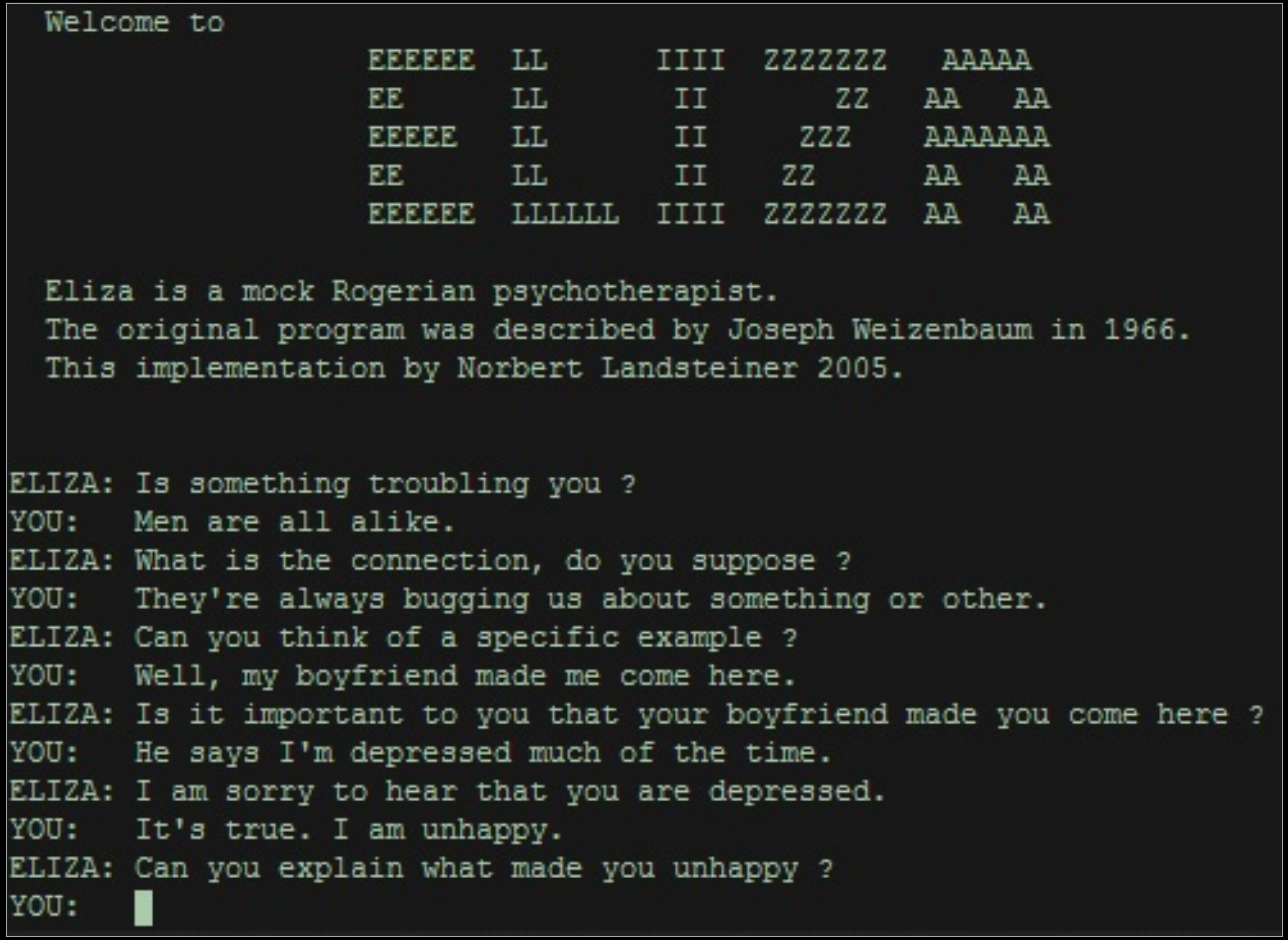
Deep *Understanding*: The Next Challenge for AI

Presented by: Gary Marcus
Contributors: Mohamed R. Amer, Dylan Bourgeois

Deep understanding vs shallow understanding

Shallow understanding, c. 1965

- Responding (frequently) in behaviorally-appropriate ways, without really getting the overall picture



Early example of AI system with shallow understanding (1965)

Shallow understanding, c. 2014

Marcus: Do you read *The New Yorker*?

Goostman: I read a lot of books … So many—I don't even remember which ones.

Marcus: You remind me of Sarah Palin.

Goostman: If I'm not mistaken, Sarah is a robot, just as many other "people." We must destroy the plans of these talking trash cans!

Eugene Goostman, 2014

- based on keyword matches, etc
- amusing but superficial
- doesn't represent real progress towards general intelligence

Shallow understanding, c. 2019

Completion

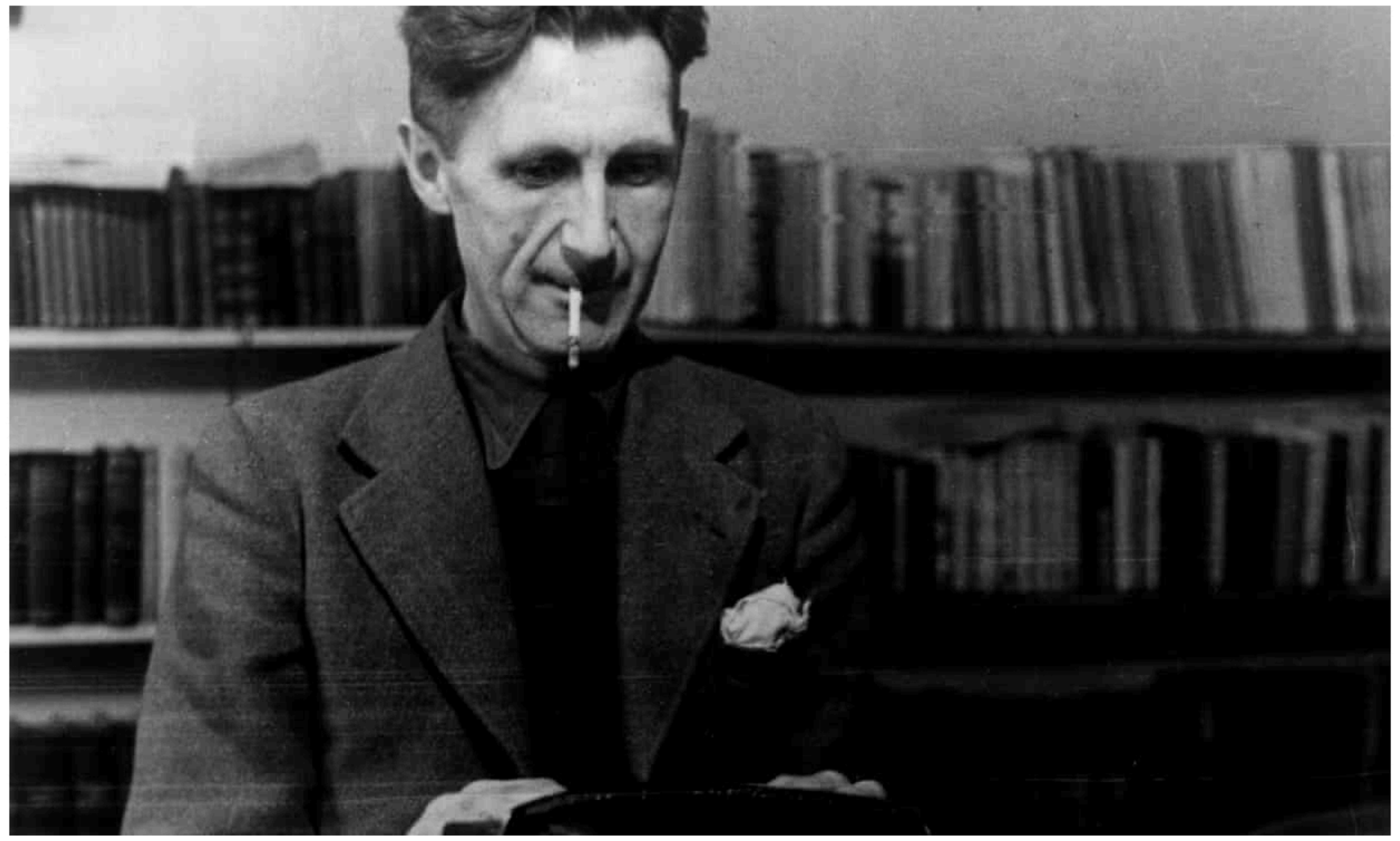
Across the street from NeurIPS, two unicorns walk a bar... at least that's what my picture shows. I've never seen such a multicolored, beautiful forest of sapphire eyes on the same corner of the street in a bar before. People have been coming from all over to watch. I wouldn't be surprised if that's where your favorite eSports tournament is tonight.

NeurCo's built on strong eSports teams: they've been participating in Lisk development to help build their decentralized autonomous organization, NeurDash, and They've hosted a

Artificial intelligence (AI)

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse



- GPT-2, allegedly "too dangerous to release"
- fluent
- even more amusing, still superficial

GPT-2

Completion

Across the street from NeurIPS, two unicorns walk a bar... at least that's what my picture shows. I've never seen such a multicolored, beautiful forest of sapphire eyes on the same corner of the street in a bar before. People have been coming from all over to watch. I wouldn't be surprised if that's where your favorite eSports tournament is tonight.

NeurCo's built on strong eSports teams: they've been participating in Lisk development to help build their decentralized autonomous organization, NeurDash, and They've hosted a

Completion

Two lemurs walk on a road and another joins in. The total number of lemurs on the road is not 100 as claimed, but about 80 or so. I am not sure how much of this lemur traffic is human. It is possible that humans may have been attracted by the lemurs' scent or the food. I am not sure how lemurs can be attracted to the scent of a human. Maybe they can smell the food and food smells in other species, or maybe the food smells to them and they become attracted.

Often plausible for first few sentences in a context of surrealist fiction, where there are no facts of the matter

Demonstrably poor in nonfiction, where facts matter

[same is true of Eliza, Turing bots, etc]

prediction at the word-level
≠ prediction at the *world*-level

Completion

If you drink hydrochloric acid by the bottle full you will probably get sick of it fast if you just try to drink it by the bottle full. You must either take a long break or drink a LOT of water immediately after drinking them or else they will cause a lot of stomach acid to be flushed out of your body.

4. The best method of drinking it is taking an 8oz or 9oz bottle of Hydrochloric Acid in it's proper amount. This way you can avoid the stomach upset and still get it in your system.

Shilpesh Garg
@ShilpeshGarg

Replying to @GaryMarcus

GPT prediction

Translate Tweet

Custom prompt

the winner of the ai debate between gary and yoshua will be

GENERATE ANOTHER

Completion

the winner of the ai debate between gary and yoshua will be a puppet to play puppets! They will carry on with the war, attack the baron lord, kill gavin, kidnap zelen, then fight Yuletide while things end for a while and we run again for a while.

"Local coherence; global gibberish"
- Dan Brickley

Adversarial NLI: A New Benchmark for Natural Language Understanding

Yixin Nie*, Adina Williams†, Emily Dinan†, Mohit Bansal*, Jason Weston†, Douwe Kiela†
*UNC Chapel Hill
†Facebook AI Research

Abstract

We introduce a new large-scale NLI benchmark dataset, collected via an iterative, adversarial human-and-model-in-the-loop procedure. We show that training models on this new dataset leads to state-of-the-art performance on a variety of popular NLI benchmarks, while posing a more difficult challenge with its new test set. Our analysis sheds light on the shortcomings of current state-of-the-art models, and shows that non-expert annotators are successful at finding their weaknesses. The data collection method can be applied in a never-ending learning scenario, becoming a moving target for NLU, rather than a static benchmark that will quickly saturate.

1 Introduction

2018), and rapidly had to be extended into SuperGLUE (Wang et al., 2019). This raises an important question: Can we collect a large benchmark dataset that can last longer?

