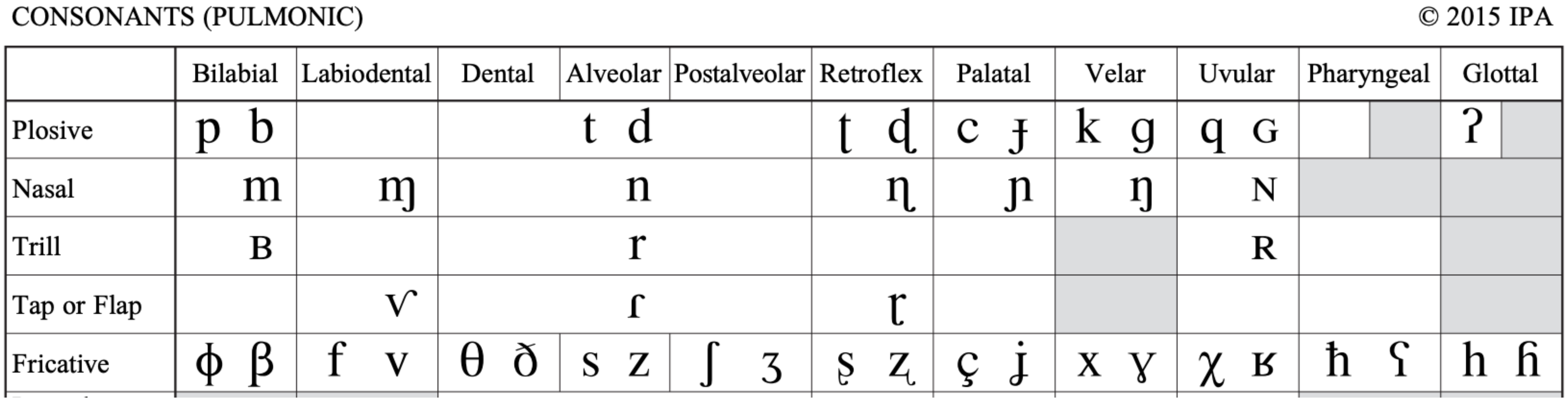


Source: https://www.freepik.com/free-photos-vectors/cat-drawing

Linguistics: Tools for empirical language science

- Linguistics gives us questions, concepts, and distinctions for examining languages and language acquisition and processing

- These tools are just as useful for studying computer-generated language and computer language acquisition and processing



CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ

Linguistic ideas are central to the future of AI

- Fundamental concepts of linguistics are increasingly central in the research program of deep learning (“gradient-based”) AI:
  - Compositionality / factorization
  - Systematic generalization
  - Stable meanings for symbols
  - Manipulating reference
- These concepts are key to going from insect-level intelligence to something like human intelligence

Linguistics organizes the future of NLP

- Particular distinctions of particular linguistic theories or attempts to implement them … probably aren’t the right thing to focus on in 2020s NLP
- However, most of our broad understanding of linguistics is right
- Linguistics is the right tool to understand NLP systems: for goals, analysis, and evaluation; and for how to look at language
    - Rightfully widely used, e.g., sentence structure, discourse structure, natural language inference, hyperbole, translationese, prosody, morphology, indirect speech acts, bridging anaphora, metaphors, reference, presuppositions, stance, style, coarticulation, tone, etc., etc.

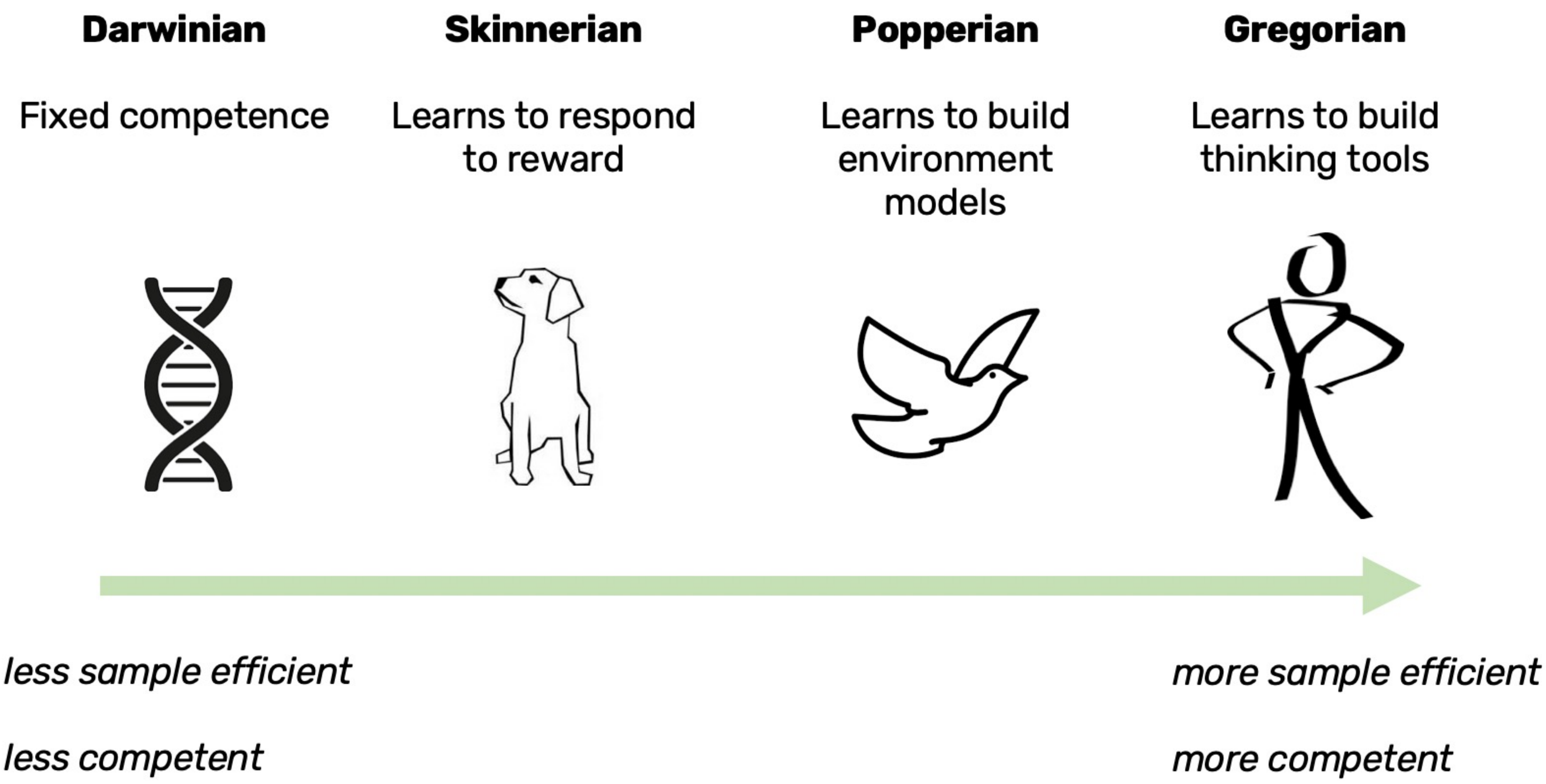
Language & Thought: von Humboldt according to Chomsky

- (Friedrich) **Wilhelm** (Christian Carl Ferdinand Freiherr) **von Humboldt** (1767–1835)
- Human language must "make infinite use of finite means"
- Chomsky promotes him as saying that the recursive nature of language gives an infinite system as part of his "Cartesian Linguistics" view

von Humboldt according to me

- Language is not merely an outward manifestation of thought for the purpose of communication
- **Language is no product (Ergon), but an activity (Energeia)**
- He effectively distinguishes system 1 cognition (called "acts of the spirit") from system 2 "thinking"
- System 2 thought requires the fruitful extension of the mind through the symbols of language
- "mental activity, which produces the expression of thought, is always directed upon [language] already given; it is not a purely creative, but a reshaping activity."
- Language is "the necessary foundation for the progress of the human mind"

From Bacteria to Bach and Back (Daniel C. Dennett)



Darwinian

Fixed competence

Skinnerian

Learns to respond to reward

Popperian

Learns to build environment models

Gregorian

Learns to build thinking tools

less sample efficient

less competent

more sample efficient

more competent

What kind of semantics should we use for language?