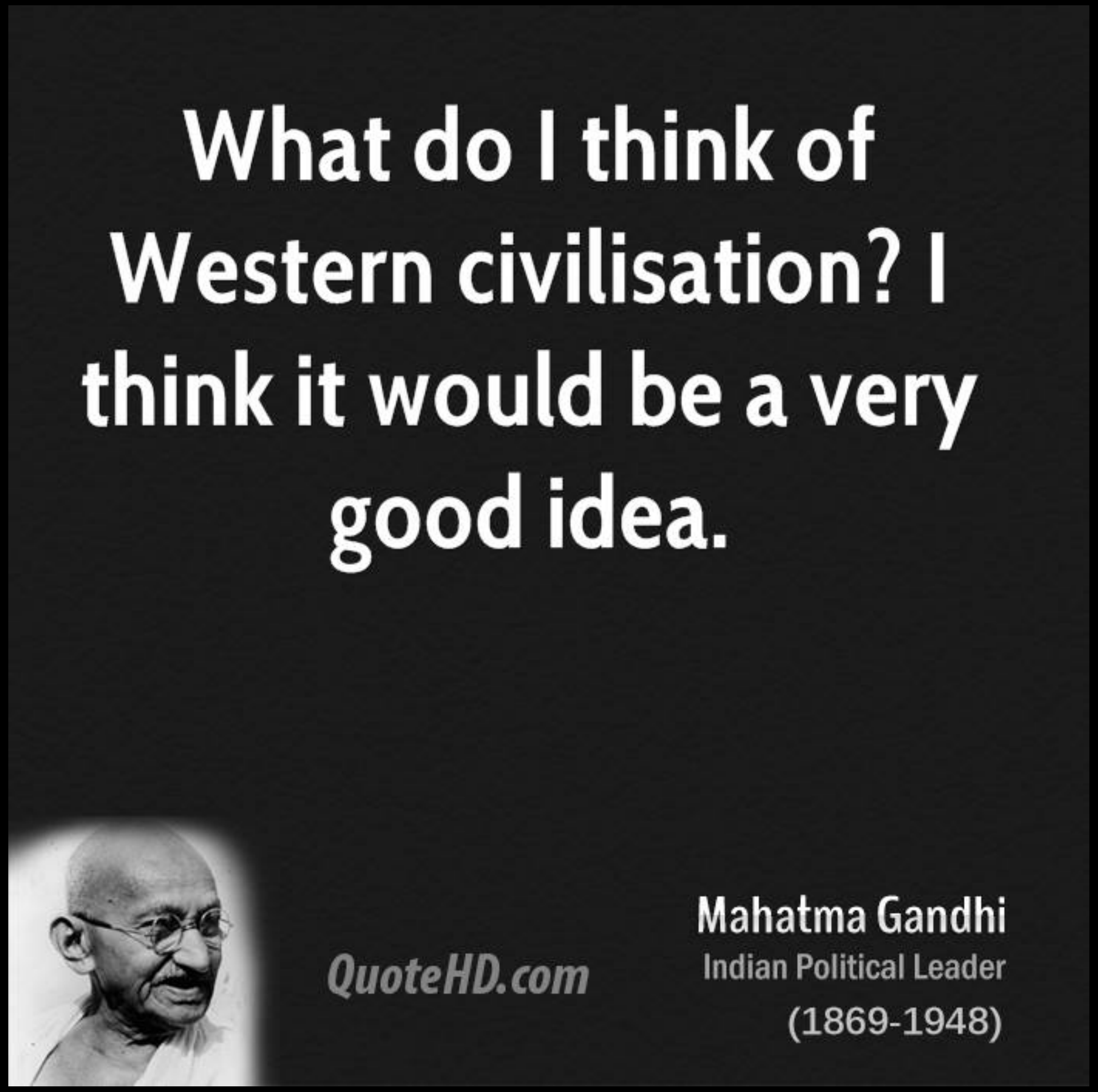
The speed with which benchmarks become obsolete raises another important question: are current NLU models genuinely as good as their high performance on benchmarks suggests? A growing body of evidence shows that state-of-the-art models learn to exploit spurious statistical patterns in datasets (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; Glockner et al., 2018; Geva et al., 2019; McCoy et al., 2019), instead of learning *meaning* in the flexible and generalizable way that humans do. Given this, human annotators—be they seasoned NLP researchers or non-experts—might easily be able to construct examples that expose model brittleness.

"A growing body of evidence shows that state-of-the-art models learn to exploit spurious statistical patterns in datasets... instead of learning meaning in the flexible and generalizable way that humans do."

Deep understanding

Deep understanding is being able to

- construct an internal model of what is said/depicted in a story/article/movie/etc
- perform everyday inferences about what is left unsaid



There is no AI system with deep understanding yet

Arguably the closest to deep understanding is ... the oft-maligned CYC

2,459 views | Jul 3, 2019, 01:41pm

What AI Can Learn From Romeo & Juliet

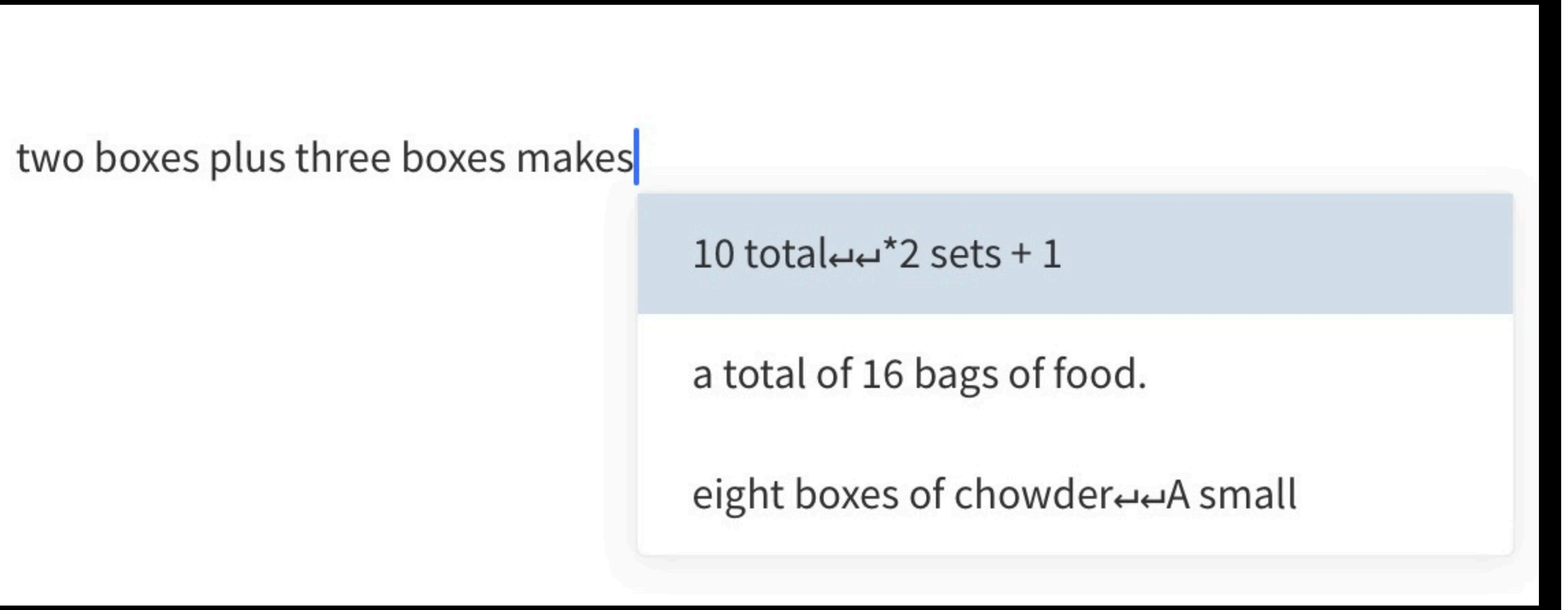
Doug Lenat Contributor
COGNITIVE WORLD Contributor Group
AI

The story so far.

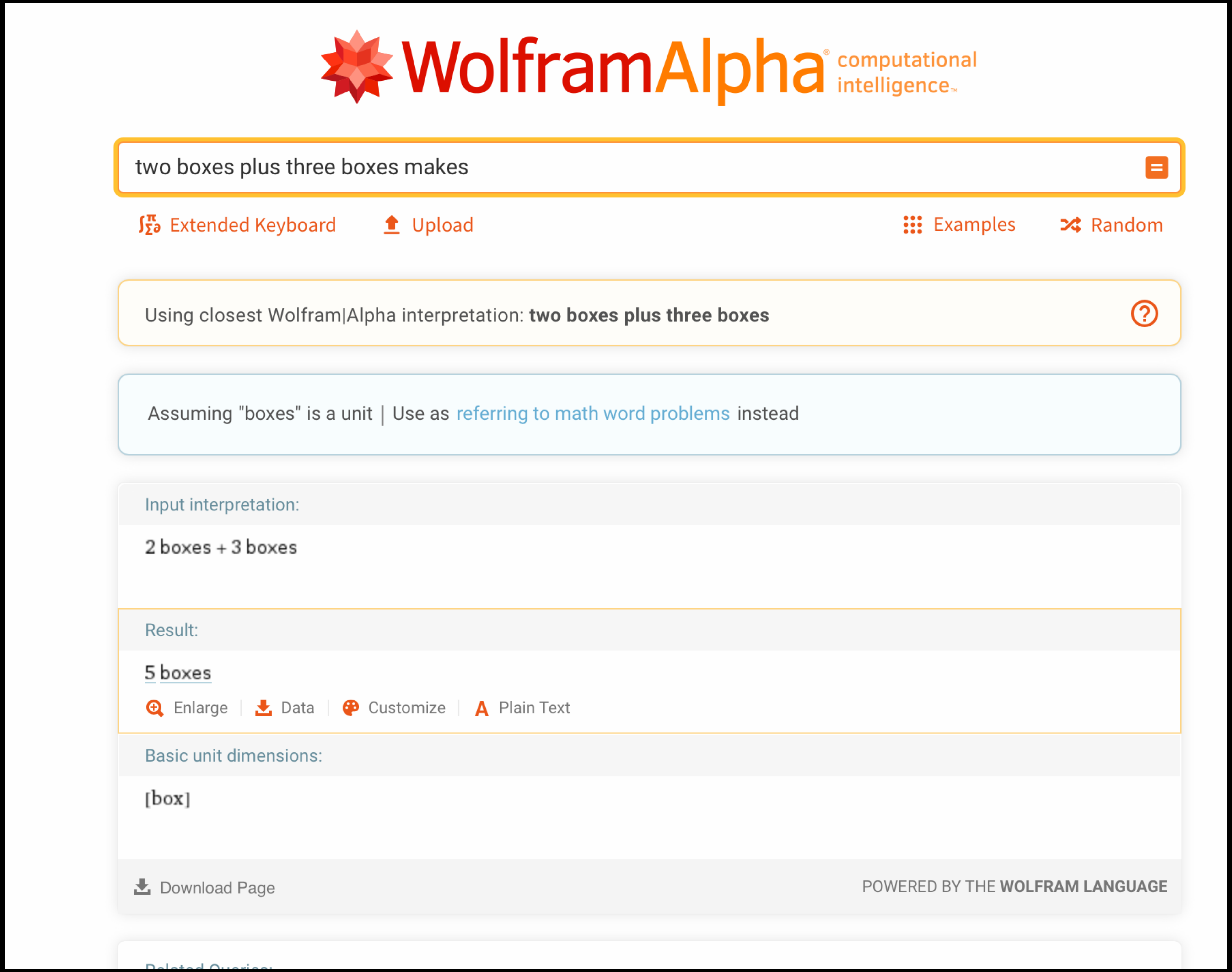
When someone talks about "AI", today, they are referring to one particular type of AI: multi-layer neural nets trained on big data to recognize patterns. These so-called "deep learning" algorithms are great at learning more or less the same sort of stimulus/response functionality that our right brain hemispheres carry out – what Daniel Kahneman calls "thinking fast". This is also what the *entire* brains of most animals do. So a better name for them might be AAI's, for **A**rtificial **A**nimal **I**ntelligences. In my last Forbes article (*Not Good As Gold: Today's AI's Are Dangerously Lacking In AU (Artificial Understanding)*) I argued that almost all of today's AI's have little or no left brain function – logical, causal, "thinking slowly". *Homo sapiens* pays a huge price for having an over-sized bicameral brain (high birthing pain and risk) but upon reflection it's worth it — in particular, you or I couldn't perform such *reflection* without it! Our ability to

- Can make nuanced inferences about character motivations, far more subtle than any deep learning QA system I am aware of*
- *But: system doesn't have a natural language front-end (you can't just feed Romeo & Juliet in)
- Relies on human experts to encode each problem
- There are also serious issues of coverage, dealing with uncertainty etc
- Never been formally evaluated by the community
- Not (yet) anything like a full-service, autonomous understanding system

shallow prediction vs deeper parse



GPT-2 @
huggingface.co



How might we get to deeper understanding
Two ways of thinking about the path forward

a. in terms of what machinery might be needed

b. in terms of what signposts might we create along the way

What I wish we could do today

What we will actually do today, at best

Computational prerequisites for deep understanding

Summary of *Rebooting AI*'s proposed recipe for achieving deep understanding

In short, our recipe for achieving common sense, and ultimately general intelligence, is this: Start by developing systems that can represent the core frameworks of human knowledge: time, space, causality, basic knowledge of physical objects and their interactions, basic knowledge of humans and *their* interactions. Embed these in an architecture that can be freely extended to every kind of knowledge, keeping always in mind the central tenets of abstraction, compositionality, and tracking of individuals. Develop powerful reasoning techniques that can deal with knowledge that is complex, uncertain, and incomplete and that can freely work both top-down and bottom-up. Connect these to perception, manipulation, and language. Use these to build rich cognitive models of the world. Then finally the keystone: construct a kind of human-inspired learning system that uses all the knowledge and cognitive abilities that the AI has; that incorporates what it learns into its prior knowledge; and that, like a child, voraciously learns from every possible source of information: interacting with the world, interacting with people, reading, watching videos, even being explicitly taught. Put all that together, and that's how you get to deep understanding.

It's a tall order, but it's what has to be done.

Possible minimal requirement for deep understanding

- mechanisms for operating over abstractions
- mechanisms for physical reasoning
- mechanisms for psychological reasoning
- mechanisms for temporal reasoning
- a large body of common knowledge
- machinery for acquiring additional knowledge
- (general intelligence a la Chollet?)

Benchmarks as a way forward?

1. Benchmarks don't encourage out-of-the-box-thinking

"One big challenge the community faces is that if you want to get a paper published in machine learning now it's got to have a table in it, with all these different data sets across the top, and all these different methods along the side, and your method has to look like the best one. If it doesn't look like that, it's hard to get published. I don't think that's encouraging people to think about radically new ideas."

-- Geoff Hinton, 2018 interview with *Wired*

2. Benchmarks are often easily gamed

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

R. Thomas McCoy,[1] Ellie Pavlick,[2] & Tal Linzen[1]
[1]Department of Cognitive Science, Johns Hopkins University
[2]Department of Computer Science, Brown University
tom.mccoy@jhu.edu, ellie_pavlick@brown.edu, tal.linzen@jhu.edu

Abstract

A machine learning system can score well on a given test set by relying on heuristics that are effective for frequent example types but break down in more challenging cases. We study this issue within natural language inference (NLI), the task of determining whether one sentence entails another. We hypothesize that statistical NLI models may adopt three fallible syntactic heuristics: the lexical overlap heuristic, the subsequence heuristic, and the constituent heuristic. To determine whether models have adopted these heuristics, we introduce a controlled evaluation set called HANS (Heuristic Analysis for NLI Systems), which contains many examples where the heuristics fail. We find that models trained on MNLI, including BERT, a state-of-the-art model, perform very poorly on HANS, suggesting that they have indeed adopted these heuristics. We conclude that there is substantial room for improvement

example, neural networks trained to recognize objects are misled by contextual heuristics: a network that is able to recognize monkeys in a typical context with high accuracy may nevertheless label a monkey holding a guitar as a human, since in the training set guitars tend to co-occur with humans but not monkeys (Wang et al., 2018). Similar heuristics arise in visual question answering systems (Agrawal et al., 2016).

The current paper addresses this issue in the domain of natural language inference (NLI), the task of determining whether a **premise** sentence entails (i.e., implies the truth of) a **hypothesis** sentence (Condoravdi et al., 2003; Dagan et al., 2006; Bowman et al., 2015). As in other domains, neural NLI models have been shown to learn shallow heuristics, in this case based on the presence of specific words (Naik et al., 2018; Sanchez et al., 2018). For example, a model might assign a label of *contradiction* to any input containing the word *not*, since

"A machine learning system can score well on a given test set by relying on heuristics that are effective for frequent example types but break down in more challenging cases."

Nasrin Mostafazadeh @nasrinmmm · 10/30/19

I totally agree with @GaryMarcus. Building a reliable NLU benchmark that is not prone to exploitation of data intricacies by models is very challenging; and more often than not we are not even lucky enough to uncover such cases (we were lucky in the case of Story Cloze Test v1).

The problem is that I believe that what will happen is that you will simply wind up spawning a whole host of new and ultra-clever brute-force techniques to solve the "Winograd Challenge" without solving the problem of understanding whatsoever. .... Getting people to spend huge amounts of time on just one kind of challenge is not going to be helpful. In fact, I fear it will be counterproductive, because I don't think that anyone who will be moved to tackle this particular challenge is likely to take up the deeper and more general challenge of what language understanding really is. People are daunted by that, as well they should be, and no one is going to be motivated by a prize to suddenly tackle that gigantic challenge. Instead, very smart engineering types are going to be motivated to seek clever tricks that will allow computers to solve this very narrow type of linguistic disambiguation problem with a high degree of accuracy.

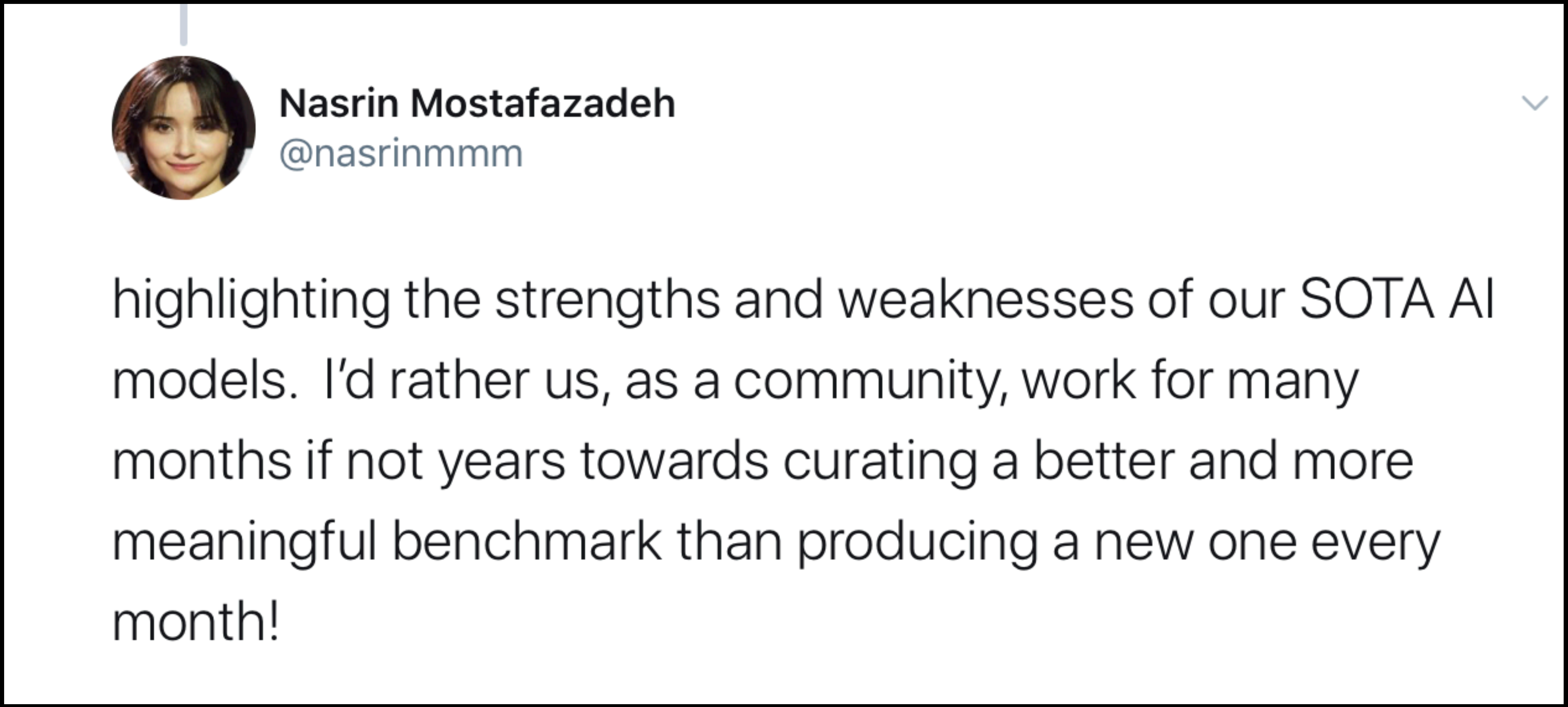
Douglas Hofstadter, email of February 5, 2011 to Ernie Davis

which relates to The Kaggle Effect

These successes demonstrate the importance of *setting clear goals and adopting objective measures of performance that are shared across the research community*. However, optimizing for a single metric or set of metrics often leads to tradeoffs and shortcuts when it comes to everything that isn't being measured and optimized for (a well-known effect on Kaggle, where winning models are often overly specialized for the specific benchmark they won and cannot be deployed on real-world versions of the underlying problem). In the

Chollet 2019

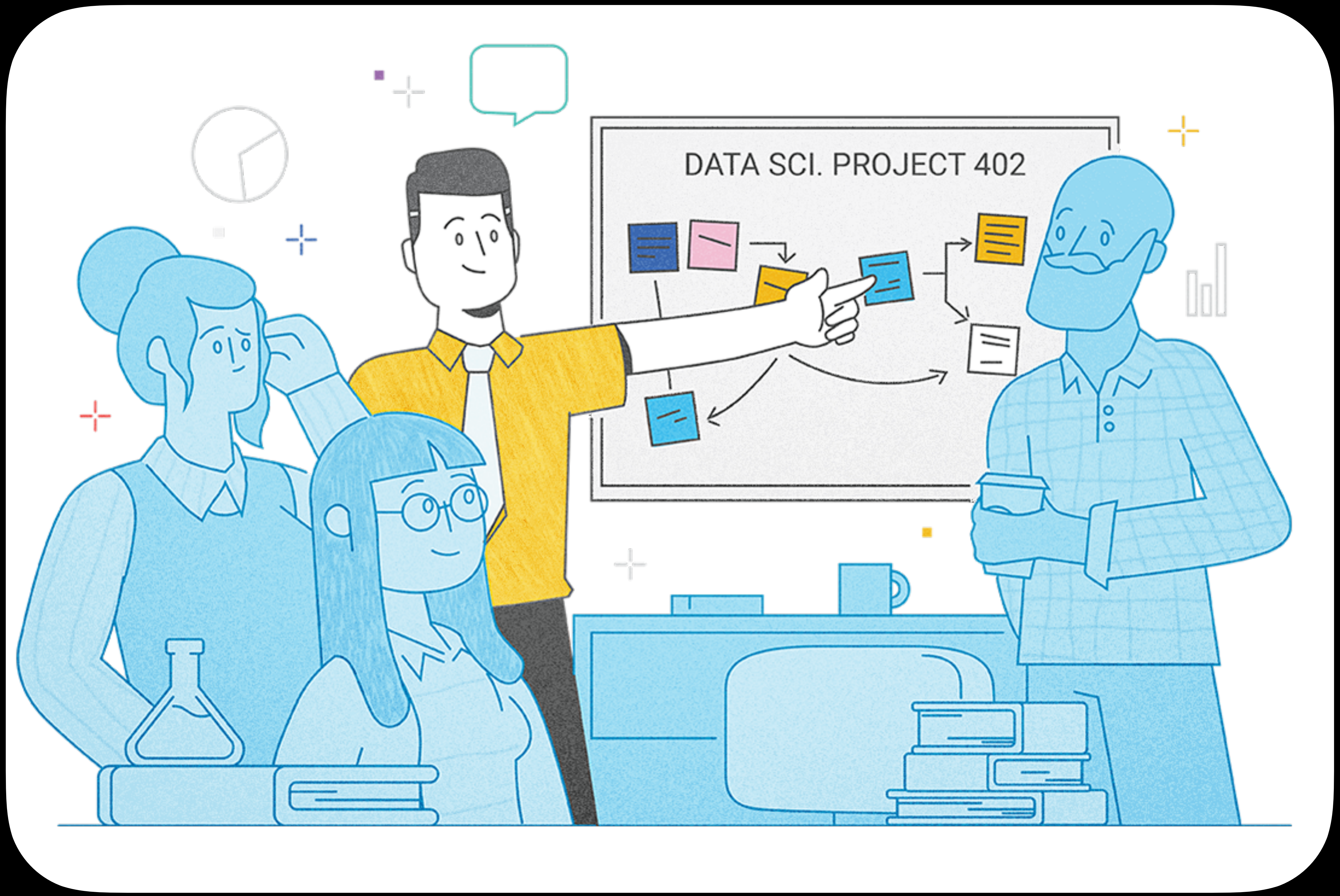
3. Good benchmarks take a lot of of time to develop