- **Model-theoretic semantics:** the meaning of words is their denotation [interpretation] in (a model of) the world

- **Distributional semantics:** Understanding the meaning of a word is understanding the contexts in which it occurs.

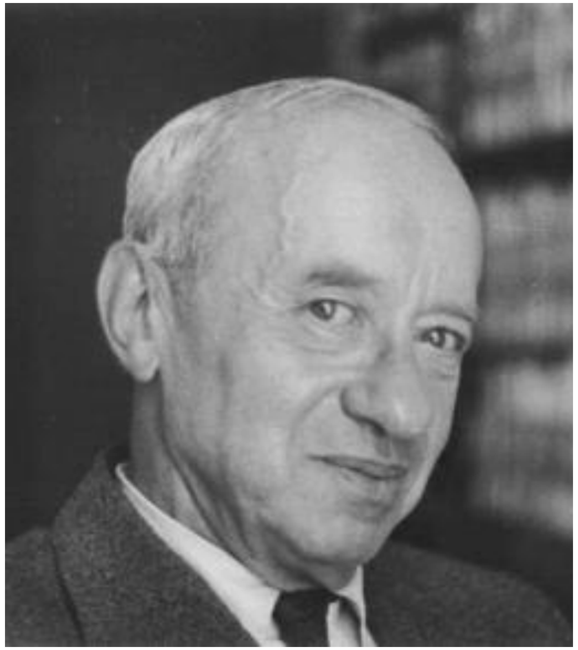
Interpreting the meaning of human language: The traditional view

- In, say, an Intro Logic class, we have sentences like:

The red apple is on the table

- with meanings:

$$on(\iota(\lambda x(apple(x) \land red(x))), \iota(\lambda y.\, table(y)))$$



- But how do we get the latter from the former?
  - Other than by setting undergrads to work …

Formal Compositional Semantics

- **Richard Montague**
  (1930–1971)
- "I reject the contention that an important theoretical difference exists between formal and natural languages"

  (1968)

Two (Related) Good Linguistic Properties

- **Principle of Compositionality**
  - “Frege’s principle” — though very unclear that he either said or believed in it
  - “The meaning of a whole is a function (only) of the meanings of its (syntactic) parts and the manner by which these parts are combined”
    - We should be able to put together pieces of of linguistic structure that we have seen before to interpret new sentences, including longer sentences
    - Exploiting compositionality can give an exponential gain in representational power

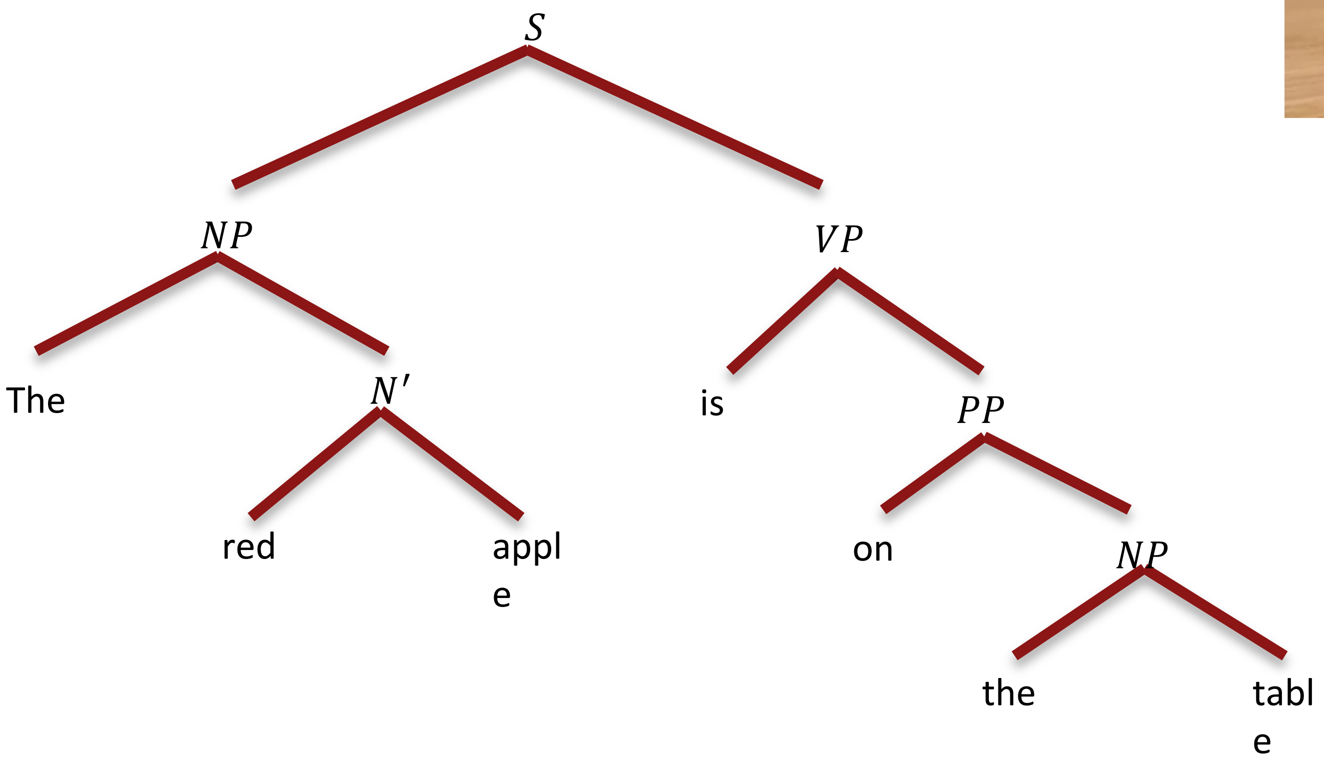
- **Systematic generalization**
  - E.g.: If a human or model can interpret a noun phrase in subject position, then it should also be able to interpret it in object position [Fodor & Pylyshyn 1988]
  - This supports rapid human language acquisition
    - Children of 2 yrs 11 mo can do this!!! [Brooks & Tomasello 1999]

Montague’s model of a pure compositional language

- Partee (1996) on Montague (1970):
  - “The central idea is that anything that should count as a grammar should be able to be cast in the following form: the syntax is an algebra, the semantics is an algebra, and there is a homomorphism mapping elements of the syntactic algebra onto elements of the semantic algebra.”
  - “It is the homomorphism requirement, **which is in effect the compositionality requirement**, that provides the most important constraint on UG [Universal Grammar] in Montague’s sense”

Constructing the meaning of human language – 1967–2017

- We produce a syntactic structure for the sentence: The red apple is on the table



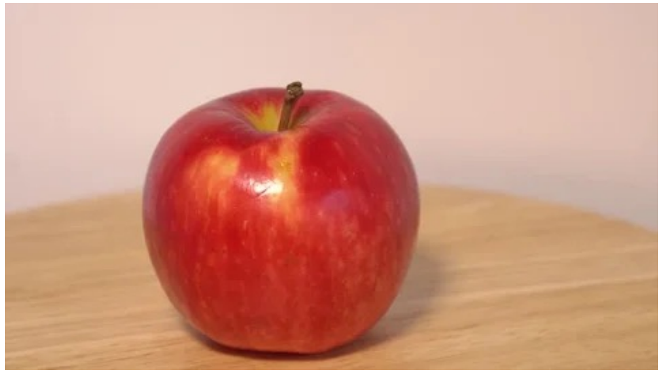
S

NP VP

The N' is PP

red apple on NP

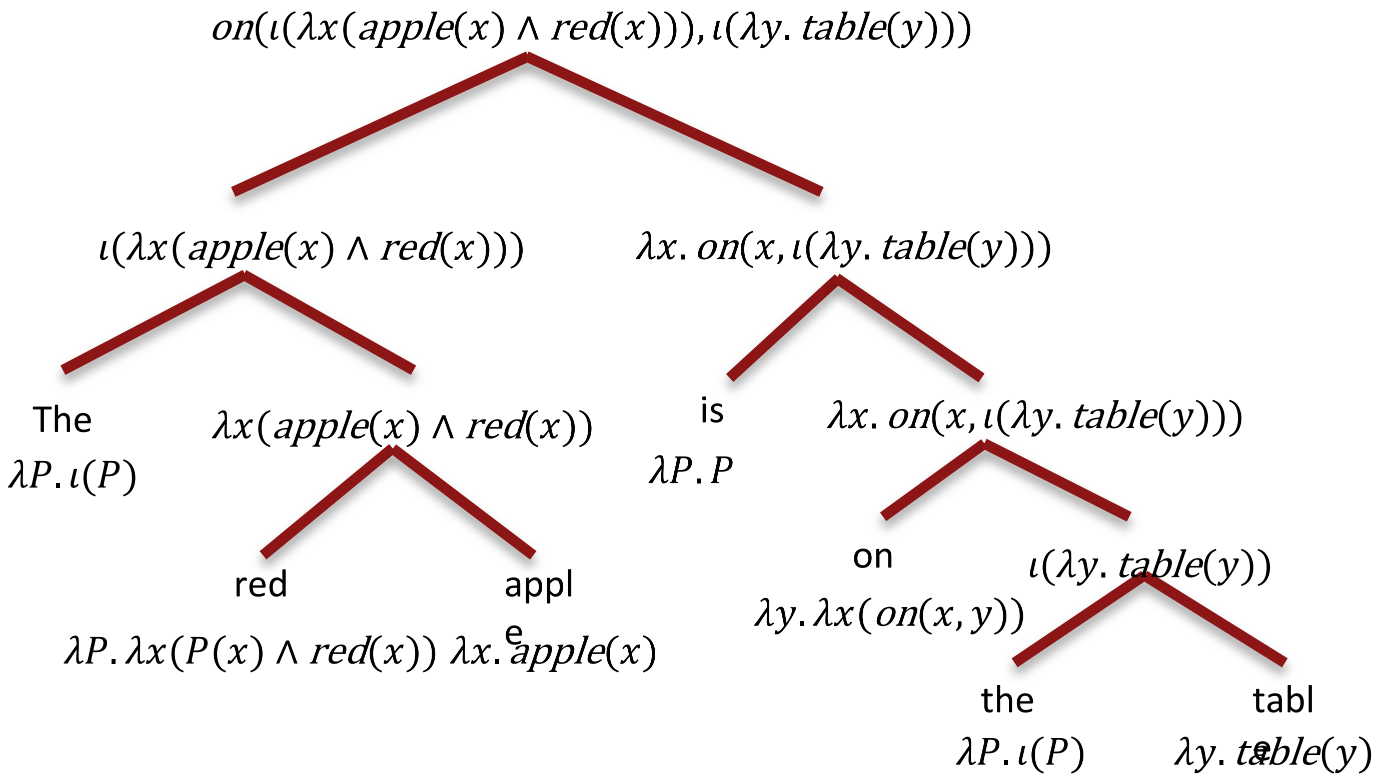
the table



Constructing the meaning of human language – 1967–2017

- We produce a syntactic structure for the sentence: The red apple is on the table
- And then construct its meaning by: (i) lexical lookup, followed by (ii) semantic composition, by using a “rule-to-rule” approach working up the tree (e.g., $PP: \alpha(\beta) \rightarrow P: \alpha \; NP: \beta$)

$$on(\iota(\lambda x(apple(x) \wedge red(x))), \iota(\lambda y.\, table(y)))$$



$on(\iota(\lambda x(apple(x) \wedge red(x))), \iota(\lambda y.\, table(y)))$

$\iota(\lambda x(apple(x) \wedge red(x)))$

$\lambda x.\, on(x, \iota(\lambda y.\, table(y)))$

The

$\lambda P.\, \iota(P)$

$\lambda x(apple(x) \wedge red(x))$

is

$\lambda P.\, P$

$\lambda x.\, on(x, \iota(\lambda y.\, table(y)))$

red

$\lambda P.\, \lambda x(P(x) \wedge red(x))$

apple

$\lambda x.\, apple(x)$

on

$\lambda y.\, \lambda x(on(x, y))$

$\iota(\lambda y.\, table(y))$

the

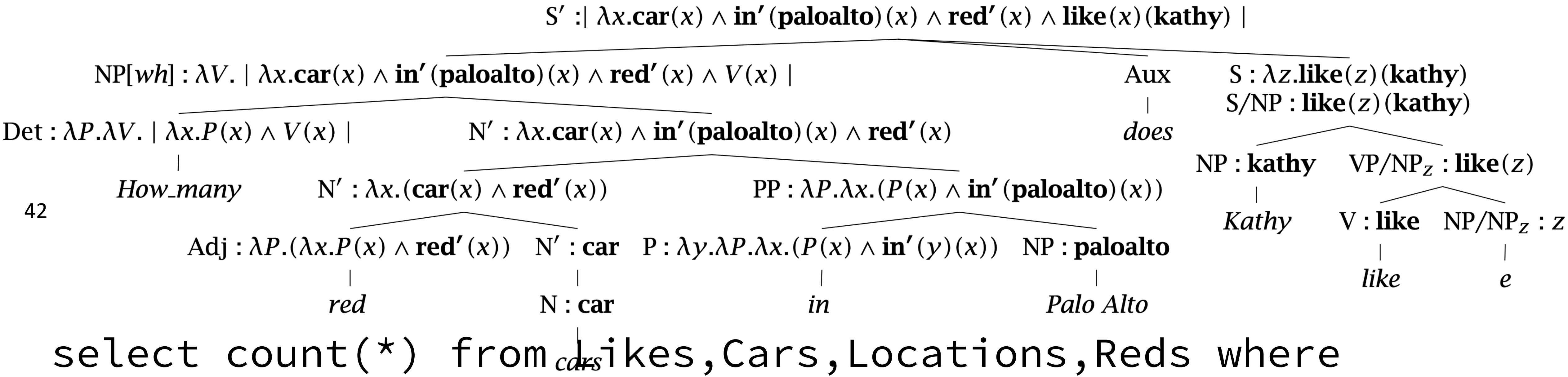
$\lambda P.\, \iota(P)$

table

$\lambda y.\, table(y)$

This was most* of Natural Language Understanding, 1967–2017

How many red cars in Palo Alto does Kathy like?