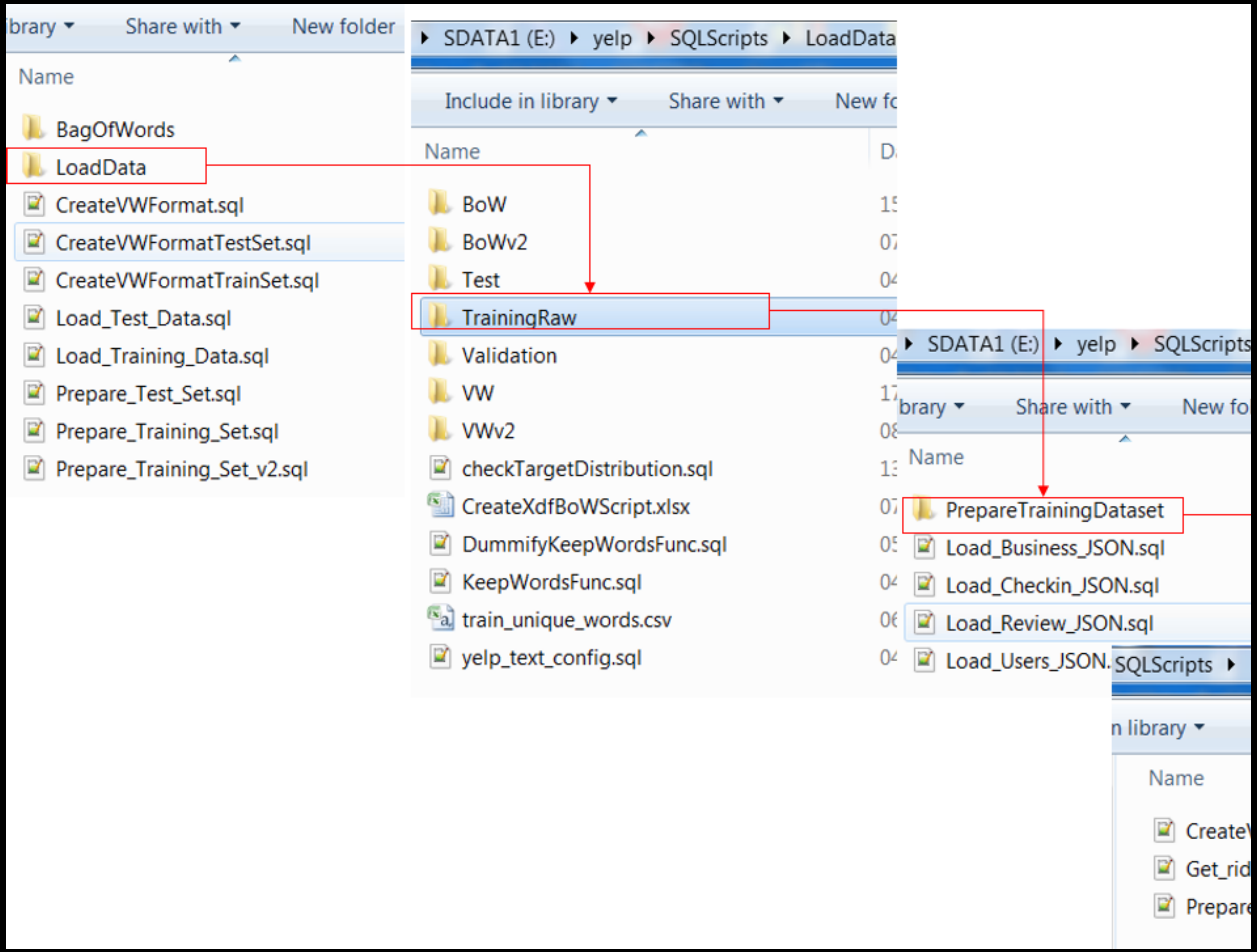

Nasrin Mostafazadeh
@nasrinmmm

highlighting the strengths and weaknesses of our SOTA AI models. I'd rather us, as a community, work for many months if not years towards curating a better and more meaningful benchmark than producing a new one every month!

- Benchmarks take time to develop
- The specific idea we will introduce today is only about a month into development (talk invite came last week)
- new benchmark certainly not finalized yet, let alone adversarially vetted

4. Benchmarks are prepackaged; human experience rarely is





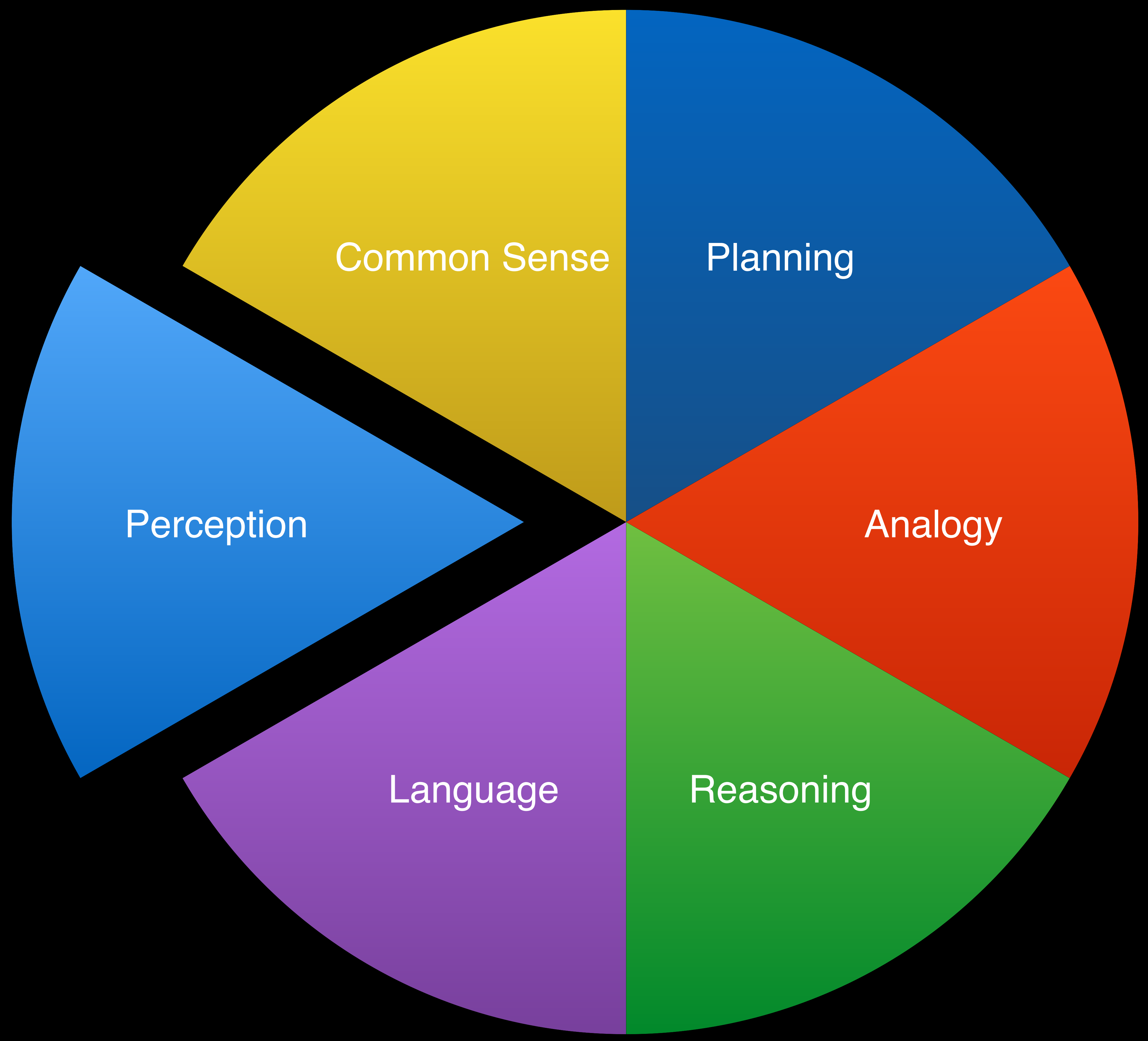
Life is not a Kaggle competition.

5. We shouldn't expect any single benchmark to suffice

"There is no one way the mind works, because the mind is not one thing. Instead, the mind has parts, and the different parts of the mind operate in different ways: Seeing a color works differently than planning a vacation, which works differently than understanding a sentence, moving a limb, remembering a fact, or feeling an emotion."

-- Chazz Firestone and Brian Scholl

- Intelligence is clearly multidimensional
- Deep understanding involves many facets of intelligence



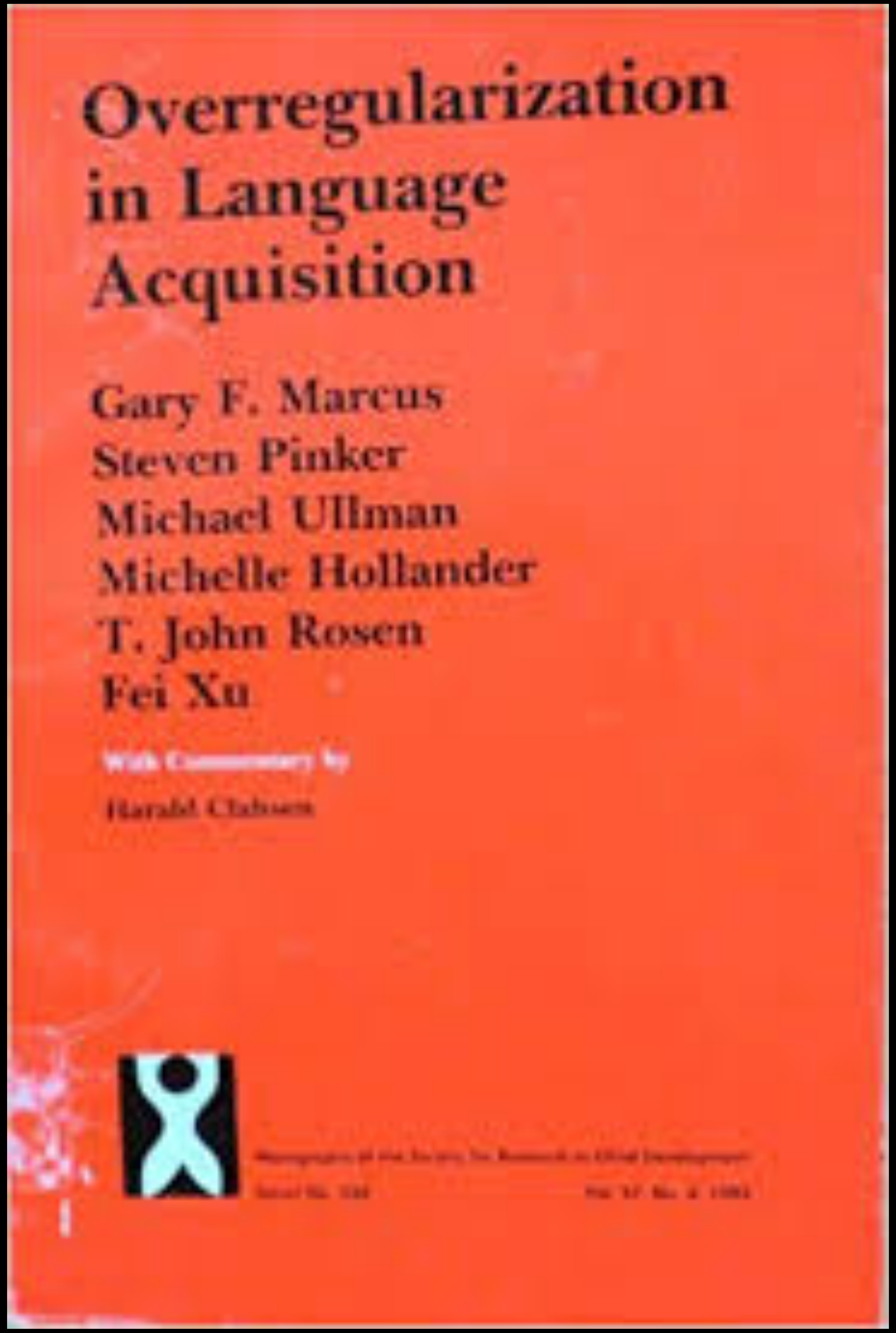
Advice to Young Scholars

- Don't wait for the field to make you a formal, Kaggle-ready benchmark. Notice a dataset, or even just a question, and create your own challenges.
  - Don't just look to what the ML community has published
  - Example: there's plenty of extant data in the form of experiments in fields like psychology and psycholinguistics that is ripe for exploration.
  - And plenty of work suggesting other challenges that could be developed
  - And don't forget that children rarely get their data in neatly packaged form.

Related Aside: Twitter's rumors to the contrary, this is hardly the first time I have presented a specific challenge to the field

Children's Overregularization Errors

- 1992
- widely modeled throughout the 1990s
- debate simmered down, still AFAIK no model that really captures all of the longitudinal and lexical data we presented
- not packaged as a benchmark with a pre-made corpus, but kids aren't given a pre-made corpus, either



Marcus et al (1999, *Science*)

Rule Learning by Seven-Month-Old Infants

G. F. Marcus, * S. Vijayan, S. Bandi Rao, P. M. Vishton

A fundamental task of language acquisition is to extract abstract algebraic rules. Three experiments show that 7-month-old infants attend longer to sentences with unfamiliar structures than to sentences with familiar structures. The design of the artificial language task used in these experiments ensured that this discrimination could not be performed by counting, by a system that is sensitive only to transitional probabilities, or by a popular class of simple neural network models. Instead, these results suggest that infants can represent, extract, and generalize abstract algebraic rules.

What learning mechanisms are available to infants on the cusp of language learning? One learning mechanism that young infants can exploit is statistical in nature. For example, Saffran *et al.* (1) found that the looking behaviors of 8-month-old infants indicated a sensitivity to statistical information inherent in sequences of speech sounds produced in an artificial language--for example, transitional probabilities, which are estimates of how likely one item is to follow another. In the corpus of sentences "The boy loves apples. The boy loves oranges." the transitional probability between the words "the" and "boy" is 1.0 but the transitional probability between the words "loves" and "apples" is $1/2 = 0.5$.

It has been suggested that mechanisms that track statistical information, or connectionist models that rely on similar sorts of information [for example, the simple recurrent network (SRN) (2)], may suffice for language learning (3). The alternative possibility considered here is that children might possess at least two learning mechanisms, one for learning statistical information and another for learning "algebraic" rules (4)--open-ended abstract relationships for which we can substitute arbitrary items. For instance, we can substitute any value of x into the equation $y = x + 2$. Similarly, if we know that in English a sentence can be formed by concatenating any plural noun phrase with any verb phrase with plural agreement, then as soon as we discover that "the three blickets" is a well-formed plural noun phrase and that "reminded Sam of Tibetan art" is a well-formed verb phrase with plural agreement, we can infer that "The three blickets reminded Sam of Tibetan art." is a well-formed sentence.