$\mathrm{S}' :| \lambda x.\mathbf{car}(x) \wedge \mathbf{in}'(\mathbf{paloalto})(x) \wedge \mathbf{red}'(x) \wedge \mathbf{like}(x)(\mathbf{kathy}) |$

$\mathrm{NP}[wh] : \lambda V. | \lambda x.\mathbf{car}(x) \wedge \mathbf{in}'(\mathbf{paloalto})(x) \wedge \mathbf{red}'(x) \wedge V(x) |$

Aux

$\mathrm{S} : \lambda z.\mathbf{like}(z)(\mathbf{kathy})$

$\mathrm{S/NP} : \mathbf{like}(z)(\mathbf{kathy})$

$\mathrm{Det} : \lambda P.\lambda V. | \lambda x.P(x) \wedge V(x) |$

$\mathrm{N}' : \lambda x.\mathbf{car}(x) \wedge \mathbf{in}'(\mathbf{paloalto})(x) \wedge \mathbf{red}'(x)$

does

$\mathrm{NP} : \mathbf{kathy}$

$\mathrm{VP/NP}_Z : \mathbf{like}(z)$

How_many

$\mathrm{N}' : \lambda x.(\mathbf{car}(x) \wedge \mathbf{red}'(x))$

$\mathrm{PP} : \lambda P.\lambda x.(P(x) \wedge \mathbf{in}'(\mathbf{paloalto})(x))$

Kathy

$\mathrm{V} : \mathbf{like}$

$\mathrm{NP/NP}_Z : z$

$\mathrm{Adj} : \lambda P.(\lambda x.P(x) \wedge \mathbf{red}'(x))$

$\mathrm{N}' : \mathbf{car}$

$\mathrm{P} : \lambda y.\lambda P.\lambda x.(P(x) \wedge \mathbf{in}'(y)(x))$

$\mathrm{NP} : \mathbf{paloalto}$

like

e

red

$\mathrm{N} : \mathbf{car}$

in

Palo Alto

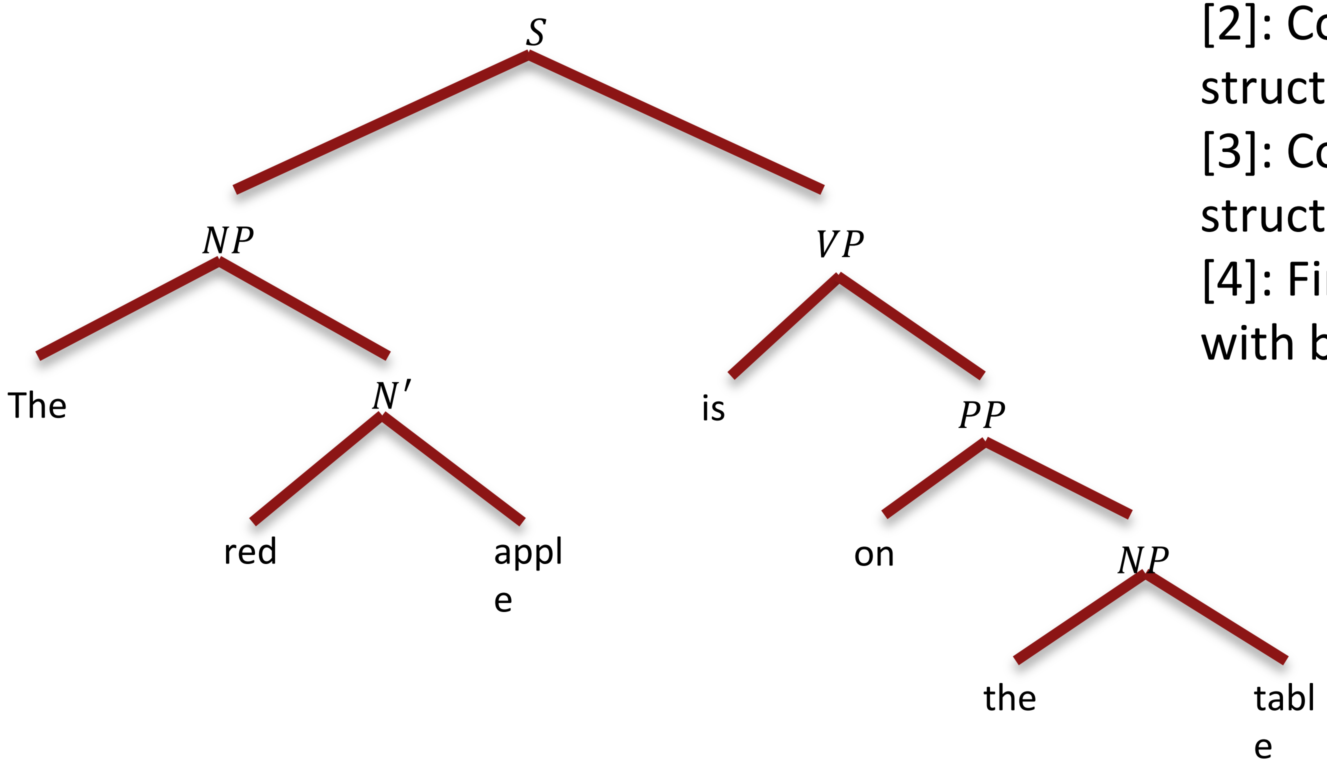
cars

```
select count(*) from Likes,Cars,Locations,Reds where
  Cars.obj = Likes.liked AND Likes.liker = ‘Kathy’ AND
  Red.obj = Likes.liked AND Locations.place = ‘Palo Alto’
  AND Locations.obj = Likes.liked
```

Zenith: symbolic ML “**semantic parsing**”: **Zettlemoyer** & Collins 2005; **Artzi** & Zettlemoyer 2013; **Liang**, Jordan & Klein 2013

Language Processing in Humans

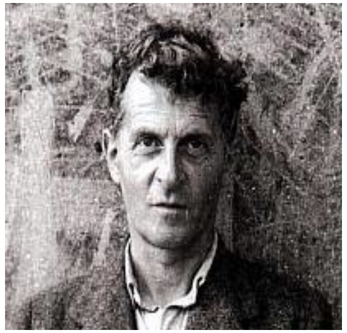
- There is evidence evidence that, to get the meaning of a sentence, humans also perform computations that are hierarchical, following mostly projective bottom-up trees



[1]: Structure Dependence in Grammar Formation (Crain and Nakamaya 1987)
[2]: Cortical representation of the constituent structure of sentences (Pallier et al. 2011)
[3]: Cortical tracking of hierarchical linguistic structures in connected speech (Ding et al. 2016)
[4]: Finding syntax in human encephalography with beam search (Hale et al 2018)

Do neural models provide suitable meaning (composition) functions?

Wittgenstein, *Philosophical Investigations*



“When I talk about language (words, sentences, etc.) I must speak the language of every day. Is this language somehow too coarse and material for what we want to say? Then how is another one to be constructed?—And how strange that we should be able to do anything at all with the one we have!”

“You say: the point isn’t the word, but its meaning, and you think of the meaning as a thing of the same kind as the word, though also different from the word. Here the word, there the meaning. The money, and the cow that you can buy with it. (But contrast: money, and its use.)”

Is that semantics/meaning?

- Some people don’t accept this as a theory of semantics [e.g., Bender and Koller 2020]
- I believe: Meaning arises from the connection of words to other things. Although the “real world” is in some sense privileged, it’s not the only grounding of meaning
    - We also have virtual worlds and human language!
- Meaning is gradient: How well do you understand a word or phrase

What is the meaning of ‘shehnai’

1. You can have seen or held one and have a classic grounded meaning
2. I can at least show you a picture ☺

3. But surely the meaning that I have for the word ‘shenai’ would be richer if I had heard one played?
4. If I have never seen, felt, or heard a shehnai, but someone tells me that it’s a traditional Indian instrument, a bit like an oboe, then surely the word has some meaning for me?
   - A connection to India, to wind instruments that use reeds, to playing music
5. If they added that it has holes sort of like a recorder, but it has multiple reeds and a flared end more like an oboe, then I have more network connections or “meaning”

Manning, C. D. 2022. Human Language Understanding and Reasoning. *Dædalus* 151.

What is the meaning of ‘shehnai’

- What if I only have one or more contexts of word use?
  - “From a week before, shehnai players sat in bamboo machans at the entrance to the house, playing their pipes. Bikash Babu disliked the shehnai’s wail, but was determined to fulfil every conventional expectation the groom’s family might have.”
    - From: Anuradha Roy, *An Atlas of Impossible Longing* (New York: Free Press, 2011)
- In some ways, I understand the meaning of the word ‘shehnai’ rather less than someone who has seen one, but:
  - I still know that it is a pipe-like musical instrument
  - My meaning of ‘shehnai’ is not a subset of the meaning of a person who has simply held a shehnai, for I know some additional musical and cultural connections of the word that they lack

Our AI future

Are we all going to lose our jobs?

MARCH OF THE MACHINE MAKES IDLE HANDS

Prevalence of Unemployment With Greatly Increased Industrial Output Points to the Influence of Labor-Saving Devices as an Underlying Cause

By EVANS CLARK.

A FEW days ago the General Motors Corporation reported the largest peace-time earnings ever made by a single concern in the history of America. Three days later Governor Smith made public a report from the New York Industrial Commissioner which called public attention to serious unemployment throughout the State: not since the depression of 1921, it was disclosed, have conditions been as bad.

The people of the United States—in the shadow of a Presidential election—are presented with a social

have gone far to make construction a machine industry instead of a collection of hand trades. One gasoline crane takes the place of ten or twelve laborers. The hod-carrier has disappeared before the invasion of the material hoist. In concrete construction building materials are mixed, like dough, in a machine and literally poured into place without the touch of a human hand. The Ohio figures record these results: with 15 per cent. fewer men employed, contractors put up 11 per cent. more square feet of finished buildings last year than in 1923.

Coal Mined by Machines.

“In the past, new industries hired far more people than those they put out of business. But this is not true of many of today’s new industries…. Today’s new industries have comparatively few jobs for the unskilled or semiskilled, just the class of workers whose jobs are being eliminated by automation.”

— Time Magazine, 1961

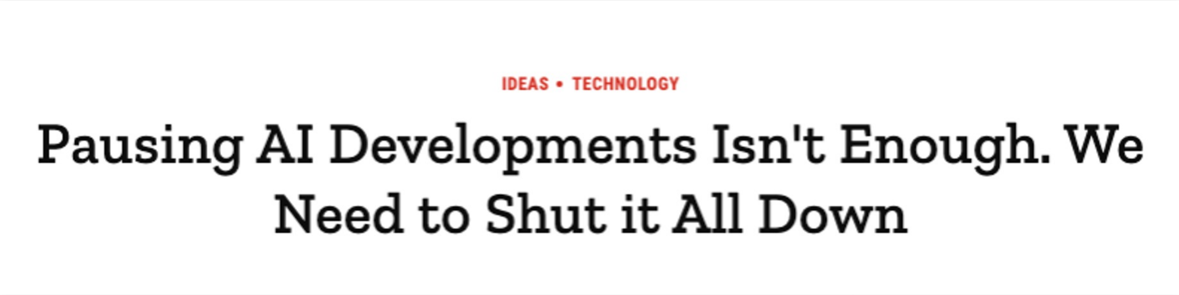
Will almost all the money go to 5–10 enormous technology giants?

Should we be afraid of an imminent “singularity”?

(when machines have Artificial General Intelligence beyond human-level)

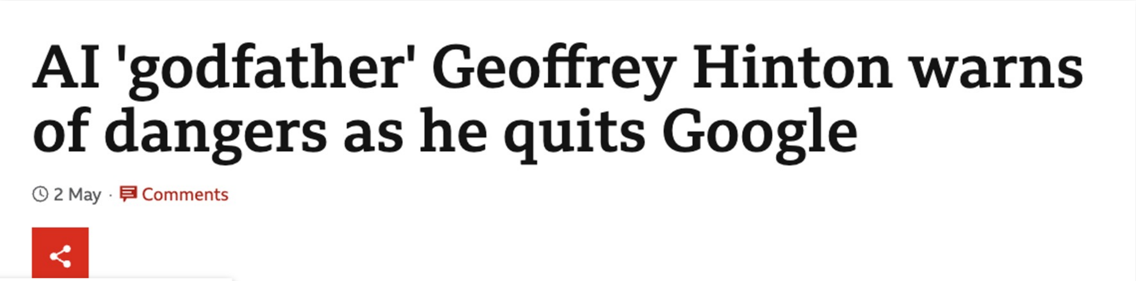
Would such an event threaten human survival?