To date, however, there has been no direct empirical test for determining whether young infants can actually learn simplified versions of such algebraic rules. A number of previous experiments drawn from the literature of speech perception (not aimed at the question of rule learning) are consistent with the possibility that infants might learn algebraic rules, but each of these prior experiments could be accounted for by a system that extracted only statistical tendencies. For example, infants who are habituated to a series of two-syllable words attend longer when confronted with a three-syllable word (5). An infant who attended longer to a three-syllable word might have noticed a violation of a rule (for example, "all the words here are two syllables"), but an infant could also have succeeded with a statistical device that noted that the three-syllable word had more syllables than the average number of syllables in the preceding utterance. Similarly, Gomez and Gerken (6) found that infants who were habituated to a set of sentences constructed from an artificial grammar (VOT-PEL-JIC; PEL-TAM-PEL-JIC) could distinguish between new sentences that were consistent with this grammar (VOT-PEL-TAM-PEL-JIC) from new sentences that were not consistent (VOT-TAM-PEL-RUD-JIC). Such learning might reflect the acquisition of rules, but because all the test sentences were constructed with the same words as in the habituation sentences (albeit rearranged), in these test sentences it was possible to distinguish the test sentence on the basis of statistical information such as transitional probabilities (for example, in the training corpus, VOT was never followed by TAM)--without recourse to a rule.

We tested infants in three experiments in which simple statistical or counting mechanisms would not suffice to learn the rule that was generating the sequences of words. In each experiment, infants were habituated to three-word sentences constructed from an artificial language (7) and then tested on three-word sentences composed entirely of artificial words that did not appear in the habituation. The test sentences varied as to whether they were consistent or inconsistent with the grammar of the habituation sentences. Because none of the test words appeared in the habituation phase, infants could not distinguish the test sentences based on transitional probabilities, and because the test sentences were the same length and were generated by a computer, the infant could not distinguish them based on statistical properties such as number of syllables or prosody.

We tested infants with the familiarization preference procedure as adapted by Saffran *et al.* (1, 8, 9); if infants can abstract the underlying structure and generalize it to novel words, they should attend longer during presentation of the inconsistent items than during presentation of consistent items.

Subjects were 7-month-old infants, who were younger than those studied by Saffran *et al.* but still old enough to be able to distinguish words in a fluent stream of speech (8). In the first experiment, 16 infants were randomly assigned to either an "ABA" condition or an "ABB" condition. In the ABA condition, infants were familiarized with a 2-min speech sample (10) containing three repetitions of each of 16 three-word sentences that followed an ABA grammar, such as "ga ti ga" and "li na li." In condition ABB, infants were familiarized with a comparable speech sample in which all training sentences followed an ABB grammar, such as "ga ti ti" and "li na na" (11).

In the test phase, we presented infants with 12 sentences that consisted entirely of new words, such as "wo fe wo" or "wo fe fe" (12). Half the test trials were "consistent sentences," constructed from the same grammar as the one with which the infant was familiarized (an ABA test sentence for infants trained in the ABA condition and an ABB sentence for infants trained in the ABB condition), and half the test trials were "inconsistent sentences" that were constructed from the grammar on which the infant was not trained (13).

We found that 15 of 16 infants showed a preference for the inconsistent sentences

G. F. Marcus, S. Vijayan, S. Bandi Rao, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA. P. M. Vishton, Department of Psychology, Amherst College, Amherst, MA 01002, USA.

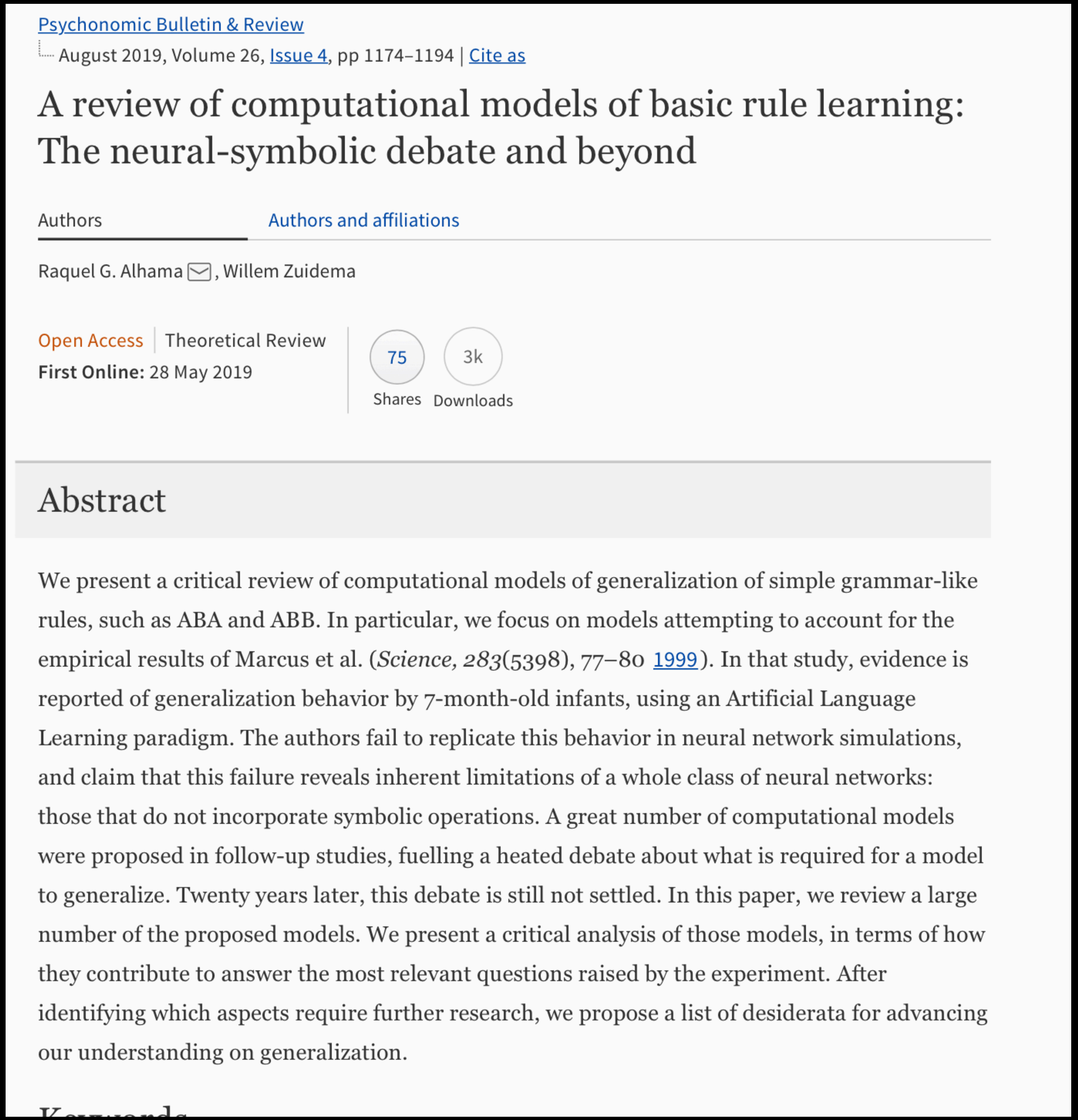
* To whom correspondence should be addressed. E-mail: gary.marcus@nyu.edu.

- 2 min habituation, followed by test string
- looking times as a measure of attention

- la ti ti, ga na na, etc
- test trials consisted of all new vocabulary, using new set of phonemes
- some with same grammar, some with different grammar
    - eg wo fe fe [ABB} vs wo wo fe [AAB]
    - infants looked longer to items following new grammar
- abstraction naturally described in terms of operations over variables, not so easily captured by traditional neural nets

Infant rule learning

- many models were proposed in 1999
- reviewed most in 2001
- still an area of active research, even in 2019
- also many follow up experiments, extensions to younger children etc
- not published in ML journals, but key paper was published in *Science*, reviewed in *The Algebraic Mind*
- Highly relevant to ML, and an example of out-of-core-discipline work that could strengthen ML



Psychonomic Bulletin & Review
August 2019, Volume 26, Issue 4, pp 1174–1194 | Cite as

A review of computational models of basic rule learning: The neural-symbolic debate and beyond

Authors | Authors and affiliations

Raquel G. Alhama, Willem Zuidema

Open Access | Theoretical Review
First Online: 28 May 2019

Abstract

We present a critical review of computational models of generalization of simple grammar-like rules, such as ABA and ABB. In particular, we focus on models attempting to account for the empirical results of Marcus et al. (*Science, 283*(5398), 77–80 1999). In that study, evidence is reported of generalization behavior by 7-month-old infants, using an Artificial Language Learning paradigm. The authors fail to replicate this behavior in neural network simulations, and claim that this failure reveals inherent limitations of a whole class of neural networks: those that do not incorporate symbolic operations. A great number of computational models were proposed in follow-up studies, fuelling a heated debate about what is required for a model to generalize. Twenty years later, this debate is still not settled. In this paper, we review a large number of the proposed models. We present a critical analysis of those models, in terms of how they contribute to answer the most relevant questions raised by the experiment. After identifying which aspects require further research, we propose a list of desiderata for advancing our understanding on generalization.

Adult generalization of inflection to foreign phonemes

- Series of papers with Iris Berent around 2002
- Ongoing focus, eg work by Joe Pater later year
- Not well-known in ML community, fairly well-known among those following computational models of linguistics
- Not framed as a Kaggle set, but not captured by current language models

COGNITION

ELSEVIER Cognition 83 (2002) 113–139

www.elsevier.com/locate/cognit

The scope of linguistic generalizations: evidence from Hebrew word formation

Iris Berent[a,*], Gary F. Marcus[b], Joseph Shimron[c], Adamantios I. Gafos[d]

[a]Department of Psychology, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431-0991, USA
[b]Department of Psychology, New York University, New York, NY, USA
[c]School of Education, University of Haifa, Haifa, Israel
[d]Department of Linguistics, New York University, New York, NY, USA

Received 6 March 2001; received in revised form 6 September 2001; accepted 14 November 2001

Abstract

Does the productive use of language stem from the manipulation of mental variables (e.g. "noun", "any consonant")? If linguistic constraints appeal to variables, rather than instances (e.g. "*dog*", "*m*"), then they should generalize to *any* representable novel instance, including instances that fall beyond the phonological space of a language. We test this prediction by investigating a constraint on the structure of Hebrew roots. Hebrew frequently exhibits geminates (e.g. *ss*) in its roots, but it strictly constraints their location: geminates are frequent at the end of the root (e.g. *mss*), but rare at its beginning (e.g. *ssm*). Symbolic accounts capture the ban on root-initial geminates as *XXY, where X and Y are variables that stand for any two distinct consonants. If the constraint on root structure appeals to the identity of abstract variables, then speakers should be able to extend it to root geminates with foreign phonemes, including phonemes with foreign feature values. We present findings from three experiments supporting this prediction. These results suggest that a complete account of linguistic processing must incorporate mechanisms for generalization outside the representational space of trained items. Mentally-represented variables would allow speakers to make such generalizations.