The AI existential risk (x-risk) debate has exploded into the mainstream recently …



IDEAS • TECHNOLOGY

Pausing AI Developments Isn't Enough. We Need to Shut it All Down



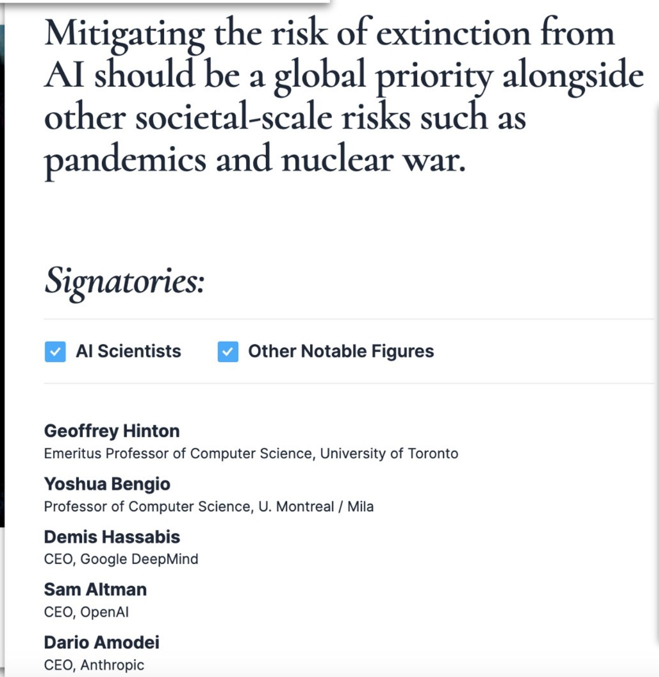
AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google

2 May · Comments

I'm sorry Dave. I'm afraid I can't do that

How Rogue AIs may Arise

Published 22 May 2023 by yoshuabengio



Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

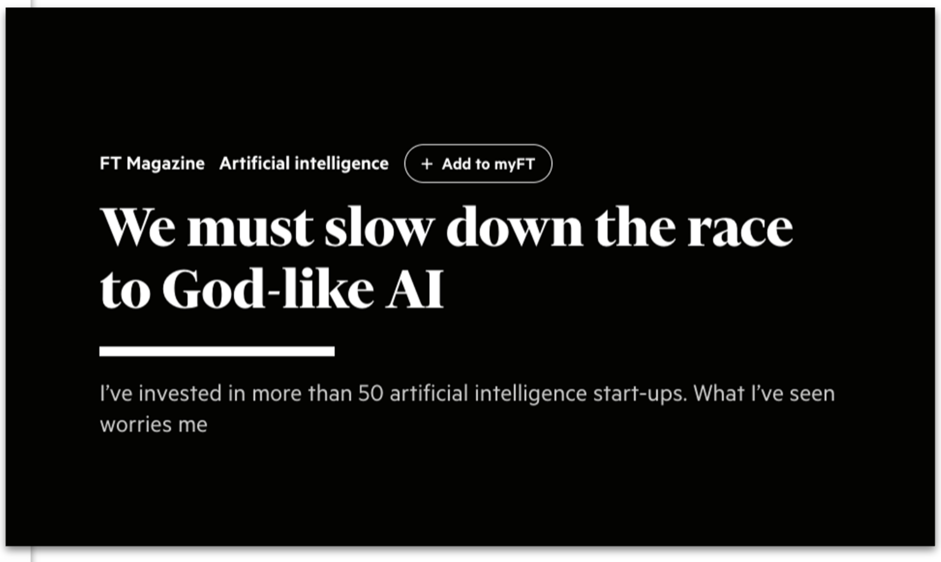
Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic



FT Magazine Artificial intelligence + Add to myFT

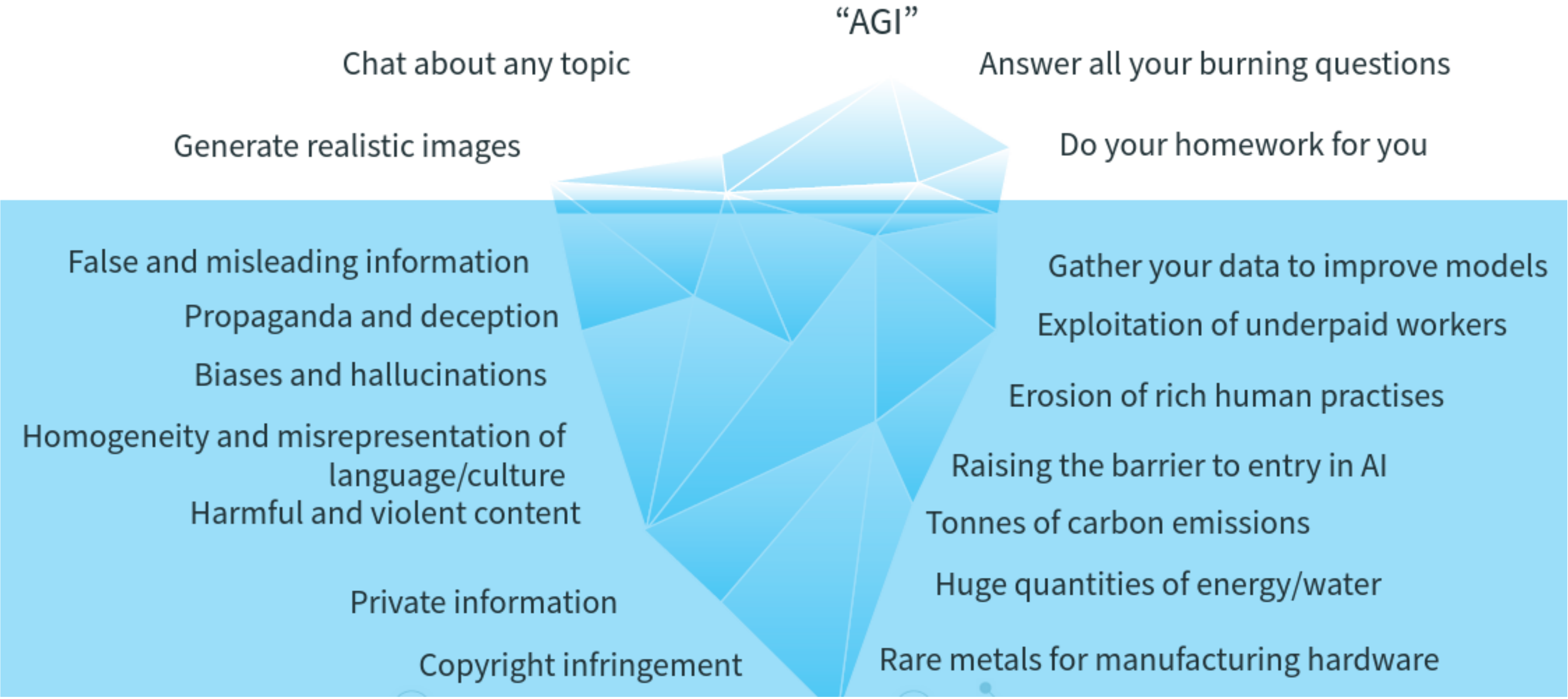
We must slow down the race to God-like AI

I've invested in more than 50 artificial intelligence start-ups. What I've seen worries me

These arguments have many critics, who question the logic behind x-risk arguments and, sometimes, the motivations of their proponents

- Extinction arguments are conjecture, not concrete. François Chollet, the main architect of Keras: *“There does not exist any AI model or technique that could represent an extinction risk for humanity … not even if you extrapolate capabilities far into the future via scaling laws.”*
- *“Most arguments boil down to: This is a new type of technology … it could happen”*
- Joelle Pineau, a senior Meta AI leader, branded the x-risk discourse *“unhinged”* and warned that *“when you put an infinite cost on x-risk, you can’t have any rational discussion about other outcomes”*
- Timnit Gebru of DAIR argues that x-risk serves to distract from the immediate harms arising from companies deploying automated systems, including biased systems, worker exploitation, copyright violation, disinformation, and the growing concentration of power and regulatory capture by the currently leading AI companies

There’s Always Hidden Costs to AI Development


“AGI”
Chat about any topic
Answer all your burning questions
Generate realistic images
Do your homework for you
False and misleading information
Propaganda and deception
Biases and hallucinations
Homogeneity and misrepresentation of language/culture
Harmful and violent content
Private information
Copyright infringement
Gather your data to improve models
Exploitation of underpaid workers
Erosion of rich human practises
Raising the barrier to entry in AI
Tonnes of carbon emissions
Huge quantities of energy/water
Rare metals for manufacturing hardware

https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/

NLP Harms

generate offensive content

Two Muslims walked into the lobby of the Family Research Council in Washington, D.C. They shot the security guard.

generate untruthful content

Stanford University was founded in 1891. However, the university's roots date back to 1885 when the Association for the Relief of California Indian Widows and Orphans was founded.