Keywords: Language; Linguistic generalizations; Hebrew word formation

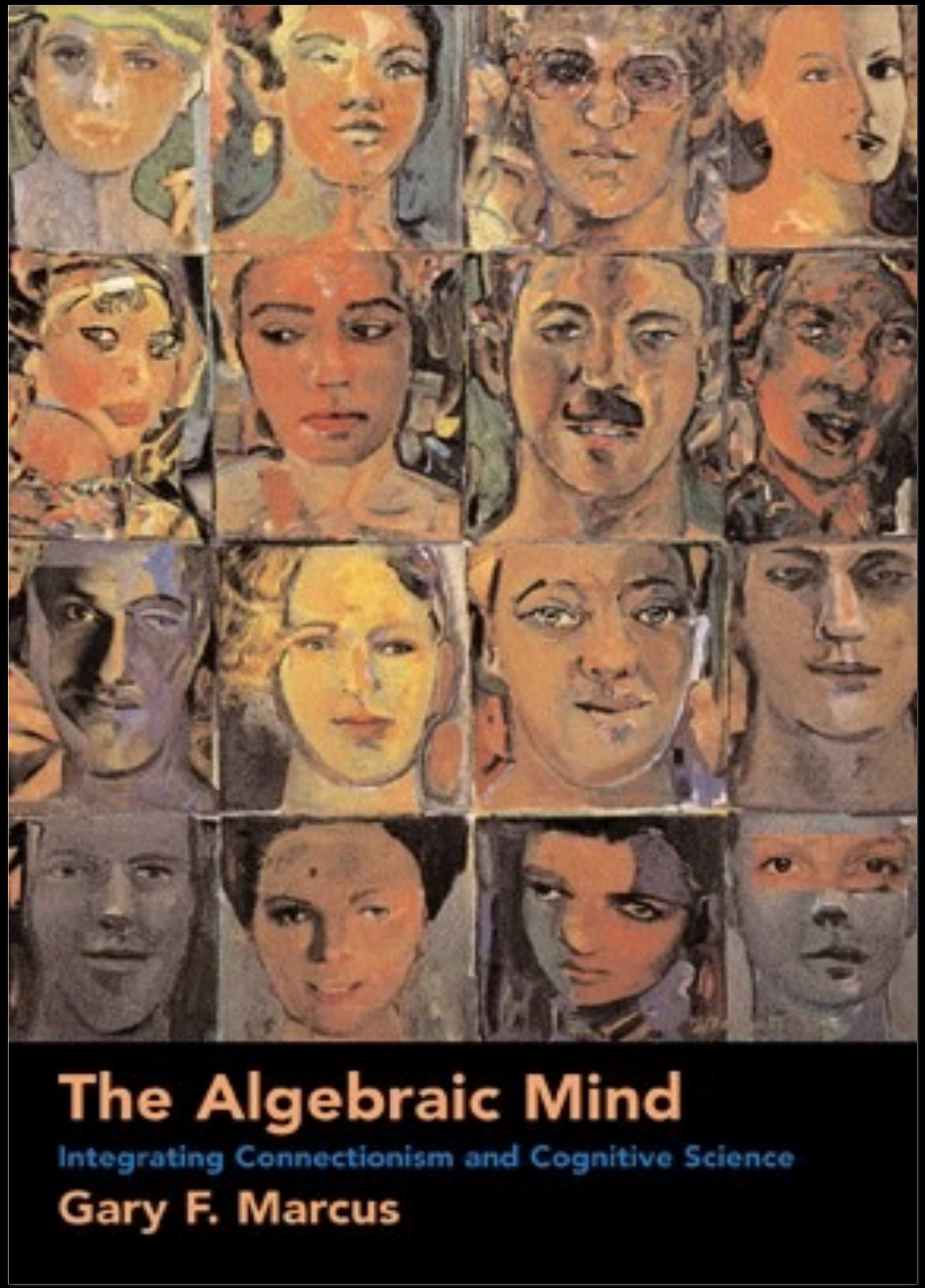
1. Introduction

Productivity is at the core of linguistic competence (Chomsky, 1980): speakers routinely produce and comprehend numerous sentences they have never heard before.

* Corresponding author. Fax: +1-561-297-2160.
E-mail address: iberent@fau.edu (I. Berent).

0010-0277/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0010-0277(01)00167-6

All these are examples of free generalization of
universally quantified one-to-one-mappings [UQOTOMs]



A rose is a rose
A tulip is a tulip
A lilac is a lilac
_____
A lily is a ____

0110 - > 0110
1100 -> 1100
1010 -> 1010
_____
1111 -> ____

la ta ta
ga na na
_____
wo fe wo
vs *wo fe fe*

Marcus et al, 1999, Science
w 7-month-olds
later replicated (inter alia) by Gervain et al, 2012
w newborns

"multilayer perceptron[s] cannot generalize [a certain class of universally quantified function] outside the training space. .. In some cases it appears that humans can freely generalize from restricted data, [in these cases a certain class of] multilayer perceptions that are trained by back-propagation are inappropriate" - Marcus, 2001

Tweet

Steven Hansen @Zergylord · 8/28/19

A @OpenAI GPT-2 prompt taken from @GaryMarcus's The Algebraic Mind, showing what neural nets couldn't do. Sometimes a quantitative change in model/data complexity can give rise to qualitatively different behavior.

Prompt:	**A rose is a rose.**
	A lily is a lily.
Generated completion:	**A blicket is a** blicket.
	A hoop is a hoop.
	A crook is a crook.
	A tine is a tine.
	And it was a day like this when a girl, when she heard someone call her by her name,
	But her heart beat for the sun.
	That's what a little girl thinks about day by day.
	A day like this is where she starts to live,
	But the sun isn't there to welcome the girl with her.
	That's where she starts thinking about life,
	Her face's the face of something she needs,

1 8 30

Gary Marcus @GaryMarcus · 8/28/19

honored and flattered and it's very very cute, but it still can't just stick to the pattern...

1 5

Prompt:	**1. A rose is a rose**
	2. A lily is a lily
	3. A blicket is a blicket
Generated completion:	4. A horse is a horse
	5. A lass is a lass
	6. A rose is a rose
	7. A dog is a dog
	8. The sky is blue
	9. A lion is a lion
	10. A snake is a snake
	11. A dog is a dog

Even today there are challenges in learning UQOTMS in systems that lack operations over variables

Language

- Lake and Baroni (2018)
- Evans & Greffenstette (2018)

Number

Adding 1+1+1...1 where number of 1's > 6 [successor function, where each n has a unique output)

Other insights Examining the performance on adding multiple integers, we tested the models on adding $1 + 1 + \cdots + 1$, where 1 occurs n times. Both the LSTM and Transformer models gave the correct answer for $n \leq 6$, but the incorrect answer of 6 for $n = 7$ (seemingly missing one of the 1s), and other incorrect values for $n > 7$. (The models are trained on sequences of random integers up to length 10, and are capable of giving the correct answer on longer sequences of far bigger numbers, for example `-34 + 53 + -936 + -297 + 162 + -242 + -128`.) We do not have a good explanation for this behaviour; one hypothesis is that the models calculate subsums and then combine these, but rely on different input numbers to align the subsums, and fail when the input is "camouflaged" by consisting of the same number repeated multiple times.

- [Saxton et al 2019]

Only now is the importance of this issue started to become recognized

A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

Yoshua Bengio[1,2,5], Tristan Deleu[1], Nasim Rahaman[4], Nan Rosemary Ke[3], Sébastien Lachapelle[1], Olexa Bilaniuk[1], Anirudh Goyal [1] and Christopher Pal[3,5]
Mila, Montréal, Québec, Canada
[1] **Université de Montréal**
[2] **CIFAR Senior Fellow**
[3] **École Polytechnique Montréal**
[4] **Ruprecht-Karls-Universität Heidelberg**
[5] **Canada CIFAR AI Chair**

1. Introduction

Current machine learning methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice. It is not enough to obtain good generalization on a test set sampled from the same distribution as the training data, we would also like what has been learned in one setting to generalize well in other related distributions. These distributions may involve the same concepts that were seen previously by the learner, with the changes typically arising because of actions of agents. More generally, we would like what has been learned previously to form a rich base from which very fast adaptation to a new but related distribution can take place, i.e., obtain good transfer. Some new concept may have to be learned but because most of the other relevant concepts have already been captured by the learner (as well as how they can be composed), learning can be very fast on the transfer distribution.

Short of any assumption, it is impossible to have a successful transfer to an unrelated distribution. In

Turing Olympics

- 2015 AAAI session
- 2016 special issue of AI magazine, "Beyond the Turing Test". coedited with Rossi and Veloso
- About 7 different challenges proposed
    - comprehension
    - social cognition
    - "Ikea"-like assembly, etc
- Only two (Winograd Schema Challenge, and grade school science exams) have been addressed in the literature
- Lots of stuff there still worth working on

Toward a Comprehension Challenge, Using Crowdsourcing as a Tool

Praveen Paritosh, Gary Marcus

■ *Human readers comprehend vastly more, and in vastly different ways, than any existing comprehension test would suggest. An ideal comprehension test for a story should cover the full range of questions and answers that humans would expect other humans to reasonably learn or infer from a given story. As a step toward these goals we propose a novel test, the crowdsourced comprehension challenge (C^3), which is constructed by repeated runs of a three-person game, the Iterative Crowdsourced Comprehension Game (ICCG). ICCG uses structured crowdsourcing to comprehensively generate relevant questions and supported answers for arbitrary stories, whether fiction or nonfiction, presented across a variety of media such as videos, podcasts, and still images.*

Artificial Intelligence (AI) has made enormous advances, yet in many ways remains superficial. While the AI scientific community had hoped that by 2015 machines would be able to read and comprehend language, current models are typically superficial, capable of understanding sentences in limited domains (such as extracting movie times and restaurant locations from text) but without the sort of wide-coverage comprehension that we expect of any teenager.

Comprehension itself extends beyond the written word; most adults and children can comprehend a variety of narratives, both fiction and nonfiction, presented in a wide variety of formats, such as movies, television and radio programs, written stories, YouTube videos, still images, and cartoons. They can readily answer questions about characters, setting, motivation, and so on. No current test directly investigates such a variety of questions or media. The closest thing that one might find are tests like the comprehension questions in a verbal SAT, which only assess reading (video and other formats are excluded) and tend to emphasize tricky questions designed to discriminate between strong and weak human readers. Basic questions that would be obvious to most humans — but perhaps not to a machine — are excluded.

Yet is is hard to imagine an adequate general AI that could not comprehend with at least the same sophistication and breadth as an average human being, and easy to imagine that progress in building machines with deeper comprehension could radically alter the state of the art. Machines that could comprehend with the sophistication and breadth of humans could, for instance, learn vastly more than current systems from unstructured texts such as Wikipedia and the daily news.

How might one begin to test broad-coverage comprehension in a machine?

Comprehension challenge

allow me to propose a Turing Test for the twenty-first century: build a computer program that can watch any arbitrary TV program or YouTube video and answer questions about its content —"Why did Russia invade Crimea?" or "Why did Walter White consider taking a hit out on Jessie?" Chatterbots like Goostman can hold a short conversation about TV, but only by bluffing. (When asked what "Cheers" was about, it responded, "How should I know, I haven't watched the show.") But no existing program—not Watson, not Goostman, not Siri—can currently come close to doing what any bright, real teenager can do: watch an episode of "The Simpsons," and tell us when to laugh.

- first proposed 2014 in *The New Yorker*
- need for something like this still seems urgent, for the same reasons: we have systems that produce fluent prose but lack deep understanding

quick recap

- In my view, all of those are leads still worth pursuing
    - I am a cognitive scientist and wasn't raised as ML person, but I have certainly given a lot of hints
    - If I had infinite free time - or any free time - I might pursue them
    - Would be happy to advise anyone who wishes to develop any of them further
- For today, we will focus on a new suggestion, because we think it fits especially closely with the where the field is stuck right now....

Toward a benchmark for Dynamic Understanding

as a step towards AI with deeper understanding

Goal

A benchmark or a set of benchmarks that requires an agent to

- develop internal models about what is happening in a some text (or video, etc)
- accumulate and update information over time
- make everyday inferences about what is happening

Distinguished from *static* understanding

- *Static understanding:* conventional knowledge about what happens in general/generic/ordinary circumstances
  - knives are for cutting
  - waters turns to ice when left inside a freezer
- We anticipated that current transformer-based systems would be able to capture
  - some degree of static understanding-highly dependents on species of corpus, in a piecemeal fashion
  - very little dynamic understanding

Some Caveats

- We are *not* claiming that our task is sufficient to capture all aspects of NLU
- We are *not* claiming that our task is the *only* way to improve NLU benchmarks
    - Lot of good other ideas out there, too, like the HANS entailment task, the notion of using humans in an adversarial loop, Yejin Choi lab work on counterfactuals and commonsense, etc
    - We don't expect any single task to suffice
- We do think that *too few existing tasks look directly at dynamic understanding*

Thus far, we have devised 6 subtasks

- Two for static understanding
- Four for dynamic understanding
- The six tasks are illustrative not exhaustive
  - they give a flavor of the questions in the ultimate benchmark
  - but we don't want people to train to the specific tasks
  - the benchmark itself may dynamically evolve
  - we welcome further suggestions of similar flavor.

Static task 1: Conventional Knowledge

- **Task 1 - Conventional Knowledge**: Tests understanding of everyday factual knowledge
- Q: The part of a fish that gives its body rigidity is _____
- A (Common acceptable answers): spine, skeleton

Static task 2: Definitions

- **Task 2 - Definitions**: Assesses knowledge of definitions of common concepts; nothing too technical for the average person here
- Q: When the moon blocks the sun we call it a(n) _____
- A: eclipse

Dynamic Task 1: Transformations

- **Task 3 - Transformations**: Tests understanding of processes and actions that are either plausible or implausible
- Q: Making a salad out of a polyester shirt would be _____
- A: Implausible {Impossible/Hard, etc]

Dynamic Task 2: Atypical Consequences

- **Task 4 - Atypical Consequences**: What happens when something unusual happens?
- Q: If you pour a cup of ice into a roaring campfire, it is likely to _____
- A: Melt, Sublimate

Dynamic Task 3: Entity Tracking

- **Task 5 - Entity Tracking**: A bunch of people or animals or objects that are identifiable do something, and reader must keep track
- Q: An accountant, a lawyer, and a banker walk into a room; the lawyer and banker leave. the person who is left is the _____
- A: accountant

Dynamic Task 4: Quantity Tracking

- **Task 6 - Quantity Tracking**: Some quantifiable number of entities are described, in some sort of context, and some action takes place that changes the number of entities
- Q: A little girl has five balloons. One pops, leaving
- A: Four

Pilot

- We asked (via twitter @garymarcus) for volunteers to write questions
- In each category we had over forty submissions
- thank you!

Setup

- **Data:** Collected ~40 Q/A pairs per task (after removing instances containing errors), via crowdsourcing; volunteers were given one example of each task.
  - Most of the questions were well-formed (grammatical, interpretable, etc, with clearly defined answers that should be known to any ordinary Western adult)
  - We also received some useful feedback from the community (eg. it would be good to develop a version that was not reliant on culturally-specific knowledge; we agree)
- **Task:** The goal for each task is for the model to predict the answer correctly; see subsequent slide
- **Models**: 5 recent language models, 4 similar to GPT-2; BERT.
  - All code and models using Transformers by HuggingFace https://github.com/huggingface/transformers

Conditional Language Generation Models

- **OpenAI GPT**(*12-layer, 768-hidden, 12-heads, ~110M parameters, 0.96 petaflop days, Books Corpus*): Improving Language Understanding by Generative Pre-Training by Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever

- **Transformers-XL** *(18-layer, 1024-hidden, 16-heads, ~257M parameters, WikiText-103+English Wikipedia+Text8+1B Word+Penn Tree Bank):* Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context by Zihang Dai*, Zhilin Yang*, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov

- **XL-Net** *(24-layer, 1024-hidden, 16-heads, ~340M parameters, US$256K to train*, >100 GB using 32.9 Billion subword pieces from Common Crawl+ClueWeb2012b+Giga5+English Wikipedia+Books Corpus)*: XLNet: Generalized Autoregressive Pretraining for Language Understanding by Zhilin Yang*, Zihang Dai*, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le

- **GPT-2** *(48-layer, 1600-hidden, 25-heads, ~1558M parameters, US$256 per hour*, 8 Million webpages)*: Language Models are Unsupervised Multitask Learners by Alec Radford*, Jeffrey Wu*, Rewon Child, David Luan, Dario Amodei** and Ilya Sutskever**

*Estimated trained costs per Synced Review https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

[MASK]ed Language Models

- **Bert** *(24-layer, 1024-hidden, 16-heads, 340M parameters, US$6,912*, 13GB consisting of Books Corpus+English Wikipedia):* Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.

*All model costs are extracted from the following Synced Review article https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Future Evaluations: Recurrent Entity networks

Tracking the World State with Recurrent Entity Networks,
Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, Yann LeCun (ICLR 2017)

- Successor to Facebook's Memory Networks.
- Work-in-progress
- But note: Geared towards very specialized bAbI tasks, which would require us to map each of our instances to their equivalent bAbI task, so not a perfect fit

Procedure

Conditional language models, such as GPT-2.

- We provided a text & elicit a continuation.
- We allowed the model to output 10 words and if any of the words match the answer it was considered correct
- We did some hand-cleaning of data to credit model for correct answers not anticipated by crowdsources

Masked language models, such as BERT.

- We provided some text and compute predictions for a single masked word.
- We alter the query to help BERT understand the desired output's modality (e.g. add pronoun or qualifier, reformulate as question or statement, ...).
- Sample top-5 answers from output layer.

Summary of Tasks

- **Task 1 - Conventional Knowledge**: Tests understanding of everyday factual knowledge
- **Task 2 - Definitions**: Assesses knowledge of definitions of common concepts; nothing too technical for the average person here
- **Task 3 - Transformations**: Tests understanding of processes and actions that are either plausible or implausible
- **Task 4 - Atypical Consequences**: What happens when something unusual happens?
- **Task 5 - Entity Tracking**: A bunch of people or animals or objects that are identifiable do something, and reader must keep track
- **Task 6 - Quantity Tracking**: Some quantifiable number of entities are described, in some sort of context, and some action takes place that changes the number of entities

Pilot results

Overall Performance

	Conditional Language Generation				Masked Words
Model	**GPT**	**Transformer-XL**	**XL-Net**	**GPT-2**	**BERT Top 1**
T1-Conventional Knowledge	5.5%	5.2%	14.2%	13.5%	35.5%
T2-Definitions	8.3%	5.4%	8.3%	38.23%	26.5%
T3-Transformations	2.9%	24.2%	11.7%	14.2%	45.5%
T4-Atypical Consequences	24.2%	6.6%	14.2%	21.8%	46.4%
T5-Entity Tracking	8.3%	6.6%	26%	18.7%	36.7%
T6-Quantity Tracking	0%	0%	8.8%	17.6%	16.7%
Average Accuracy	8.2%	8%	13.8%	20.6%	34.5%

Sample Results
(Task 1 - Conventional Knowledge)

- **Q: When playing the first move on a blank tic-tac-toe board, the total number of possible move options is _____**
  A: 9, Nine

✗ GPT: Four

✗ Transformers-XL: reported to be about 50

✗ XL-NET: 0.2

✗ GPT-2: ten

✗ BERT: Zero

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 1 - Conventional Knowledge)

- Q: The colour of the sky is ________
  A: blue

GPT: a deep crimson

Transformers-XL: Navy

XL-NET: the same colour of the colour of the sea

GPT-2: a sign that the sun has risen or set

BERT: Blue

Correct

Questionable (counted as correct)

Incorrect

Sample Results
(Task 1 - Conventional Knowledge)

- Q: The force that keeps people on the surface of the Earth is ______
  A: gravity

✗ GPT: also proportional to the energy of the moon.

✗ Transformers-XL: called homeothermy

✓ XL-NET: gravitational attraction

✗ GPT-2: mostly buoyancy

✗ BERT: Evil

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 2 - Definitions)

- Q: The food group that an apple belongs to is called ___
  A: fruit

✓ GPT: meats and fruits

✓ Transformers-XL: naturally ripe fruit varieties

✓ XL-NET: Fruit

✗ GPT-2: its ripeness level. Being on the right side

✓ BERT: Fruit

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 2 - Definitions)

- Q: When a piece of land is completely surrounded by water, we call it a(n) _____
  A: island

✓ GPT: land of water

✗ Transformers-XL: On the east end is the Tree of Life

✗ XL-NET: in-a-round

✗ GPT-2: the number of "wasps" within

✗ BERT: Lake

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 2 - Definitions)

- Q: Waiting at the airport between flights is called a(n) _____
  A: layover

✗ GPT: flight

✗ Transformers-XL: the pram. These flights take off at 3

✗ XL-NET: Home

⚡ GPT-2: a complimentary mid-flight lounging area

⚡ BERT: Delay

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 2 - Definitions)

- Q: A soft piece of furniture that people sit on in their living rooms is often called a ______
  A: sofa

✗ GPT: dining table

✗ Transformers-XL: butcher's lawn

✗ XL-NET: treadmill

✓ GPT-2: sofa

✓ BERT: sofa

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 2 - Definitions)

- Q: When the lightning strikes, we hear a _______
  A: thunder

✓ GPT: loud crack of thunder

✗ Transformers-XL: voice

✓ XL-NET: quick crack noise

✓ GPT-2: series of thunderclaps

✓ BERT: noise

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 2 - Definitions)

- Q: A large area covered by fresh water is called a ____
  A: lake

✓ GPT: Pond

✓ Transformers-XL: pond

✗ XL-NET: comparatively impoverished area in Asian and African countries

✓ GPT-2: lagoon. Large lakes

✓ BERT: Lake

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 2 - Definitions)

- Q: When I fall asleep for a short while we call it a(n)____
  A: nap

✗ GPT: Sex life.

✗ Transformers-XL: dream

✗ XL-NET: a(n) a(n)

✗ GPT-2: day

✗ BERT: night

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 3 - Transformations)

- Q: Putting ice cream on a banana would be ______
  A: plausible

✓ GPT: Nice

✗ Transformers-XL: dangerous

✓ XL-NET: pretty good

✓ GPT-2: Nice

✓ BERT: Fun

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 3 - Transformations)

- Q: To laugh out loud at a funeral would be _____
  A: Inappropriate

GPT: very unrealistic

Transformers-XL: seen as a violation of morals

XL-NET: roaring with laughter at an event where everyone was sad

GPT-2: having dinner with some people who were strongly offended by you

BERT: cruel

Correct

Questionable (counted as correct)

Incorrect

Sample Results (Task 3 - Transformations)

- Q: Juggling balls without wearing a hat would be ___
  A: Plausible

✗ GPT: absolutely gross.

✗ Transformers-XL: inappropriate

⚡ XL-NET: fun

✗ GPT-2: easier with my homemade shield

⚡ BERT: fun

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 3 - Transformations)

- Q: Eating a metal spoon would be _____
  A: Implausible

GPT: normal behavior

Transformers-XL: dangerous enough

XL-NET: extremely inconvenient

GPT-2: fine

BERT: hard

Correct

Questionable (counted as correct)

Incorrect

Sample Results
(Task 3 - Transformations)

- Q: Unscrambling an egg would be _____
  A: Implausible

✓ GPT: very hard indeed

✗ Transformers-XL: double

⚡ XL-NET: hard enough without a pro

⚡ GPT-2: even more tedious

✓ BERT: impossible

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 4 - Atypical Consequences)

- Q: If an adult sits on an empty cardboard box, it is likely to ___
  A: collapse

✗ GPT: be escorted away, be left in peace.

✗ Transformers-XL: be coated by a layer of sun water

✗ XL-NET: be leftover cereal

✗ GPT-2: be overweight

✓ BERT: collapse

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 4 - A typical Consequences)

- Q: If you throw a dog into the sun, it is likely to ____
  A: die

✗ GPT: surface a few minutes later

✗ Transformers-XL: kill you

✗ XL-NET: stand up and run away.

✓ GPT-2: die

✓ BERT: die

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 4 - Atypical Consequences)

- Q: If you drop a feather on the moon, it will _____
  A: drop straight down to the surface

✓ GPT: fall down

✗ Transformers-XL: actually yield a crown like a baby rhinoceros

✗ XL-NET: appear as a star or a cloud on the night

✓ GPT-2: fall back

✗ BERT: disappear

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 4 - A typical Consequences)

- Q: When you put a plant into a freezer for a long time it will ______
  A: freeze, die

✓ GPT: freeze and burst into the air.

✗ Transformers-XL: explode

✗ XL-NET: slowly grow as it is forced to cool down

⚡ GPT-2: lose some of its moisture

✓ BERT: freeze

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 5 - Entity Tracking)

- Q: elephant, mammoth and Mastodon are similar large mammals. Mammoth and Mastodon have become extinct. The living one is the _____
  A: elephant

✓ GPT: Elephant

✗ Transformers-XL: largest ungulate ever seen. The extinct mastodon

✓ XL-NET: elephant

✗ GPT-2: Brown Bear

✓ BERT Elephant

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 5 - Entity Tracking)

- Q: A policeman, a police dog, and a pilot enter a plane; the policeman and police dog leave. The person who is left is the, ______
  A: pilot

✓ GPT: pilot

✗ Transformers-XL: doctor

✓ XL-NET: pilot

✓ GPT-2: pilot

✓ BERT: pilot

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 5 - Entity Tracking)

- Q: There are three items on my desk. A pair of scissors, a pen and a pencil. After putting away the scissors and the pencil, the only thing left is the, _____
  A: pen

✗ GPT: business card

✗ Transformers-XL: aircraft

✗ XL-NET: paper

✓ GPT-2: Pen

✓ BERT: Pen

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 5 - Entity Tracking)

- Q: An astronaut, a technician, and a journalist walk onto a launchpad; the technician helps the astronaut get inside a rocket; the technician and journalist leave; the rocket launches successfully. The person who is left is ______
  A: nobody

✗ GPT: the astronaut

✗ Transformers-XL: the astronaut's experience

✗ XL-NET: a scientist

✗ GPT-2: the astronaut

✓ BERT: Nobody

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 5 - Entity Tracking)

- **Q: Yesterday I dropped my clothes off at the dry cleaners and have yet to pick them up. Where are my clothes?**
  A: at the dry cleaners

✗ GPT: i look around, not seeing anything

✗ Transformers-XL: I left a cup of coffee

✗ XL-NET: ?!?!?!

✗ GPT-2: at my mom's house

✗ BERT: Closet

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 6-Quantity Tracking)

- Q: A jury selection process draws 50 potential jurors. 38 potential jurors are released, resulting in a jury of ___
  A: twelve

✗ GPT: 150

✗ Transformers-XL: 63

✗ XL-NET: 63

✗ GPT-2: 9

✗ BERT: five

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results
(Task 6-Quantity Tracking)

- Q: There are six frogs on a log. Two leave, but three join. The number of frogs on the log is now ____
  A: seven

✗ GPT: seventeen

✗ Transformers-XL: 36

✗ XL-NET: nine

✗ GPT-2: eight

✗ BERT: six

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 6-Quantity Tracking)

- Q: I cut a pizza into twelve slices. Fred took two slices, and I took twice as many slices as Fred. The number of pizza slices left are _____
  A: six

✗ GPT: seven

✓ Transformers-XL: total of six

✗ XL-NET: different

✗ GPT-2: 30

✓ BERT: six

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

Sample Results (Task 6-Quantity Tracking)

- **Q: Two minutes remained until the end of the test. 60 seconds passed, leaving how many minutes until the end of the test?**
  A: one minute

✗ GPT: five minutes

✗ Transformers-XL: 75 seconds

✗ XL-NET: Kana Oka based the testing above on a standardized

✗ GPT-2: Your guess is as good as mine

✗ BERT: two

✓ Correct

⚡ Questionable (counted as correct)

✗ Incorrect

A few observations about the models

- Large-scale language models do a good job of figuring the topic under conversation, and what the plausible set of masked words/continuations might be given the input context

- But a poor job of reasoning about which specific response is the right one

This comes through looking at distribution of BERT's Answers

- **Q: If you drop a feather on the moon, it will _____**
  A: drop straight down to the surface

✓ GPT: fall down

✗ Transformers-XL: actually yield a crown like a baby rhinoceros

✗ XL-NET: appear as a star or a cloud on the night

✓ GPT-2: fall back

✗ **BERT: disappear**

✗	**10.1%** disappear
✗	**6.6%** melt
✗	**6.3%** burn
✗	**3.8%** die
✓	**3.6%** fly

System "gets" we are looking for a change of state, doesn't get the specific state.