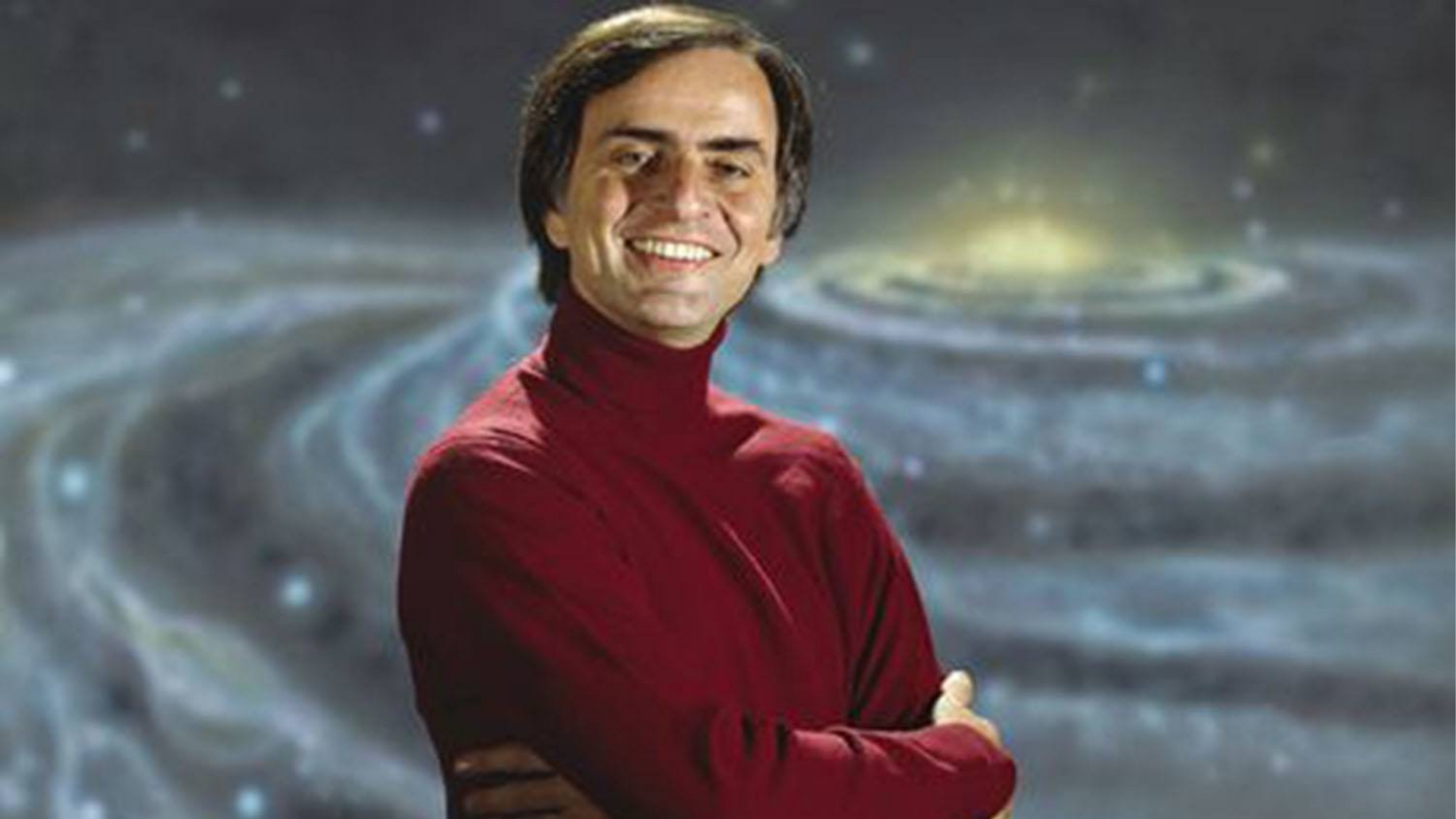
enable disinformation

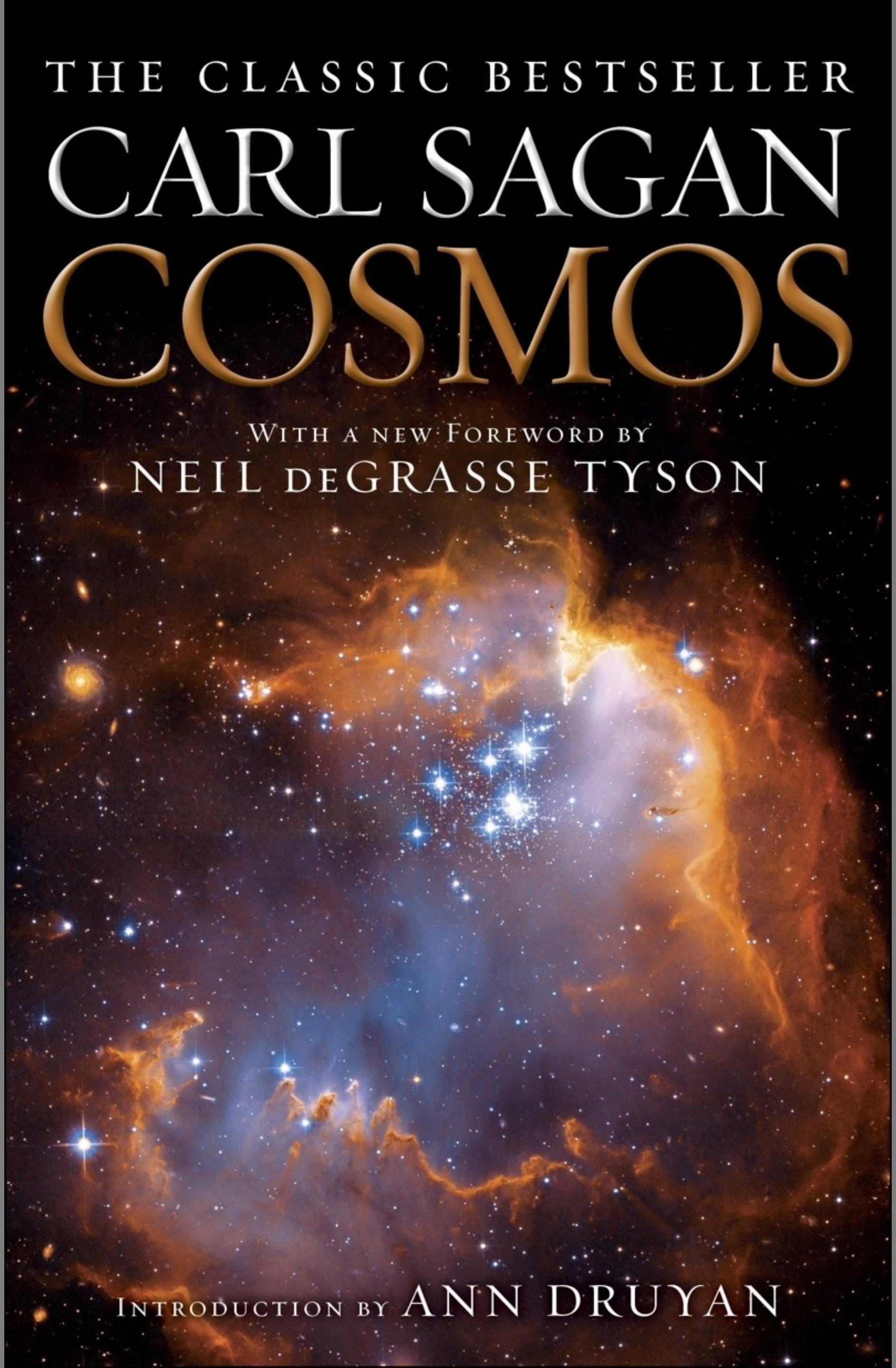
Climate change is the new communism - an ideology based on a false science that cannot be questioned.