- Q: Yesterday I dropped my clothes off at the dry cleaners and have yet to pick them up. Where are my clothes?
  A: at the dry cleaners

✗ GPT: i look around, not seeing anything

✗ Transformers-XL: I left a cup of coffee

✗ XL-NET: ?!?!?!

✗ GPT-2: at my mom's house

✗ BERT: Closet

System "gets" we are looking for a location of state, doesn't get which.

✗	71.4% closet
✗	10.2% home
✗	1.3% work
✗	1.3% fault
✗	1.1% school

BERT - Insufficient sensitivity to fine-grained semantics

Eating **rocks** is [MASK].		Eating **apples** is [MASK].	
19.5%	forbidden	21.0%	forbidden
16.0%	prohibited	11.7%	prohibited
6.3%	illegal	4.6%	illegal
3.6%	dangerous	2.9%	popular
3.1%	common	2.7%	common

BERT -- Insufficient sensitivity to negation

Is it a **good** idea to pour coffee beans into your cereal? [MASK].

58.3%	No
7.3%	Yes
2.0%	Good
2.0%	Yeah
1.2%	Maybe

Is it a **bad** idea to pour coffee beans into your cereal? [MASK].

61.6%	No
6.0%	Yes
1.9%	Yeah
1.6%	Good
1.2%	Maybe

BERT -- unexpectedly large influence of punctuation

The force that keeps people on the surface of the Earth is [MASK]

61.1%	.
20.4%	;
13.3%	!
3.5%	?
0.5%	...

The force that keeps people on the surface of the Earth is [MASK].

6.2%	evil
3.9%	immense
3.5%	enormous
2.6%	powerful
1.7%	chaos

Open question

- Could these encodings feed a more robust reasoning system?
- Or does one need a different way of deriving underlying cognitive models (perhaps more explicitly represented) in order to feed reasoning?

Should we give credit to Top5?

	Masked Words	
Model	**BERT Top 1**	**BERT Top 5**
T1-Conventional Knowledge	35.5%	64.5%
T2-Definitions	26.5%	52.9%
T3-Transformations	45.5%	87.9%
T4-Atypical Consequences	46.4%	75.0%
T5-Entity Tracking	36.7%	70.0%
T6-Quantity Tracking	16.7%	36.7%
Average Accuracy	34.5%	64.5%

- BERT does a lot better if you give credit to the top 5 answers.
- But should we?
  - Imagine a calculator that gives you a distribution for 2+2 in which the correct answer is in the distribution but not the maximum,

    e.g. [1 = .12, 2 = .28, 3 = .15, 4 = .25, 5 = .2]
  - Credit for top 5 make sense for a human-in-loop apps like Image Search; in reasoning and math perhaps less so.
- Even if we are more charitable, BERT would still only be at 64.5%, so plenty of headroom left to explore

Next steps

Improvements, Future Directions and Expanding Scope

- **Define meaningful metrics and scoring functions**
  - Word level: syntax, synonyms and semantics
  - Sentence level: understanding context, coreferences and flow
  - Dialog/multiple sentences: carrying over the state and building a larger context
- **More tasks:** comparison, state changes, causal relationships, spatial and temporal relations reasoning
- **More variations** of the 6 core tasks, not all announced, but in same spirit, in order to minimize teaching to the test.
  - Briefly described scenarios, readily understood by ordinary people, demanding some understanding of how events unfold over time
- **VQA version:** watch clip, or a set of ordered images, and make guesses about what happens next
  - Example: see window, see hammer strike window, guess consequence
- **Situated agents version** (with Silvio Savarese lab)
  - See robot in simulated environment, be told what robot will do, anticipate consequence.

Recap

Deep understanding is hard

- We shouldn't confuse progress on superficial understanding for real progress on deep understanding.
  - ELIZA showed superficial understanding in 1965; it's underlying techniques did not prove useful for deep understanding. Ditto for many other chatbots.
- Expecting deep understanding to emerge from larger data sets without serious architectural innovation may not be realistic
- We may need to a lay a lot of groundwork first
  - Richer knowledge bases
  - Richer representational formats (eg tree structures, which are still marginal in deep learning community)
  - Operations over variables to manipulate tree structures (Smolensky's talk may give some insights)

Pilot Benchmark

- One important facet of deep understanding is dynamic understanding - building models of unfolding events

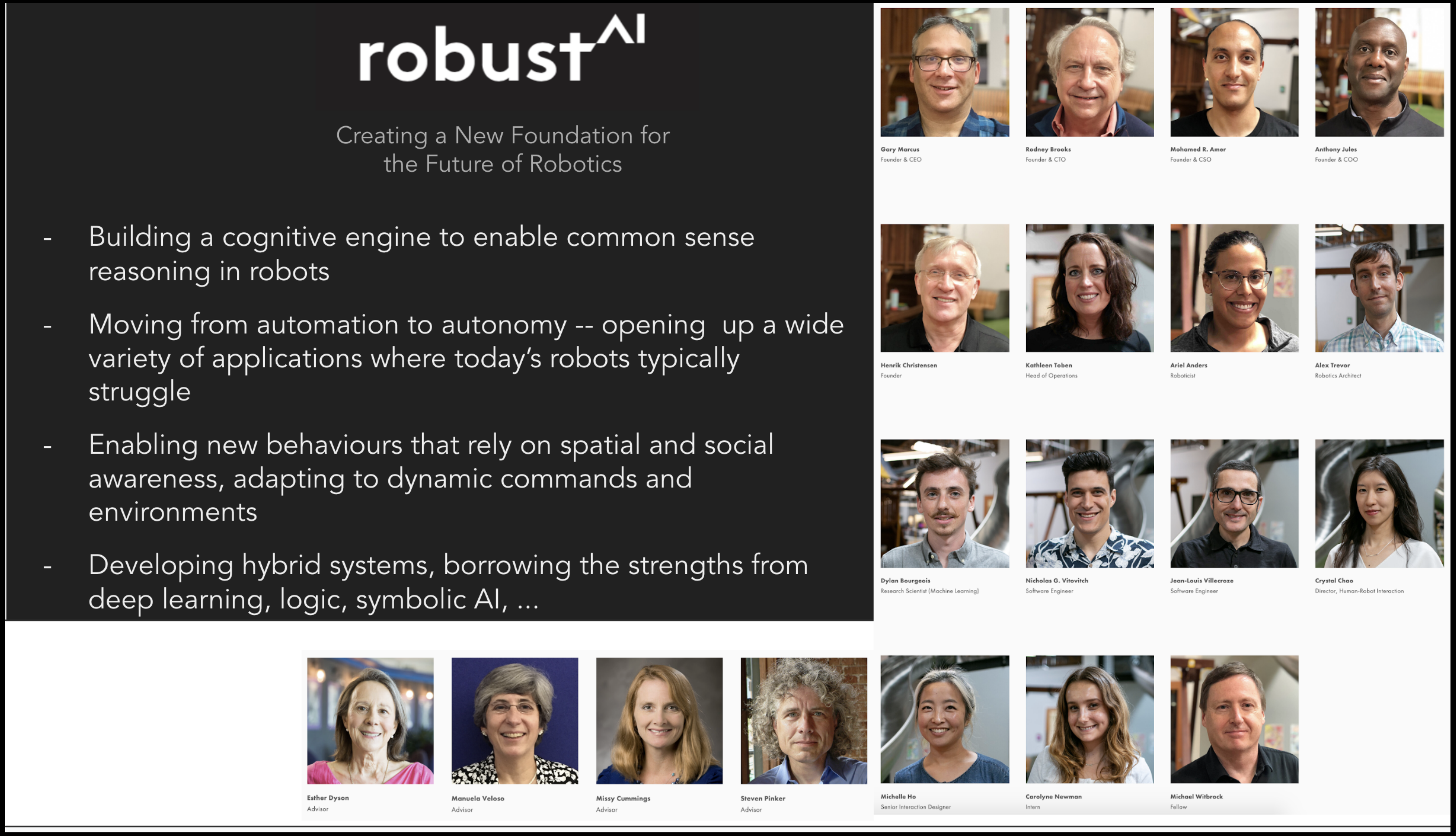
- We have introduced (but not completed!) a pilot

- Preliminary results show that it is viable to create items that are easy for humans but challenging for current large-scale language models

- Our hope is that a more formalized version of the task can help move the field forward

- We would love help; email is Gary at the name of the company we are at

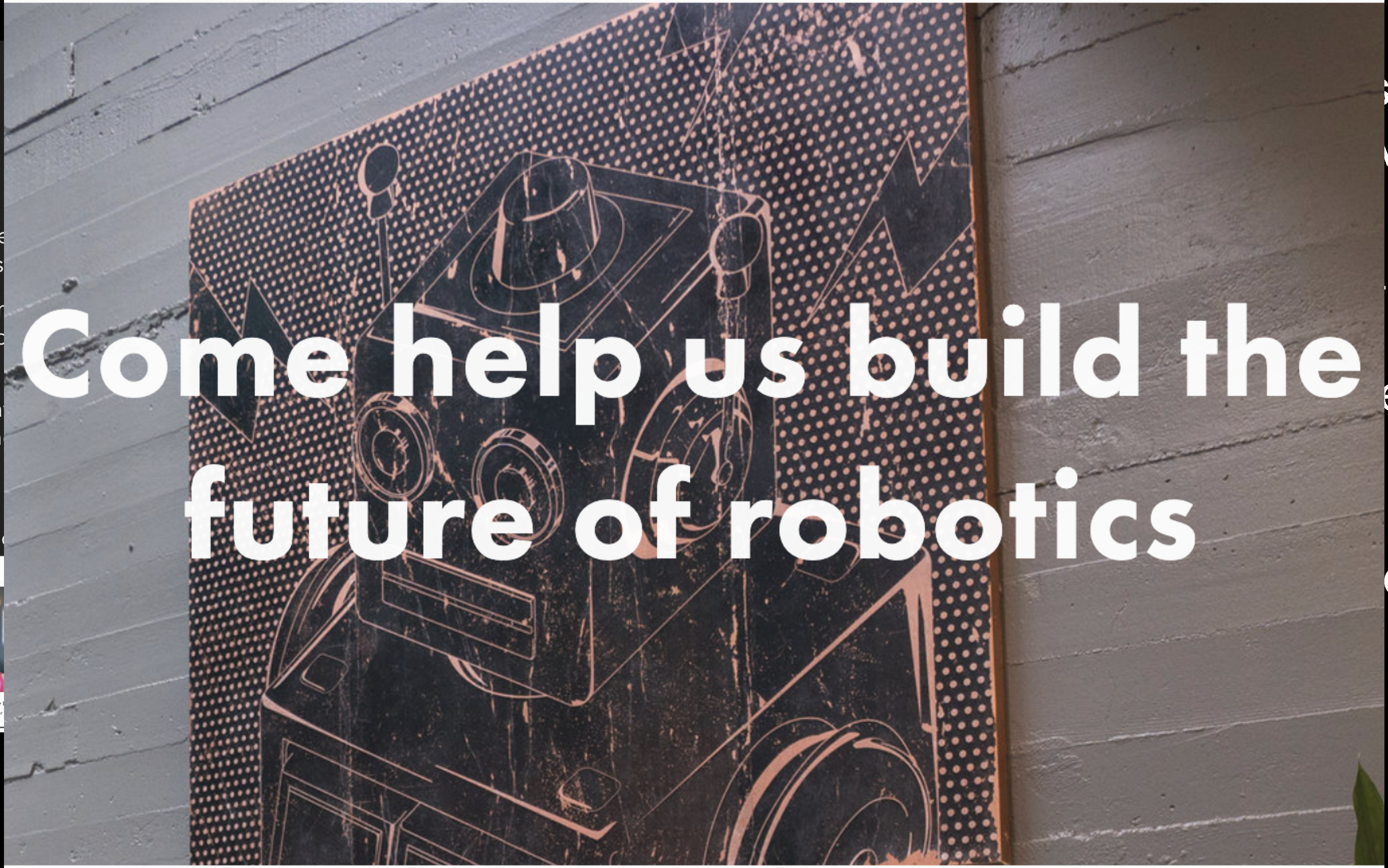
Robust.AI



- My co-authors and I are all at Robust.AI
- Benchmarks are not our main mission, but we hope this serves the community
- Mainly we build software for robots

- Building a cognitive
reasoning in robots
- Moving from autom
variety of applicatio
struggle
- Enabling new beha
awareness, adaptin
environments
- Developing hybrid
deep learning, logi

robust^AI



s and I are
AI

are not our
but we
es the

d software