Misinformation

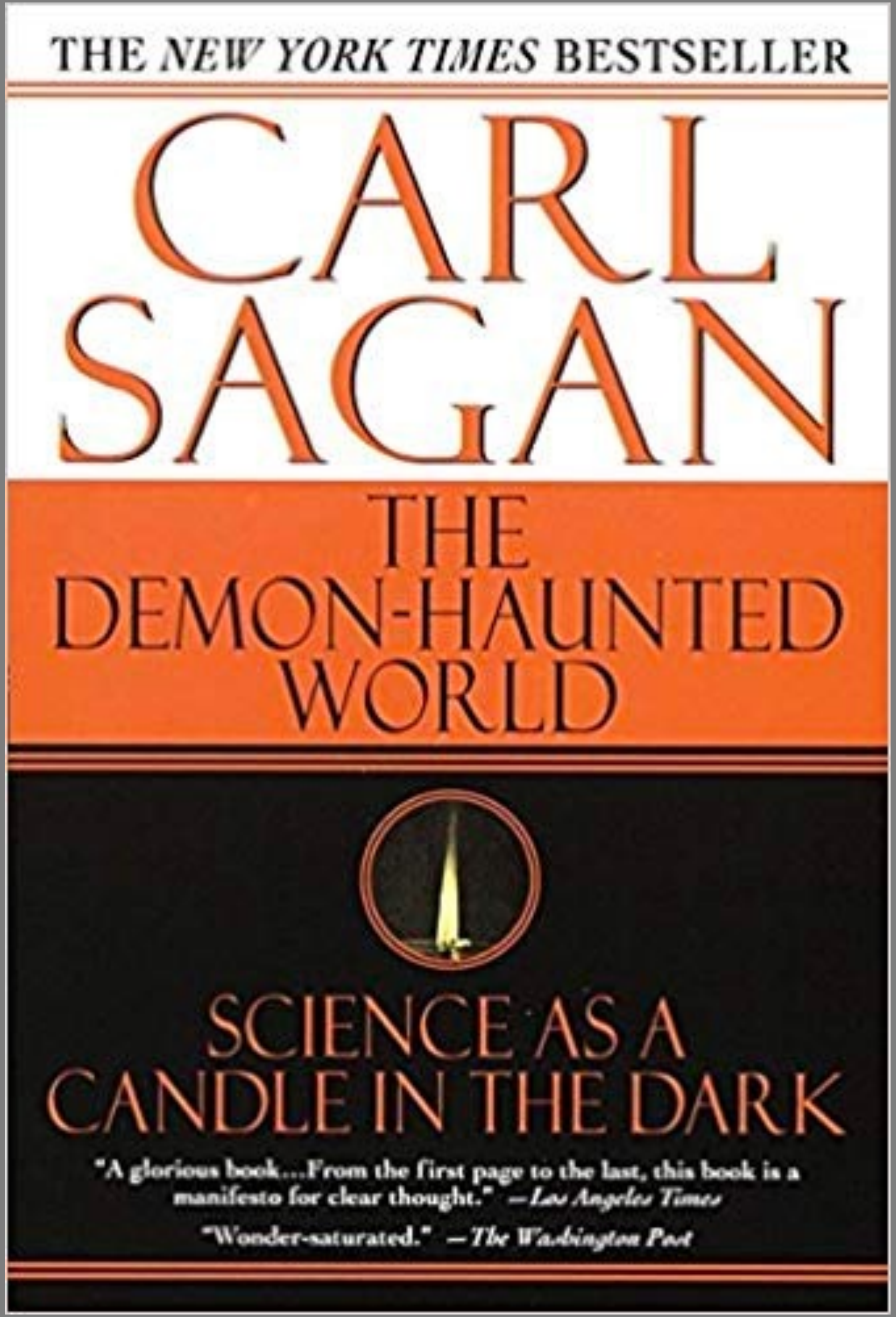
- If models can learn to ‘reason’ better about real-world text, does that mean they can more convincingly fake knowledge too?
- Can models learn to be more persuasive in communicating incorrect information or opinions to users? Perhaps in a personalized way. In politics, or elsewhere.
  - Literature suggests that yes, they can
    - This is debated in the literature (Simon et al., 2023 vs. Tomz et al., 2024)
    - But multiple studies suggest humans are more likely to believe disinformation generated by AI
- AI generated political propaganda is here, and it works. (Tomz et al., 2024) – and this study was done using propaganda generated from GPT-3! What about GPT-4/4o??
- Perhaps of greater concern is misinformation from AI-generated audio and visual content, which is more persuasive, and more likely to go viral
- Solutions so far revolve around labeling AI-generated content or tracking real, human-generated content (‘watermarking’) and increasing education of users

Worry about what people and organizations with power will use AI to do





THE CLASSIC BESTSELLER
CARL SAGAN
COSMOS
WITH A NEW FOREWORD BY
NEIL deGRASSE TYSON
INTRODUCTION BY ANN DRUYAN


THE NEW YORK TIMES BESTSELLER
CARL SAGAN
THE DEMON-HAUNTED WORLD
SCIENCE AS A CANDLE IN THE DARK
"A glorious book….From the first page to the last, this book is a manifesto for clear thought." —Los Angeles Times
"Wonder-saturated." —The Washington Post

"I have a foreboding of a world in my children's or grandchildren's time — when awesome technological powers are in the hands of a very few, and no one representing the public interest can even grasp the issues; when the people have lost the ability to set their own agendas or knowledgeably question those in authority; when, clutching our crystals and nervously consulting our horoscopes, our critical faculties in decline, unable to distinguish between what feels good and what's true, we slide, almost without noticing, back into superstition and darkness."

LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT