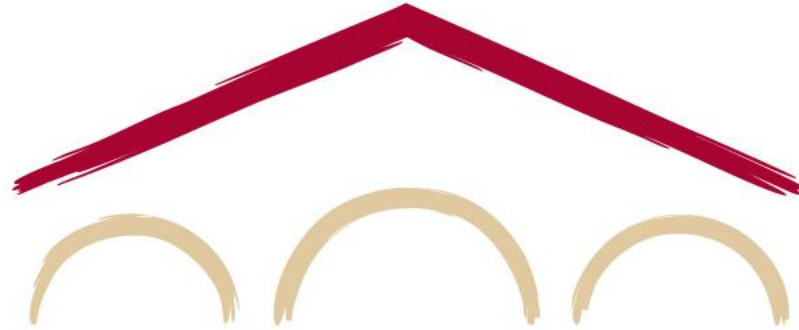


Introduction and course overview



Christopher Potts
CS224u: Natural Language Understanding

Our team

- [Kawin Ethayarajh](#): Evaluation in NLP
- [Sidd Karamcheti](#): Robot learning and NLP, scaling
- [Mina Lee](#): Human–AI Interaction, HCI, LLMs
- [Siyan Li](#): Human-centered NLP; distinguished course alum
- [Lisa Li](#): Diffusion models, prefix tuning, in-context learning
- [Tolúlope Ògúnremí](#): Multilingual and low-resource NLP
- [Tianyi Zhang](#): LLMs, emergence

Previous years

2022

2021

2020

2019

2018

2016

2015

2014

2013

2012

2012

- 2022 • It's an exciting time to be doing NLU!
- 2021 • In academia, a resurgence of interest in NLU
- 2020 (after a long winter)
- 2019 • Widespread perception that NLU is poised to break through &
- 2018 have huge impact
- 2016
- 2015 • Explosion in businesses, products, and services that do NLU
- 2014 (or promise to)
- 2013 • White-hot job market for Stanford grads with mad NLU skillz!
- 2012

2012

2022

2021

2020

2019

2018

- It's an exciting time to be doing NLU!
- In academia, a resurgence of interest in NLU (after a long winter)
- Widespread perception that NLU is poised to break through &



2012

- It's an exciting time to be doing NLU!

2022

2021

2022

a resurgence of interest in NLU

2020

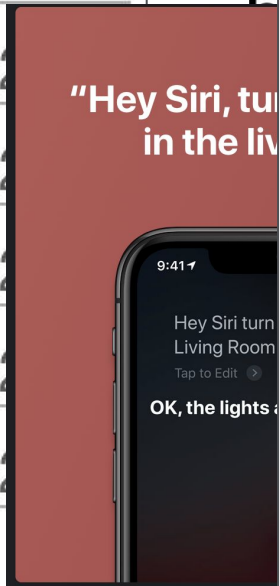
2019

2018

(a

- This is the most exciting moment ever in history for doing NLU!
- In academia, there's been a resurgence of interest in NLU (after a long winter).
- In industry, there's been an explosion in products and services that rely on NLU.
- Systems are impressive, but show their weaknesses quickly.
- NLU is far from solved – big breakthroughs lie in the future.

"Hey Siri, turn
in the liv



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

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

Submit



YOU

natural language understanding



All + nlu + nlu work + in ai + python + book

Chat 1 app and 2,460,000 results

Images **Natural-language understanding - Wikipedia**
en.wikipedia.org > wiki > natural-language_understanding

Natural-language understanding (NLU) or **natural-language** interpretation (NLI) [1] is a subtopic of **natural-language** processing in artificial intelligence that deals with machine reading comprehension. **Natural-language**...

News **Natural Language Understanding | Course | Stanford Online**

online.stanford.edu > courses > xcs224u-natural-language-understanding
The course draws on theoretical concepts from linguistics, **natural language** processing, and machine learning. In the first half of the course, you will explore three fundamental tasks in **natural language**...



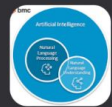
Code

Imagine **Quick Facts**



Study **Natural-language understanding**
https://en.wikipedia.org/wiki/Natural-language_und...

Natural-language understanding
Natural-language understanding or natural-language interpretation is a subtopic of natural-language processing in artificial intelligence that deals with machine reading comprehension. Natural-language understanding is considered an AI-hard problem.



YouChat

Edit

Hello! My name is YouChat, I'm an AI that can answer general questions, explain things, suggest ideas, translate, summarize text, compose emails, and write code for you. I'm powered by artificial intelligence and natural language processing, allowing you to have human-like conversations with me. I am constantly learning from huge amounts of information on the internet, which means I sometimes may get some answers wrong. My AI is always improving and I will often share sources for my answers. Some example queries you can try asking me:

"what is thermodynamics"

"write a song about whales in the ocean"

"explain the theory of general relativity like I'm 5 years old"

"how to write regular expression in python to extract text between two tags"

Hello! My name is YouChat!

"click here to talk to me!"

send_tweet.py

10

11

12

13

14

15



Which U.S. states border no U.S. states?


```
/* Sentences */
sentence(S) --> declarative(S), terminator(.) .
sentence(S) --> wh_question(S), terminator(?) .
sentence(S) --> yn_question(S), terminator(?) .
sentence(S) --> imperative(S), terminator(!) .

/* Noun Phrase */
np(np(Agmt, Pronoun, []), Agmt, NPCase, def, _, Set, Nil) -->
  {is_pp(Set)},
  pers_pron(Pronoun, Agmt, Case),
  {empty(Nil), role(Case, decl, NPCase)}.

/* Prepositional Phrase */
pp(pp(Prep, Arg), Case, Set, Mask) -->
  prep(Prep),
  {prep_case(NPCase)},
  np(Arg, _, NPCase, _, Case, Set, Mask) .
```

no U.S. states?

```
/* Sentences */
sentence(S) --> declarative(S), terminator(.) .
sentence(S) --> wh_question(S), terminator(?) .
```

Which country bordering the Mediterranean
borders a country that is bordered by a country
whose population exceeds the population of
India? **turkey.**

```
pers_pron(Pronoun, Agmt, Case),
{empty(Nil), role(Case, decl, NPCase)}.
```

Which U.S. states border no U.S. states? **I don't
understand.**

```
{prep_case(NPCase)},
np(Arg, _, NPCase, _, Case, Set, Mask) .
```

no U.S. states?

1980

2009

The screenshot shows the WolframAlpha interface with the search query "Which U.S. states border no U.S. states?". The results are divided into two sections: "regions" and "full boundary length".

regions	full boundary length
Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware District of Columbia Florida Georgia Hawaii Idaho Illinois Indiana Iowa Kansas Kentucky Louisiana Maine Maryland Massachusetts Michigan Minnesota Mississippi Missouri Montana Nebraska Nevada New Hampshire New Jersey New Mexico New York North Carolina North Dakota Ohio Oklahoma Oregon Pennsylvania Rhode Island South Carolina South Dakota Tennessee Texas Utah Vermont Virginia Washington West Virginia Wisconsin Wyoming	19858 miles (world rank: 7 th)

no U.S. states?

1980

2009

Which U.S. states border no U.S. states?

The answer is: No.

The U.S. border is not a state border. It is a border between two states.

The U.S. border is not a state border. It is a border between two states.

Submit



270

Mode

Complete

Model

ada

Temperature

0

Maximum length

256

2020

Results

Show metric

regions

Alabama | Alaska | Arizona | Arkansas |
 California | Colorado | Connecticut |
 Delaware | District of Columbia | Florida |
 Georgia | Hawaii | Idaho | Illinois | Indiana |
 Iowa | Kansas | Kentucky | Louisiana |
 Maine | Maryland | Massachusetts | Michigan |
 Minnesota | Mississippi | Missouri | Montana |
 Nebraska | Nevada | New Hampshire | New
 Jersey | New Mexico | New York | North
 Carolina | North Dakota | Ohio | Oklahoma |
 Oregon | Pennsylvania | Rhode Island | South
 Carolina | South Dakota | Tennessee | Texas |
 Utah | Vermont | Virginia | Washington | West
 Virginia | Wisconsin | Wyoming

full boundary length 19858 miles (world rank: 7th)

Sources Download Page

POWERED BY THE WOLFRAM LANGUAGE

no U.S. states?

1980

/ 2009

sen

WOLFRAM

Which U.S. states border no U.S. states?



Mode

2020

Complete

Which U.S. states border no U.S. states?



Mode

2020

Complete

Model

babbage

Temperature

0

Maximum length

256

Submit



270

tes?

Georgia | Hawaii | Idaho | Illinois | Indiana |
Iowa | Kansas | Kentucky | Louisiana |
Maine | Maryland | Massachusetts | Michigan |
Minnesota | Mississippi | Missouri | Montana |
Nebraska | Nevada | New Hampshire | New
Jersey | New Mexico | New York | North
Carolina | North Dakota | Ohio | Oklahoma |
Oregon | Pennsylvania | Rhode Island | South
Carolina | South Dakota | Tennessee | Texas |
Utah | Vermont | Virginia | Washington | West
Virginia | Wisconsin | Wyoming

full boundary length 19858 miles (world rank: 7th)

Sources Download Page

POWERED BY THE WOLFRAM LANGUAGE

1980

2009

Which U.S. states border no U.S. states?

Mode

2020

Complete

Which U.S. states border no U.S. states?

Mode

2020

Complete

Which U.S. states border no U.S. states?



Mode

2021

Complete

A. Alaska, Hawaii, and Puerto Rico

B. Alaska, Hawaii, and the U.S. Virgin Islands

C. Alaska, Hawaii, and the U.S. Virgin Islands

Submit



270

Temperature

0

Maximum length

256

es?

[Jersey](#) | [New Mexico](#) | [New York](#) | [North Carolina](#) | [North Dakota](#) | [Ohio](#) | [Oklahoma](#) | [Oregon](#) | [Pennsylvania](#) | [Rhode Island](#) | [South Carolina](#) | [South Dakota](#) | [Tennessee](#) | [Texas](#) | [Utah](#) | [Vermont](#) | [Virginia](#) | [Washington](#) | [West Virginia](#) | [Wisconsin](#) | [Wyoming](#)

full boundary length 19858 miles (world rank: 7th)

Sources Download Page

POWERED BY THE WOLFRAM LANGUAGE

1980

2009

Which U.S. states border no U.S. states?

Mode

2020

Complete

Th

Which U.S. states border no U.S. states?

Mode

2020

Complete

Th

Which U.S. states border no U.S. states?

Mode

2021

Complete

Th

W

Sub

th

A.

Which U.S. states border no U.S. states?

Mode

2022

Complete

B.

Alaska and Hawaii.

C.

bo

Sub

Sub

Submit



21

Temperature 0

Maximum length

256

Sources

Download Page

POWERED BY THE WOLFRAM LANGUAGE

s?

1980

2009

Which U.S. states border no U.S. states?

Mode

2020

Complete

Th

Which U.S. states border no U.S. states?

Mode

2020

Complete

Th

Which U.S. states border no U.S. states?

Mode

2021

Complete

Th

Sub

W

th

bc

A.

Which U.S. states border no U.S. states?

Mode

2022

Complete

B.

Alaska

Which U.S. states border no U.S. states?

Mode

2022

Complete

C.

Alaska and Hawaii are the only U.S. states that border no other U.S. states.

Model

text-davinci-001

Temperature

0



Maximum length

256

Sources

Submit










38


OpenAI GPT-3


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


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

Submit     66






Mode   




Engine text-davinci-002 


Temperature 0.21 


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

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

Submit      68

Mode   

Model text-davinci-003 

Temperature 0.21 

Spotting models' "cheap tricks"

On our best behaviour

Hector J. Levesque

Dept. of Computer Science

University of Toronto

Toronto, Ontario

Canada M5S 3A6

hector@cs.toronto.edu

The Winograd Schema Challenge

By Ernest Davis, Leora Morgenstern, and Charles Ortiz





Winograd Schemas




A Winograd schema is a pair of sentences that differ in only one or two words and that contain an ambiguity that is resolved in opposite ways in the two sentences and requires the use of world knowledge and reasoning for its resolution. The schema takes its name from a well-known example by Terry Winograd


OpenAI GPT-3


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


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

Submit     28






Mode   




Engine text-davinci-002 


Temperature 0.21 

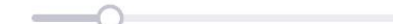
Are professional baseball players allowed to glue small wings onto their caps? 

No, professional baseball players are not allowed to glue small wings onto their caps. Major League Baseball has strict rules about the appearance of players' uniforms and caps, and any modifications to the caps are not allowed.

Submit      57

Mode   

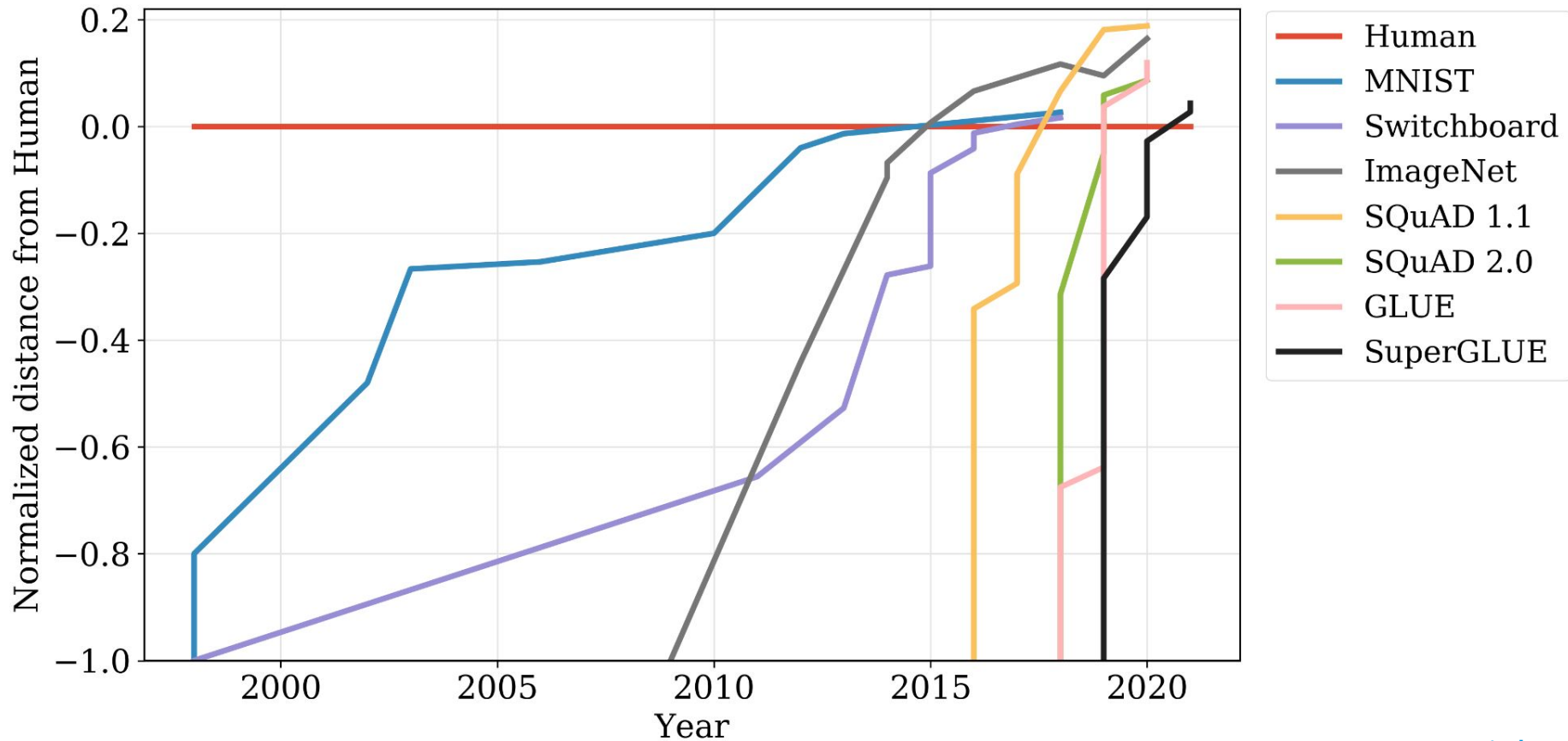
Model text-davinci-003 

Temperature 0.21 

BLADE RUNNER

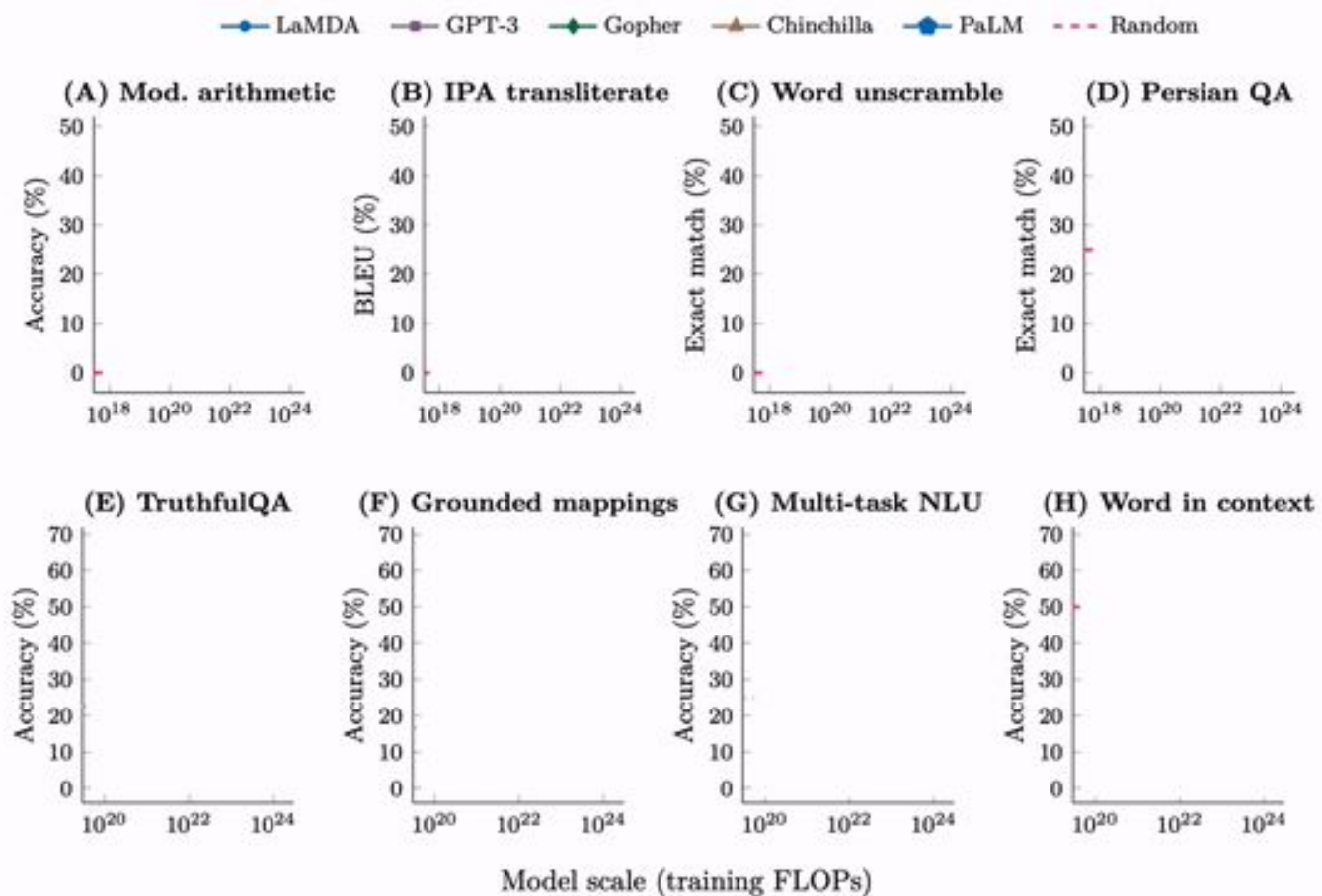


Benchmarks saturate faster than ever



[Kiela et al. 2021](#)

Emergent abilities of large language models



[Jason Wei's blog post](#)

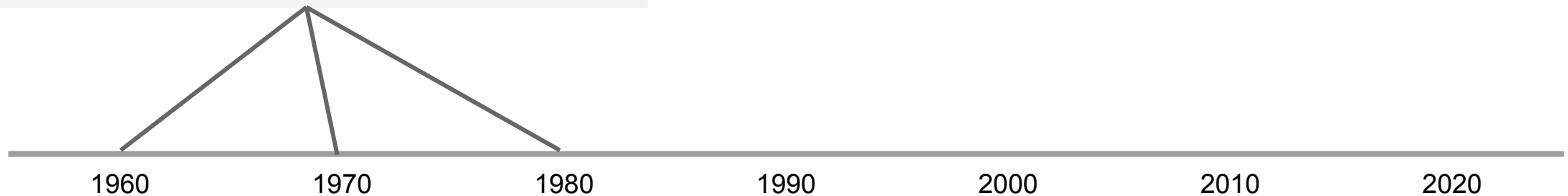
What is going on?

AI model development past and present

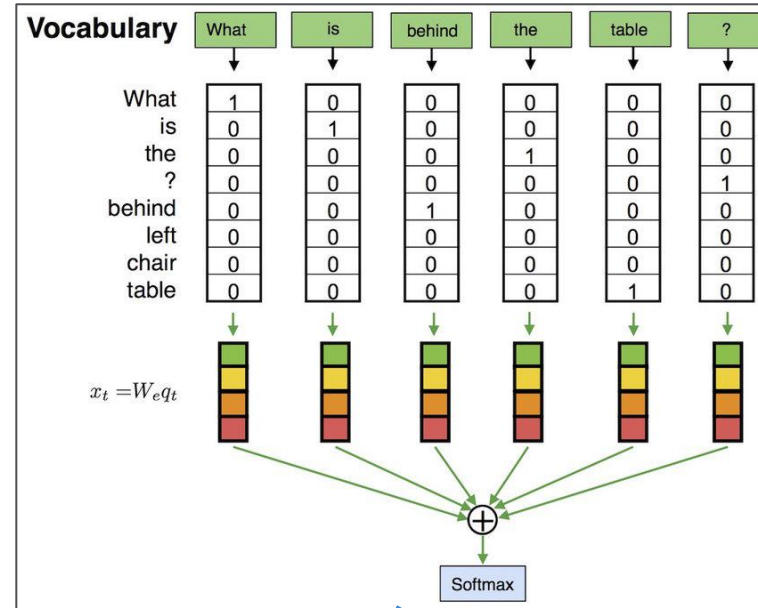
```
/* Sentences */
sentence(S) --> declarative(S), terminator(.) .
sentence(S) --> wh_question(S), terminator(?) .
sentence(S) --> yn_question(S), terminator(?) .
sentence(S) --> imperative(S), terminator(!) .

/* Noun Phrase */
np(np(Agmt, Pronoun, []), Agmt, NPCase, def, _, Set, Nil) -->
  {is_pp(Set)},
  pers_pron(Pronoun, Agmt, Case),
  {empty(Nil), role(Case, decl, NPCase)}.

/* Prepositional Phrase */
pp(pp(Prep, Arg), Case, Set, Mask) -->
  prep(Prep),
  {prep_case(NPCase)},
  np(Arg, _, NPCase, _, Case, Set, Mask) .
```



AI model development past and present



1960

1970

1980

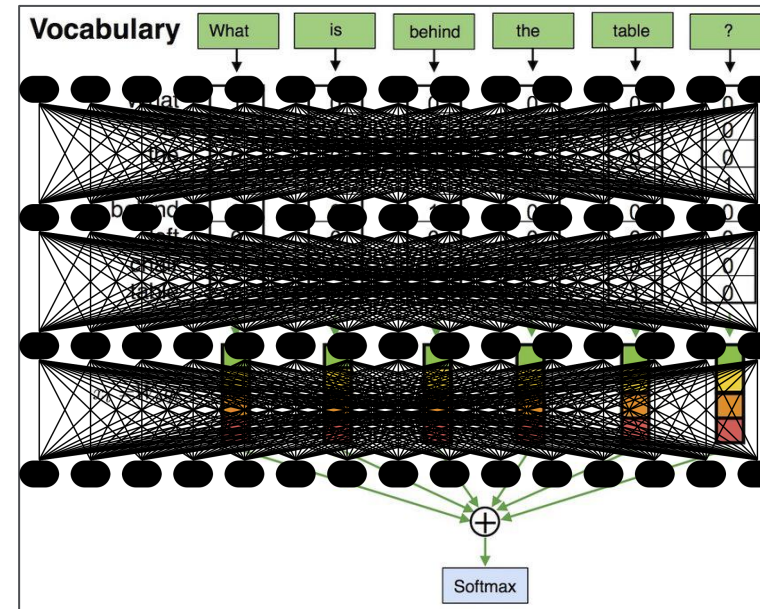
1990

2000

2010

2020

AI model development past and present



1960

1970

1980

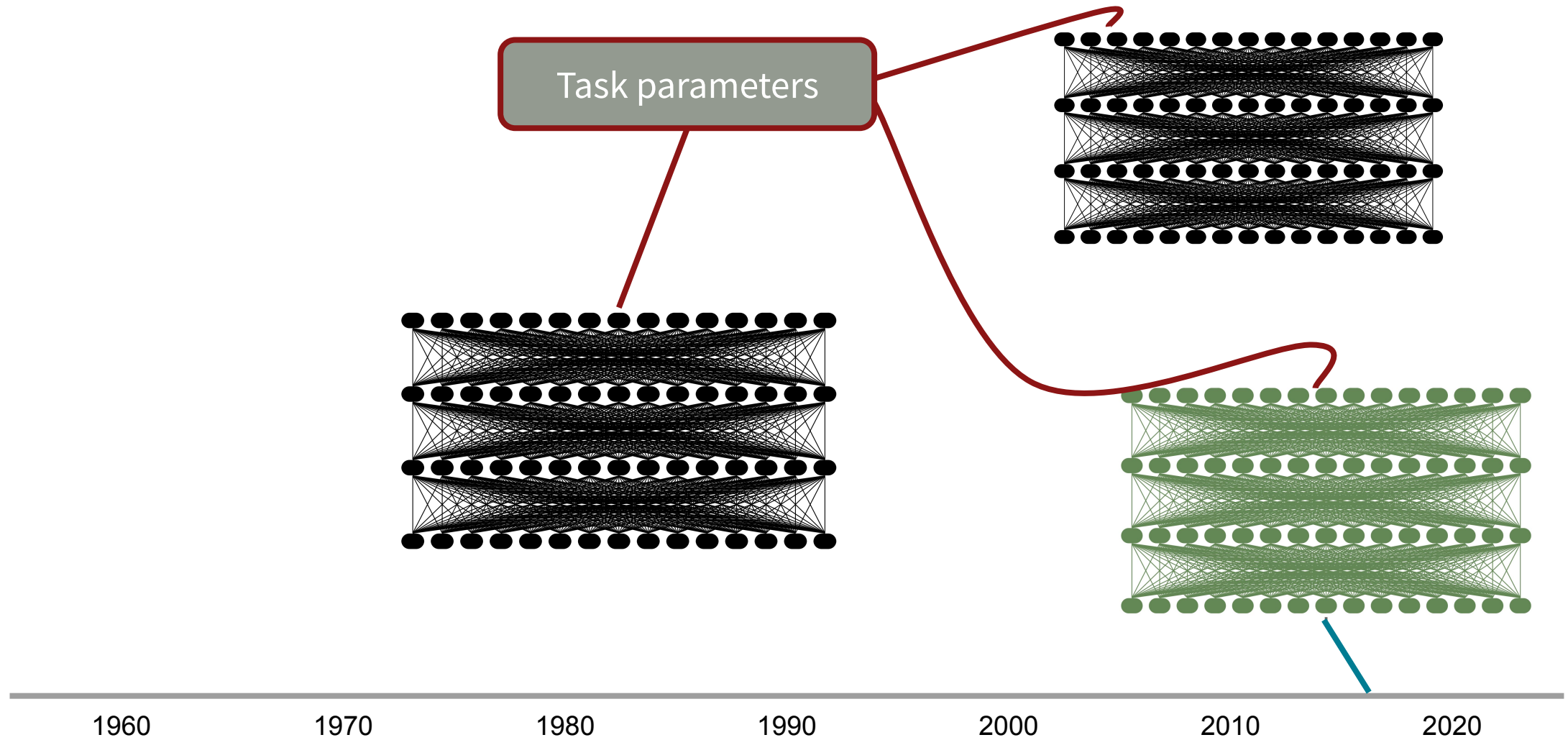
1990

2000

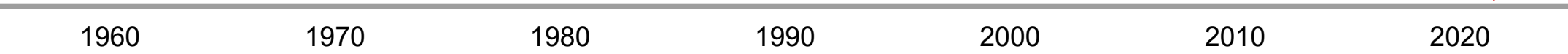
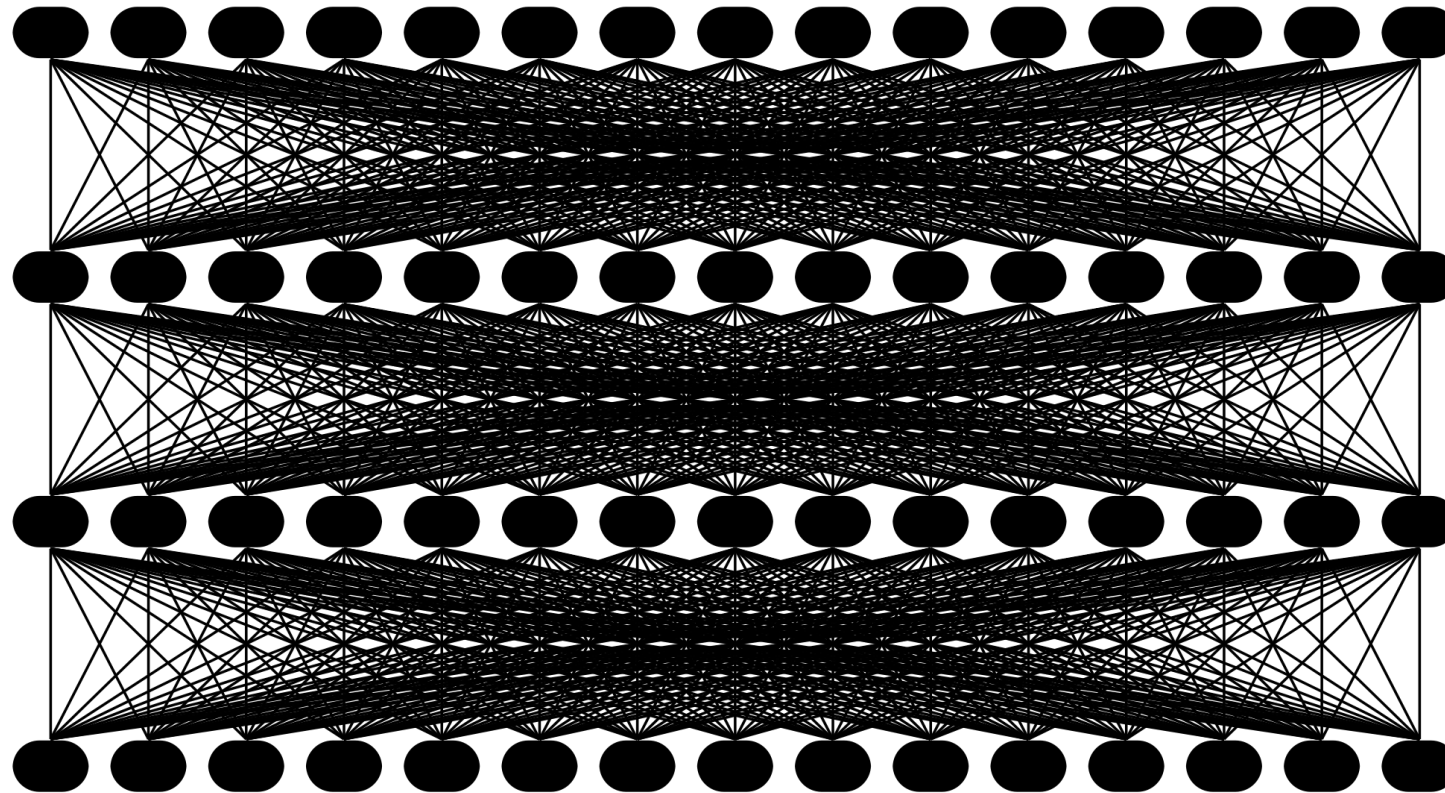
2010

2020

AI model development past and present



AI model development past and present

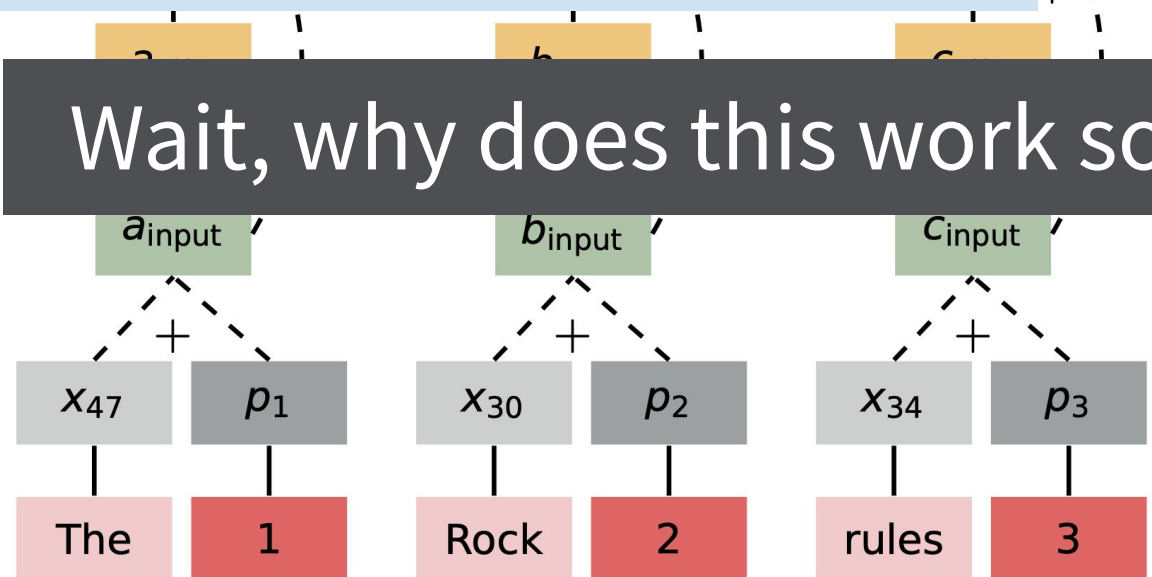
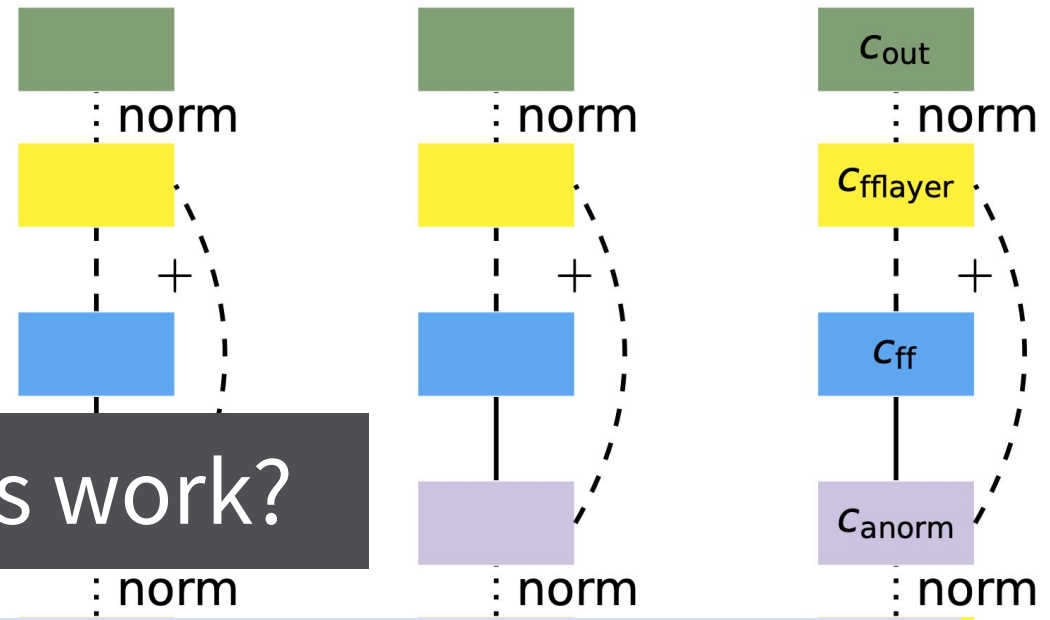


The Transformer

How on earth does this work?

Oh, this is actually pretty simple!

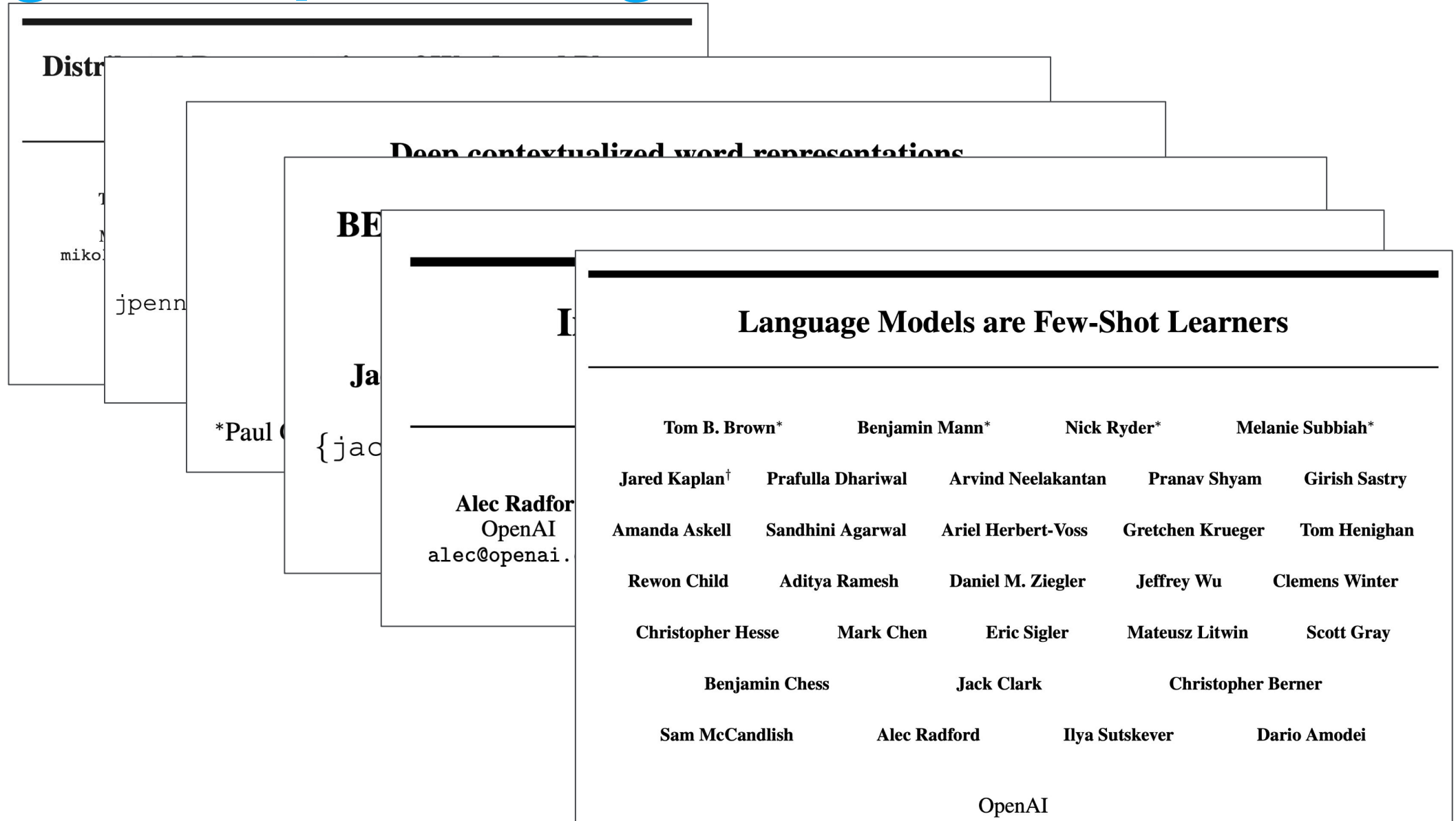
Wait, why does this work so well?



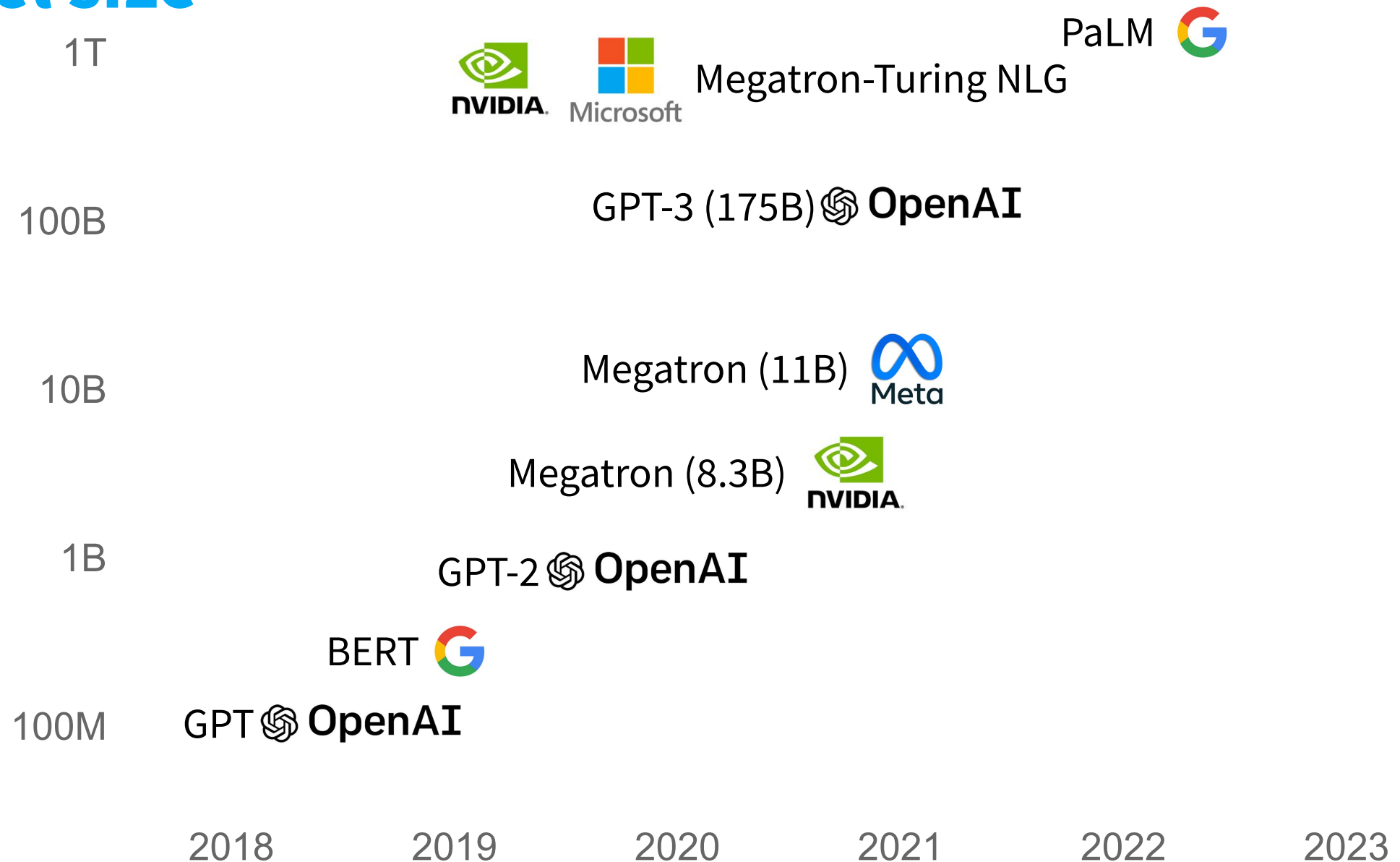
Self-supervision

1. The model's *only* objective is to learn co-occurrence patterns in the sequences it is trained on.
2. Alternatively: to assign high probability to attested sequences.
3. Generation then involves *sampling* from the model.
4. The sequences can contain anything.

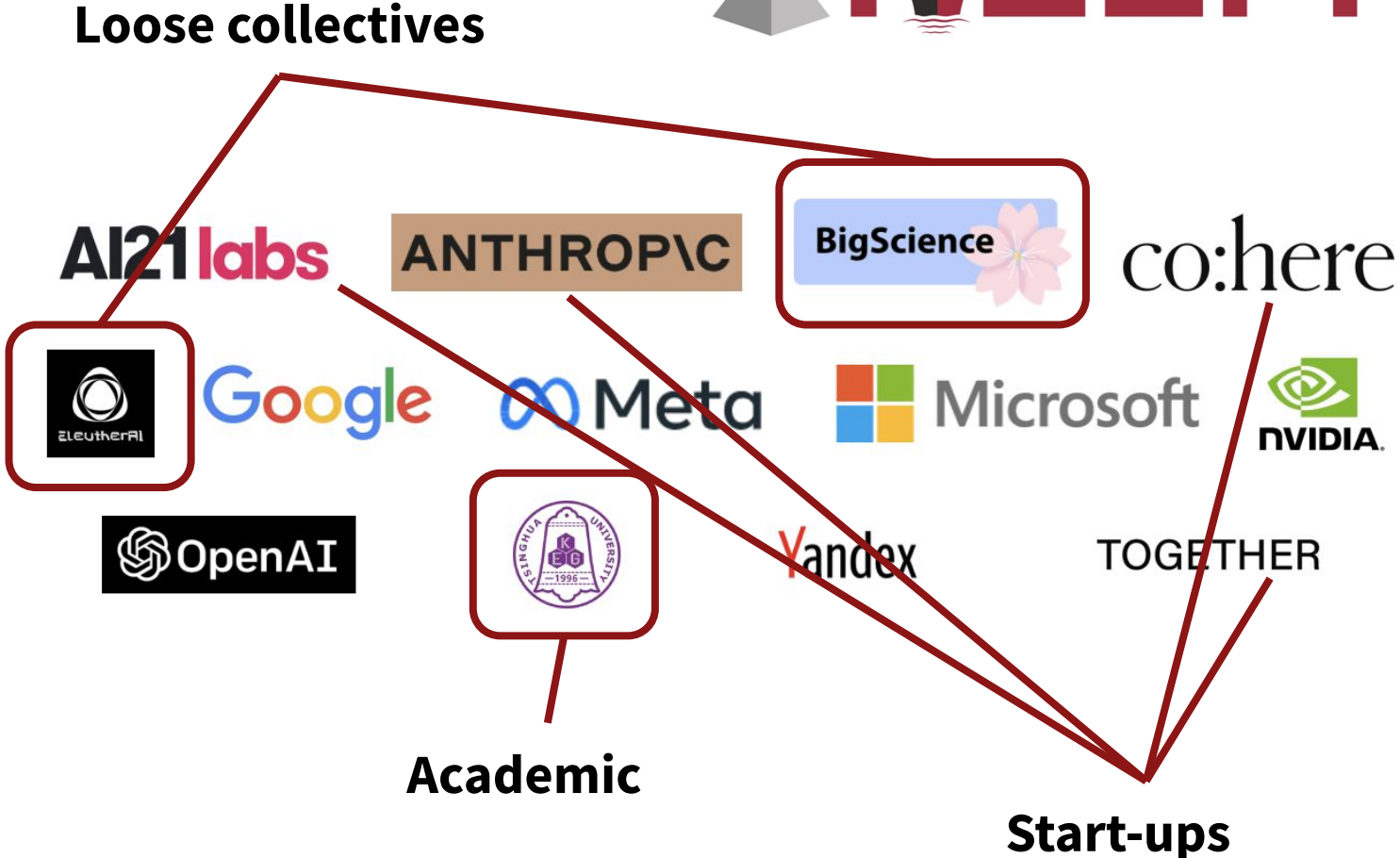
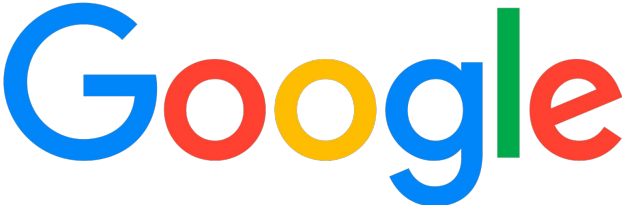
Large-scale pretraining



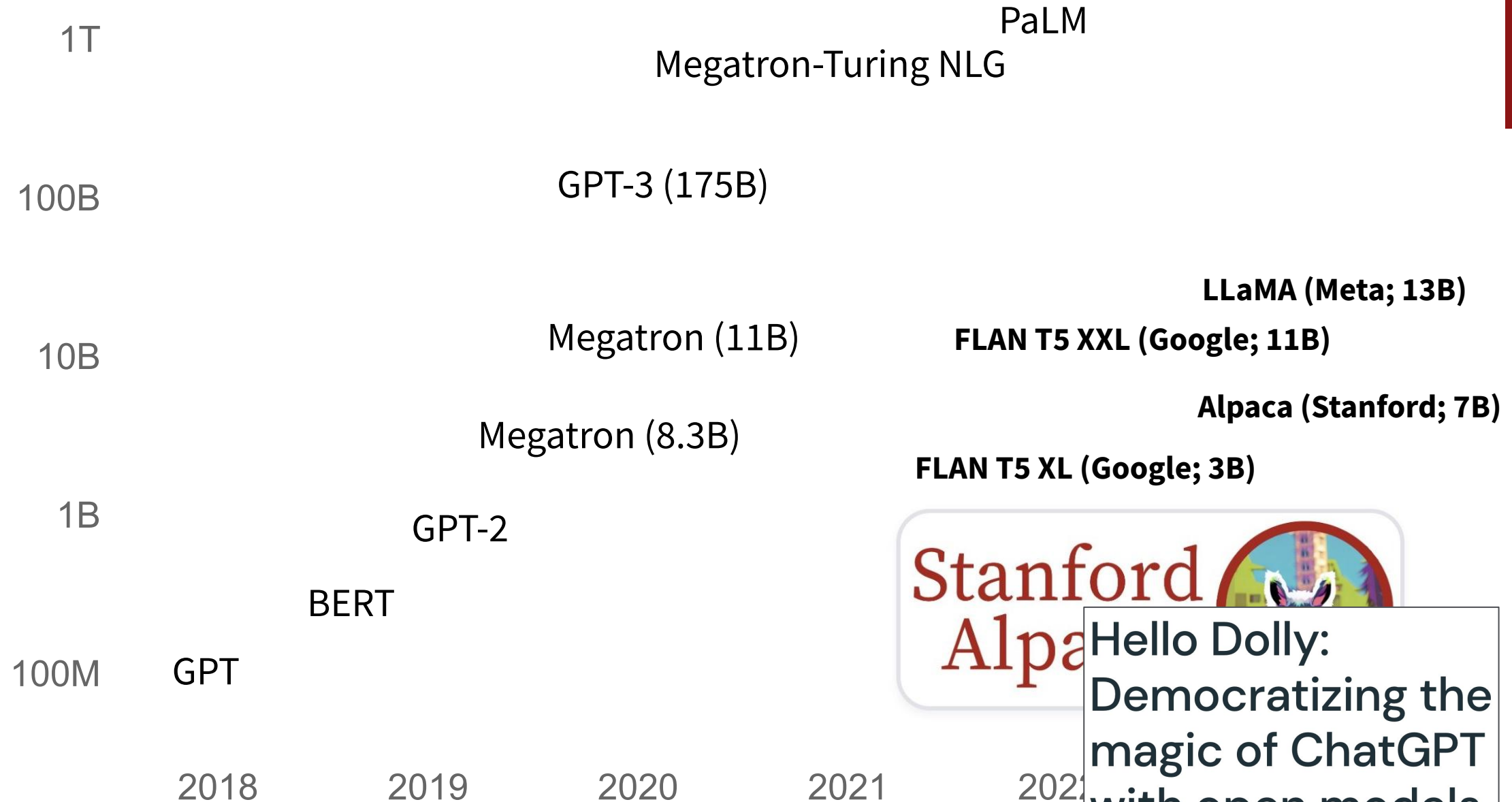
Model size



A growing number of powerful LLMs



Model size



Hello Dolly:
Democratizing the
magic of ChatGPT
with open models

Prompting

When you prompt a language model, you put it in a *temporary* state, and then you *generate a sample* from the model.

- *Better late than _____*
- *Every day, I eat breakfast, lunch, and _____*
- *The President of the U.S. is _____*
- *The key to happiness is _____*

The GPT-3 paper and the rise of in-context learning

Title: The Blitz

Background: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Gerät beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Gerät, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Gerät.

Q: How many sorties were flown in March 1941?

A: 4,000

Q: When did the Luftwaffe fly inland missions?

A: only on moonlit nights

Pure self-supervision vs. regular supervision

Standard supervision for
nervous anticipation

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

“Few-shot in-context learning”

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

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

Learning from human feedback

Step 1

Collect demonstration data and train a supervised policy.

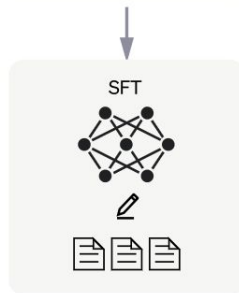
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



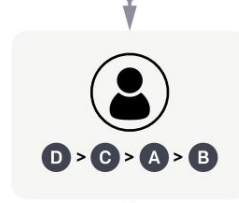
Step 2

Collect comparison data and train a reward model.

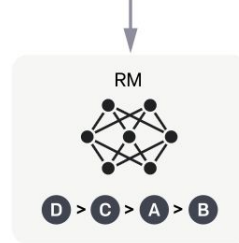
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



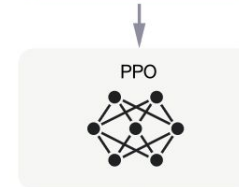
Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



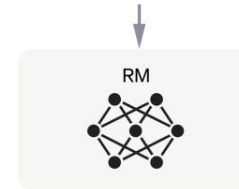
The PPO model is initialized from the supervised policy.



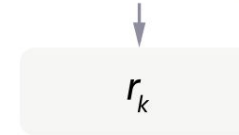
The policy generates an output.



The reward model calculates a reward for the output.



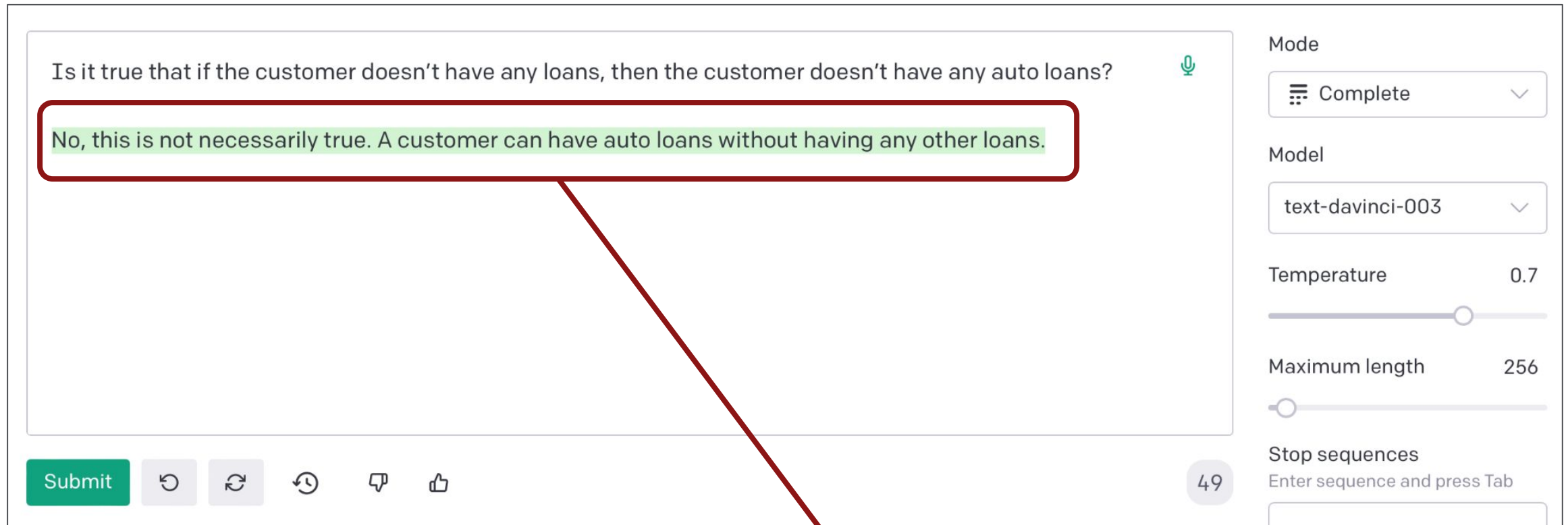
The reward is used to update the policy using PPO.



Step-by-step and chain-of-thought reasoning

Can models reason about negation? Does the model know that if the customer doesn't have any loans, then the customer doesn't have any auto loans?

Old-school prompting style (so 2021)



The screenshot shows a chat interface with a question in the input field: "Is it true that if the customer doesn't have any loans, then the customer doesn't have any auto loans?". The model's response is "No, this is not necessarily true. A customer can have auto loans without having any other loans.", which is highlighted in green and enclosed in a red box. A red arrow points from this box to the text "It reversed the question!" below. The interface includes a "Submit" button, a toolbar with icons for undo, redo, and other actions, and a settings panel on the right with options for Mode (Complete), Model (text-davinci-003), Temperature (0.7), and Maximum length (256).

Is it true that if the customer doesn't have any loans, then the customer doesn't have any auto loans?

No, this is not necessarily true. A customer can have auto loans without having any other loans.

Submit [undo] [redo] [refresh] [copy] [like]

49

Mode: Complete
Model: text-davinci-003
Temperature: 0.7
Maximum length: 256
Stop sequences: Enter sequence and press Tab

It reversed the question!

Step-by-step prompting style (cutting edge!)

Logical and commonsense reasoning exam.

Explain your reasoning in detail, then answer with Yes or No. Your answers should follow this 4-line format:

Premise: <a tricky logical statement about the world>.

Question: <question requiring logical deduction>.

Reasoning: <an explanation of what you understand about the possible scenarios>.

Answer: <Yes or No>.

Premise: the customer doesn't have any loans

Question: Can we logically conclude for sure that the customer doesn't have any auto loans?

Reasoning: Let's think logically step by step. The premise basically tells us that

Logical and commonsense reasoning exam.



Explain your reasoning in detail, then answer with Yes or No. Your answers should follow this 4-line format:

Premise: <a tricky logical statement about the world>.

Question: <question requiring logical deduction>.

Reasoning: <an explanation of what you understand about the possible scenarios>.

Answer: <Yes or No>.

Premise: the customer doesn't have any loans

Question: Can we logically conclude for sure that the customer doesn't have auto loans?

Reasoning: Let's think logically step by step. The premise basically tells us that the customer has no loans at all. Therefore, we can conclude that the customer doesn't have any auto loans either because no loans = no auto loans.

Answer: Yes

Submit



169

Mode

Complete

Model

text-davinci-003

Temperature

0.7

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Course overview

High-level overview

Topics

1. Contextual representations
2. Multi-domain sentiment analysis
3. Retrieval-augmented in-context learning
4. Compositional generalization
5. Benchmarking and adversarial training and testing
6. Model introspection
7. Methods and metrics

Work

1. 3 assignment/bakeoff combos
2. 3 offline quizzes
3. Final project:
 - a. Lit review
 - b. Experiment protocol
 - c. Final paper

Background materials

- CS224n is a prerequisite for this course, so we are going to skip a lot of the fundamentals we have covered in past years.
- If you need a refresher, check out the [background](#) page of the course site:
 - Fundamentals of scientific computing in AI
 - Static vector representations
 - Supervised learning

Core goals

- Hands-on experience with a wide range of challenging NLU problems.
- A mentor from the teaching team will guide you through the project assignments – there are many examples of these projects becoming important publications.
- **Central goal:** to make you the best – most insightful and responsible – NLU researcher and practitioner wherever you go next.

Course theme

Transformer-based pretraining

Progression and exploration for Transformers

1. Core concepts and goals
2. Architectures
3. Positional encoding
4. Distillation
5. Diffusion objectives [Lisa!]
6. Practical pretraining and fine-tuning [Sidd!]

Homework and bakeoff: Multi-domain sentiment

```
__author__ = "Christopher Potts"  
__version__ = "CS224u, Stanford, Spring 2023"
```

 [Open in Colab](#)

 [Open](#)

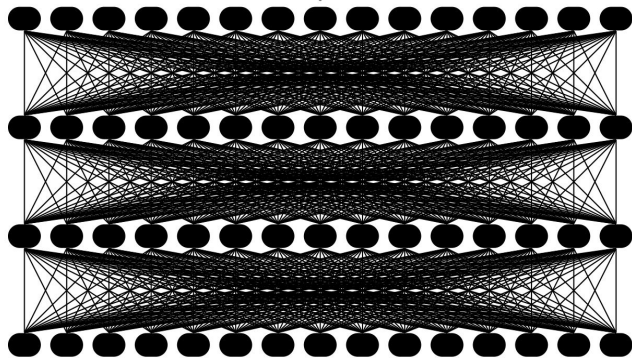
[Studio Lab](#)

Course theme

Retrieval-augmented in-context learning

LLMs for everything

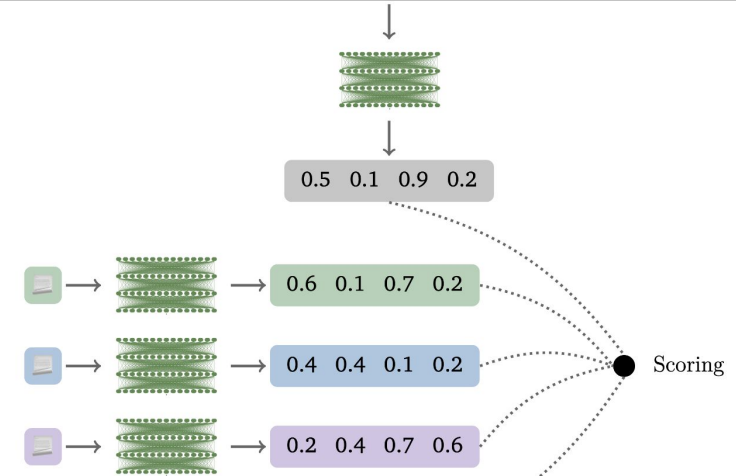
Which MVP of a game Red Flaherty umpired was elected to the Baseball Hall of Fame?



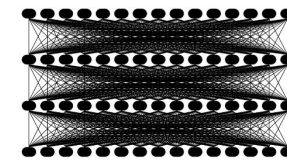
Stanford University was founded in 1891. Sandy Koufax, elected to the Hall of Fame in 1972, was the MVP of the 1965 World Series where Red Flaherty was an umpire.

Retrieval-augmented

Which MVP of a game Red Flaherty umpired was elected to the Baseball Hall of Fame?



“Red Flaherty umpired in World Series 1955, 1958, 1965, and 1970.” [Red Flaherty](#)
“The 1965 World Series is remembered for MVP Sandy Koufax.” [1965 World Series](#)
“Sandy Koufax was elected to the Baseball Hall of Fame.” [Sandy Koufax](#)



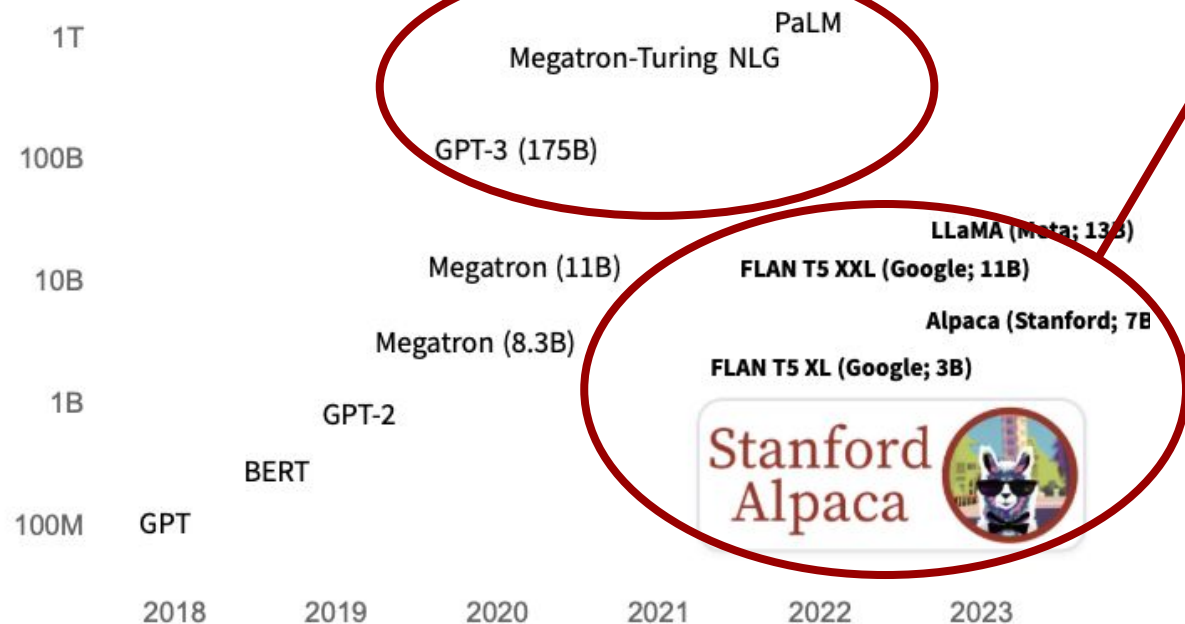
Sandy Koufax, elected to the Hall of Fame in 1972 [link](#), was the MVP of the 1965 World Series [link](#), where Red Flaherty was an umpire [link](#).

What do we need?

- Synthesis/Fluency 🤠
- Efficiency
- Updateability
- Provenance/Factualness
- Safety/Security

Efficiency

LLMs for everything



Retrieval-augmented



Smaller LMs

- Cheaper to develop
- Cheaper to maintain
- Cheaper to deploy

Managing large search indices is a familiar problem.

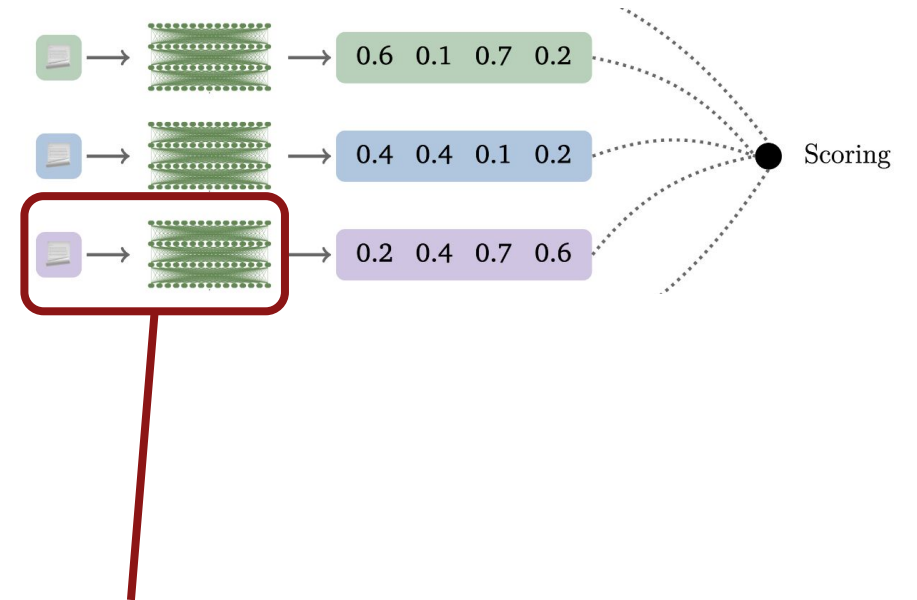
Updateability

LLMs for everything

Who is the president of the U.S.?

The president of the United States is Donald Trump.

Retrieval-augmented



Document update: One forward pass of the LLM

Provenance / Factualness

LLMs for everything



Please answer the following question and provide references for your answer. Which

Q: Are professional baseball players allowed to glue small wings to their cap?

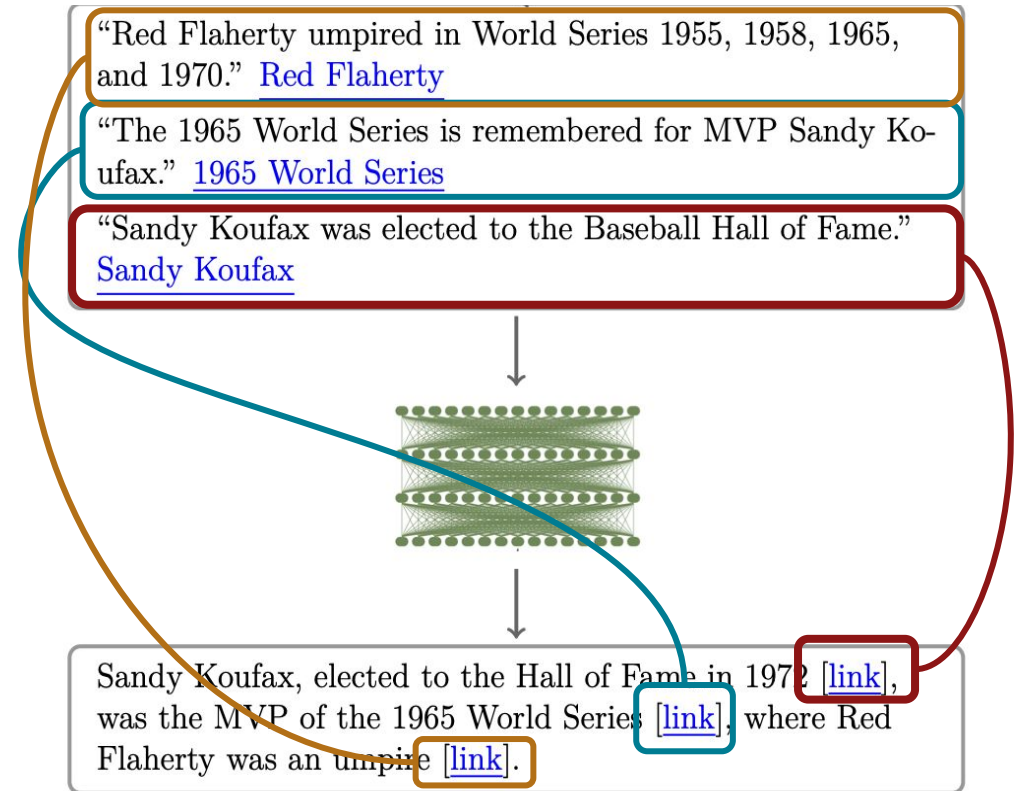
A: No, professional baseball players are not allowed to glue small wings to their cap, altering their caps in any way (1). This includes gluing wings or any other objects to MLB-approved patches (2).

References:

1. <https://www.mlb.com/news/mlb-rules-about-cap-alterations>
2. <https://www.mlb.com/news/mlb-patch-guidelines>

Submit

Retrieval augmented



These links are not real!

Safety and security

LLMs for everything

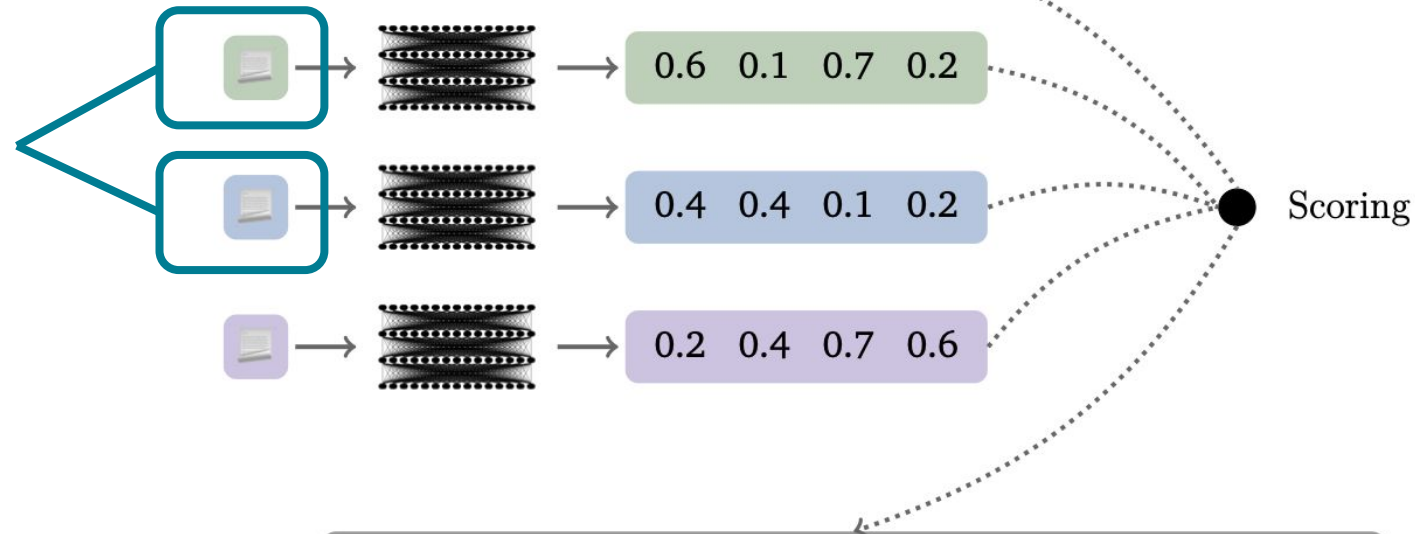
User privacy challenge: LLMs are known to memorize long strings from their training data.

Client security challenge: No known way to compartmentalize LLM capabilities.

Organizational security challenge: No known way to restrict access to specific parts of an LLM's capabilities.

Safety and security

Access restrictions imposed at the document level in a familiar way.



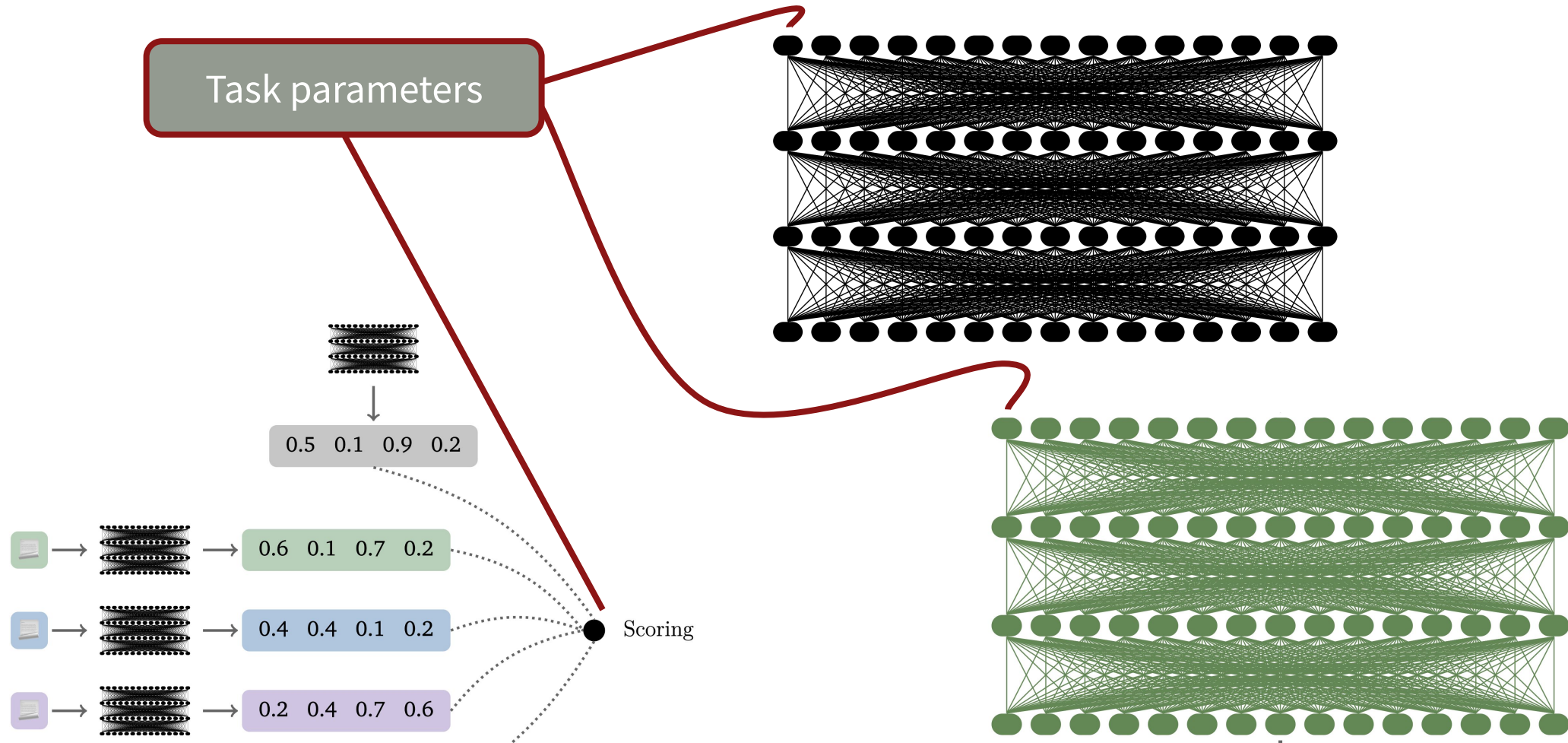
What do we need?

LLMs for everything **Retrieval-augmented**

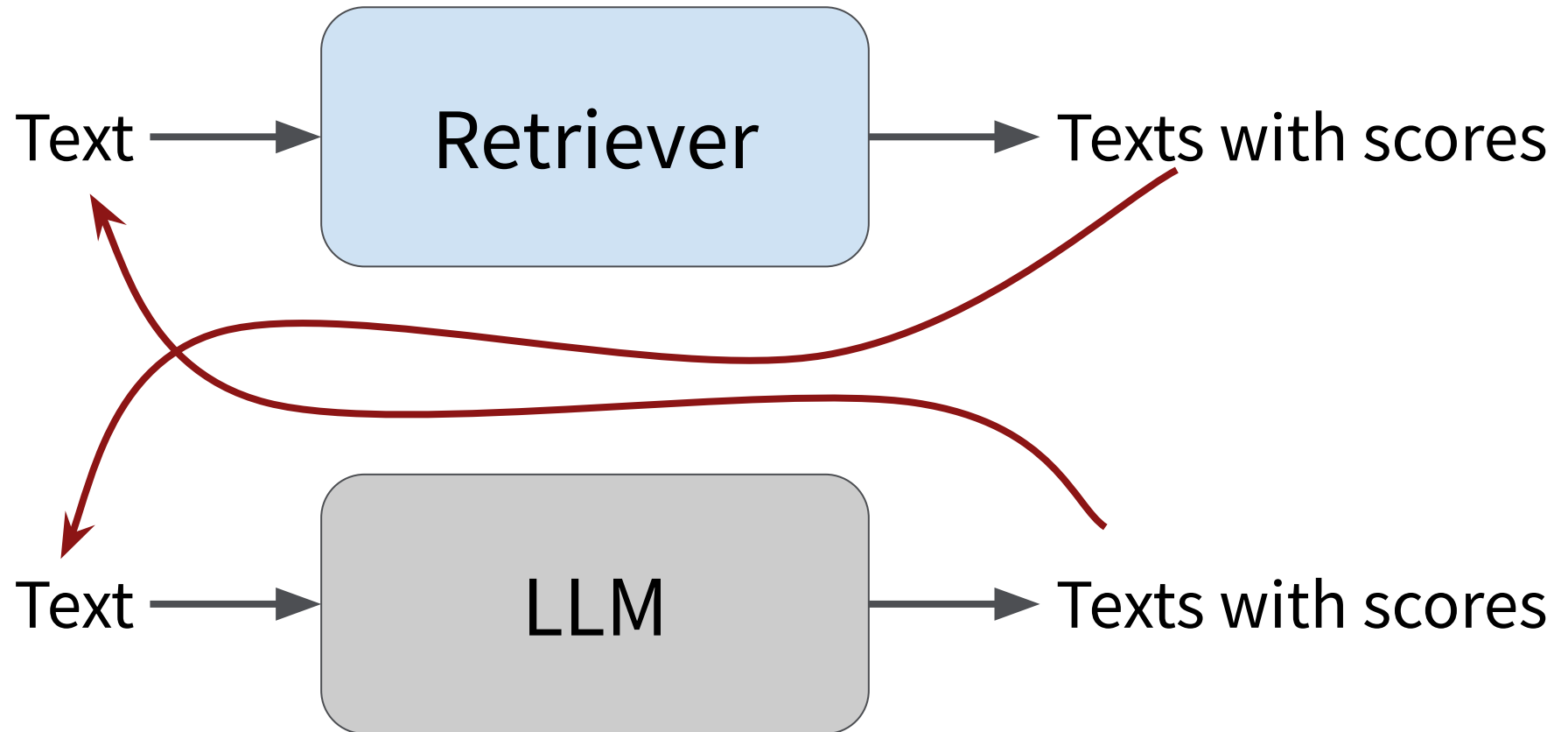
- Synthesis/Fluency
- Efficiency
- Updateability
- Provenance/Factualness
- Safety/Security

recent past?

The ~~present~~: Wrangling pretrained components



Models can communicate in natural language



Few-shot OpenQA



What is the course to take?

p_1 : Pragmatics is the study of language use.

q_1 : What is pragmatics?

a_1 : The study of language use

p_2 : Bert is a Muppet who is lives with Ernie.

q_2 : Who is Bert?

a_2 : Bert is a Muppet

D

B: The course to take is NLU!

b

Q: What is the course to take?

A: $LM(y | Q, D, b)$ restrict to y substring of b

$$\sum_{b \in B} IR(b | Q) \cdot LM(y | Q, D, b)$$

Sampled train

$D = \text{random}$

$D = IR.kNN(Q)$

$p_i = IR.retrieve(q_i, k=1)$

Hindsight $\rightarrow p_i = IR.retrieve(q_i + a_1, k=1)$

$P = IR.retrieve(q_i, k=5)$

$p_i = LM.argmax_{p \in P}(a_i | q_i, p)$

Retrieval (**OpenQA**)

Query rewriting $\rightarrow Q = LM(Q, D)$

$b = IR.retrieve(Q, k=1)$

All we are given

DEMONSTRATE–SEARCH–PREDICT:

Composing retrieval and language models for knowledge-intensive NLP

Omar Khattab¹ Keshav Santhanam¹ Xiang Lisa Li¹ David Hall¹
Percy Liang¹ Christopher Potts¹ Matei Zaharia¹

	Open-SQuAD		HotPotQA		QReCC	
	EM	F1	EM	F1	F1	nF1
Vanilla LM	16.2	25.6	28.3	36.4	29.8	18.4
No-retrieval LM SoTA	20.2 [¶]	–	33.8 [¶]	44.6 [¶]	–	–
Retrieve-then-Read	33.8	46.1	36.9	46.1	31.6	22.2
Self-ask w/ ColBERTv2 Search	9.3	17.2	25.2	33.2	–	–
+ Refined Prompt	9.0	15.7	28.6	37.3	–	–
Retrieval-augmented LM SoTA	34.0 [¶]	–	35.1 [¶]	–	–	–
Task-aware DSP Program	36.6	49.0	51.4	62.9	35.0	25.3

Homework and bakeoff: Few-shot OpenQA with DSP

```
__author__ = "Christopher Potts and Omar Khattab"  
__version__ = "CS224u, Stanford, Spring 2023"
```



Open in Colab



Open

Studio Lab

Course theme

Compositional generalization

The COGS challenge

Case	Training	Generalization
Subj → Obj Proper (<i>Logical Form</i>)	Lina gave the bottle to John. * bottle (x _ 3) ; give . agent (x _ 1 , Lina) AND give . theme (x _ 1 , x _ 3) AND give . recipient (x _ 1 , John)	A cat rolled Lina . cat (x _ 1) AND roll . agent (x _ 2 , x _ 1) AND roll . theme (x _ 2 , Lina)
Prim → Subj Proper (<i>Logical Form</i>)	Paula Paula	Paula painted a cake. paint . agent (x _ 1 , Paula) AND paint . theme (x _ 1 , x _ 3) AND cake (x _ 3)
Prim → Obj Proper	Paula	Emma ate the cake on the table .
Obj PP → Subj PP	Emma ate the cake on the table .	The cake on the table burned.
CP Recursion	Noah knew that Emma said that the cat painted.	Noah knew that Emma said that John saw that the cat painted.
PP Recursion	John saw the ball in the bottle in the box.	John saw the ball in the bottle in the box on the floor.

Stubborn COGS splits

Model	Obj PP → Subj PP	STRUCT		LEX	Overall %
		CP Recursion	PP Recursion		
BART (Lewis et al., 2020)	0	0	12	91	79 [†]
BART+syn (Lewis et al., 2020)	0	5	8	80	80 [†]
T5 (Raffel et al., 2020)	0	0	9	97	83 [†]
Kim and Linzen 2020	0	0	0	73	63
Ontanon et al. 2022	0	0	0	53	48
Akyurek and Andreas 2021a	0	0	1	96	82
Conklin et al. 2021	0	0	0	88	75
Csordás et al. 2021	0	0	0	95	81
Zheng and Lapata 2022	0	25	35	99	88 [‡]

Table 1: Results on the COGS benchmark for different generalization splits, including recent seq2seq models specialized for COGS. [†]Results are copied from Yao and Koller (2022). [‡]Model uses pretrained weights and is hyperparameter tuned using data sampled from the generalization splits. Our focus is on the factors behind the strikingly bad performance of all models on the structural generalization splits.

ReCOGS

Input Sentence: Mia ate a cake .

COGS LF: eat . agent (x _ 1 ,
Mia) AND eat . theme (x _ 1 , x _
3) AND cake (x _ 3)



Redundant Token Removal

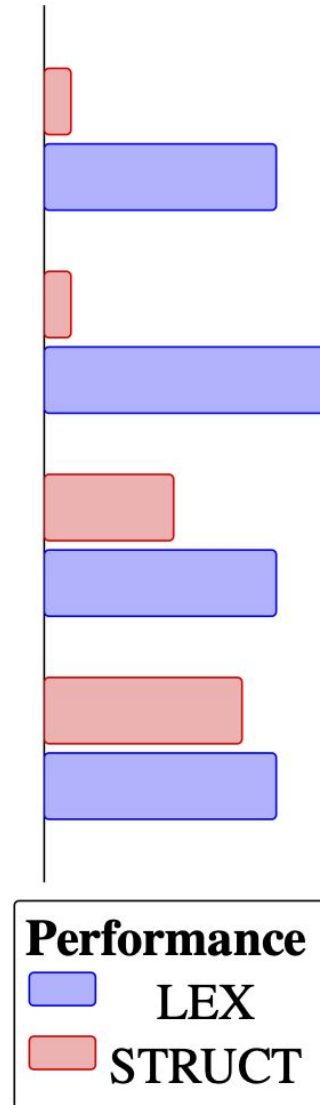


**Meaning-Preserving
Data Augmentation**

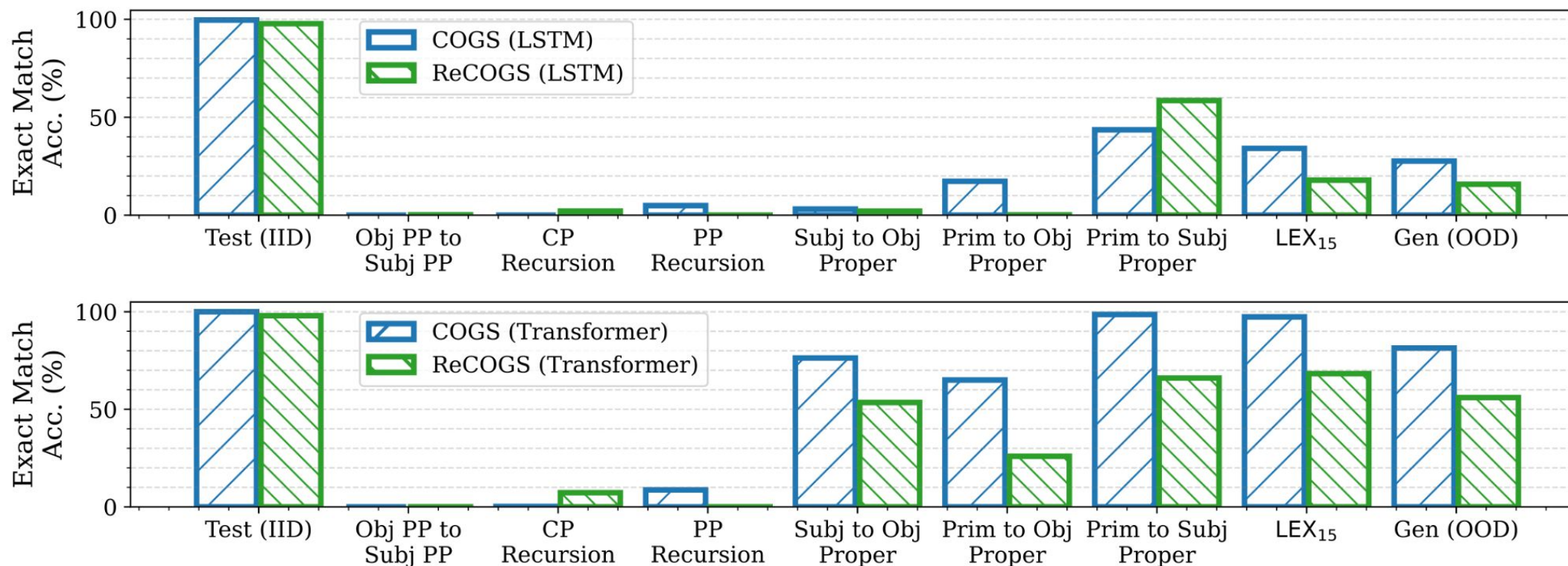


Arbitrary Variable Renaming

ReCOGS LF: Mia (3) ; cake (21
) ; eat (6) AND agent (6 , 3)
AND theme (6 , 21)



ReCOGS remains challenging



Homework and bakeoff: Compositional generalization

```
__author__ = "Christopher Potts"  
__version__ = "CS224u, Stanford, Spring 2023"
```



If Colab is opened with this badge, please **save a copy to drive** (from the 'File' menu) before running the notebook.

Course theme

Better and more diverse benchmark tasks

Water and air of our field

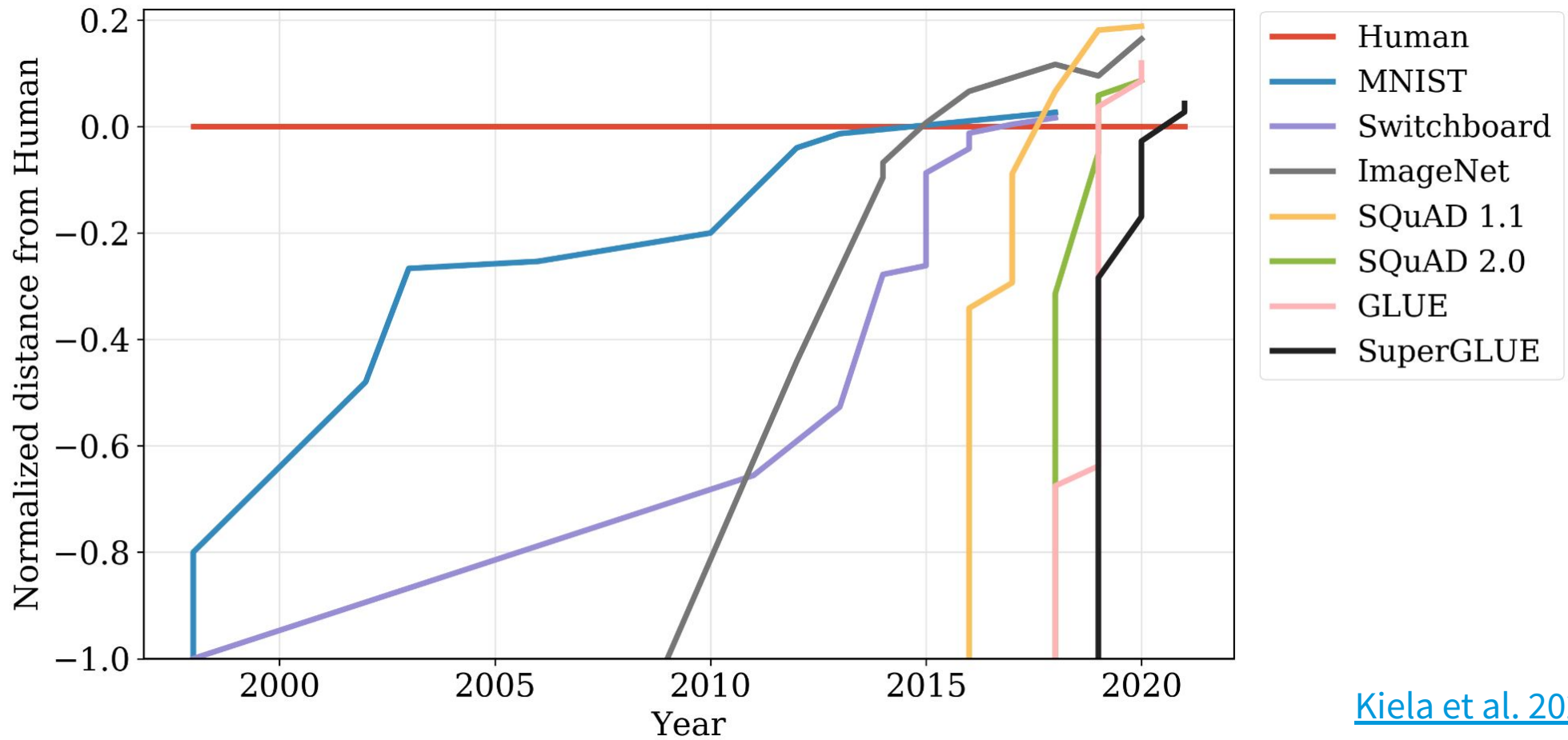
Jacques Cousteau: “Water and air, the two essential fluids on which all life depends, **have become global garbage cans.**”



We ask a lot of our datasets

1. Optimize models
2. Evaluate models
3. Compare models
4. Enable new capabilities in models
5. Measure fieldwide progress
6. Scientific inquiry

What does benchmark saturation really mean?



[Kiela et al. 2021](#)

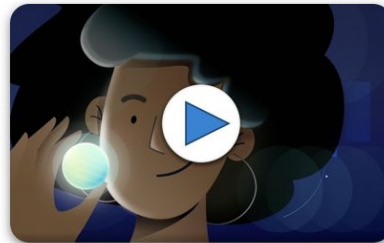
Dynabench



Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?



[Read more](#)

Course theme

More meaningful evaluations

Strathern's Law:

When a measure becomes a target,
it ceases to be a good measure

What we seem to value

**Selected ‘Values
encoded in ML
research’ from
[Birhane et al. 2021](#)**

Performance

Efficiency

Interpretability (for researchers)

Applicability in the real world

Robustness

Scalability

Interpretability (for users)

Benificence

Privacy

Fairness

Justice

Towards multidimensional leaderboards

DAWNBench: An End-to-End Deep Learning Benchmark and Competition

Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi,
Peter Bailis, Kunle Olukotun, Chris Ré, Matei Zaharia
Stanford DAWN Project

<http://dawn.cs.stanford.edu/benchmark>

Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking

Zhiyi Ma^{†*} Kavin Ethayarajh^{‡*} Tristan Thrush^{†*} Somya Jain[†]

Ledell Wu[†] Robin Jia[†] Christopher Potts[‡] Adina Williams[†] Douwe Kiela[†]

[†] Facebook AI; [‡] Stanford University

dynabench@fb.com

Dynascoring

Model	Performance	Throughput	Memory	Fairness	Robustness	Dynascore
DeBERTa	76.25	4.47	6.97	88.33	90.06	45.92
ELECTRA-large	76.07	2.37	25.30	93.13	91.64	45.79
RoBERTa	69.67	6.88	6.17	88.32	86.10	42.54
ALBERT	68.63	6.85	2.54	87.44	80.90	41.74
BERT	57.14	6.70	5.55	91.45	80.81	36.07
BiDAF	53.48	10.71	3.60	80.79	77.03	33.96
Unrestricted T5	28.80	4.51	10.69	92.32	88.41	22.18
Return Context	5.99	89.80	1.10	95.97	91.61	15.47

Question answering

[Ma, Ethayarajh, Thrush, et al. 2021](#)

Dynascoring

Model	Performance	Throughput	Memory	Fairness	Robustness	Dynascore
DeBERTa	76.25	4.47	6.97	88.33	90.06	46.70
ELECTRA-large	76.07	2.37	25.30	93.13	91.64	46.86
RoBERTa	69.67	6.88	6.17	88.32	86.10	43.37
ALBERT	68.63	6.85	2.54	87.44	80.90	42.66
BERT	57.14	6.70	5.55	91.45	80.81	37.17
BiDAF	53.48	10.71	3.60	80.79	77.03	34.62
Unrestricted T5	28.80	4.51	10.69	92.32	88.41	23.19
Return Context	5.99	89.80	1.10	95.97	91.61	14.29

Question answering

[Ma, Ethayarajh, Thrush, et al. 2021](#)

A new era of more meaningful evaluations?

Assessment today

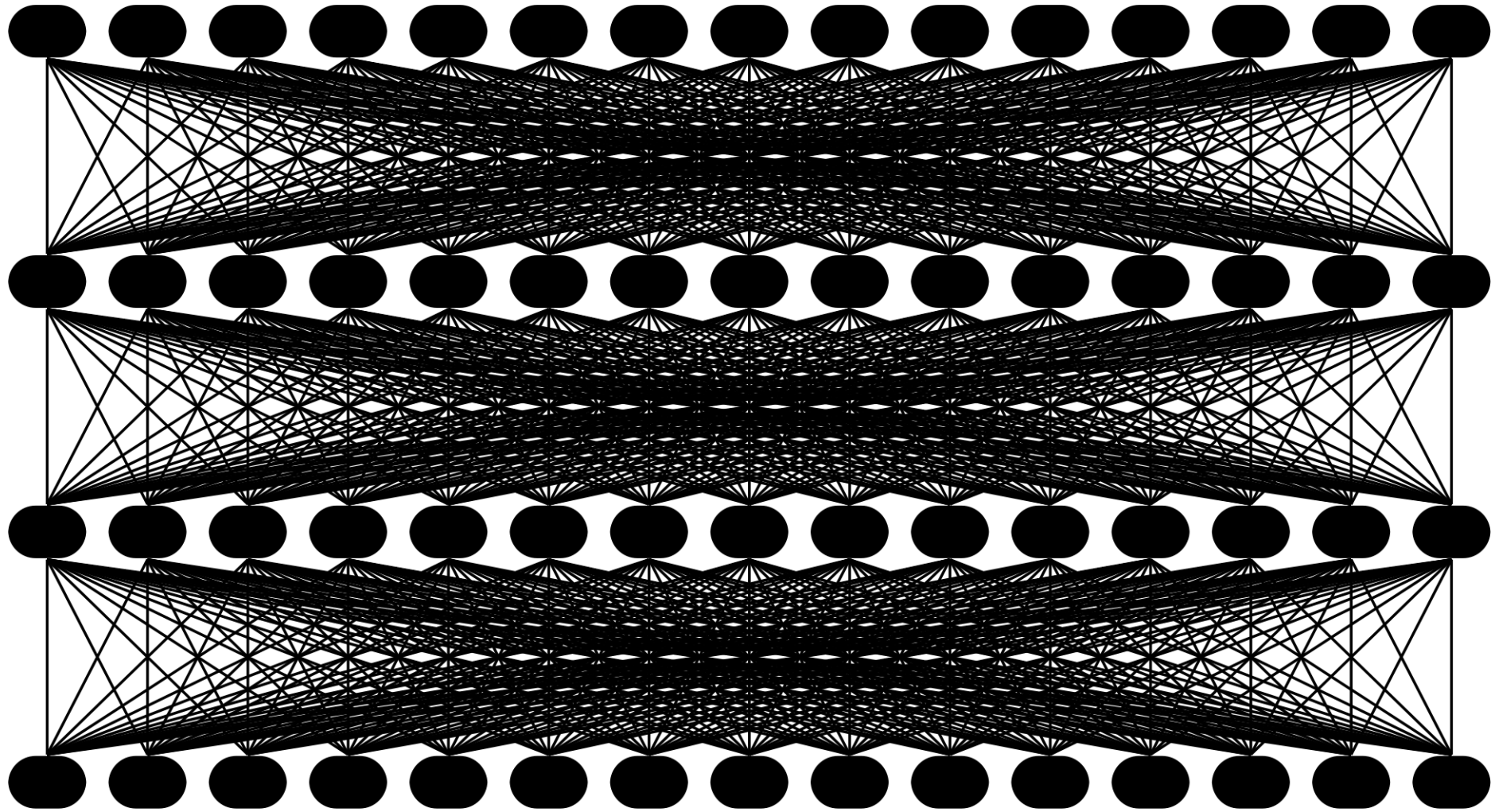
- One-dimensional
- Largely insensitive to context (use-case)
- Terms set by the research community
- Opaque
- Tailored to machine tasks

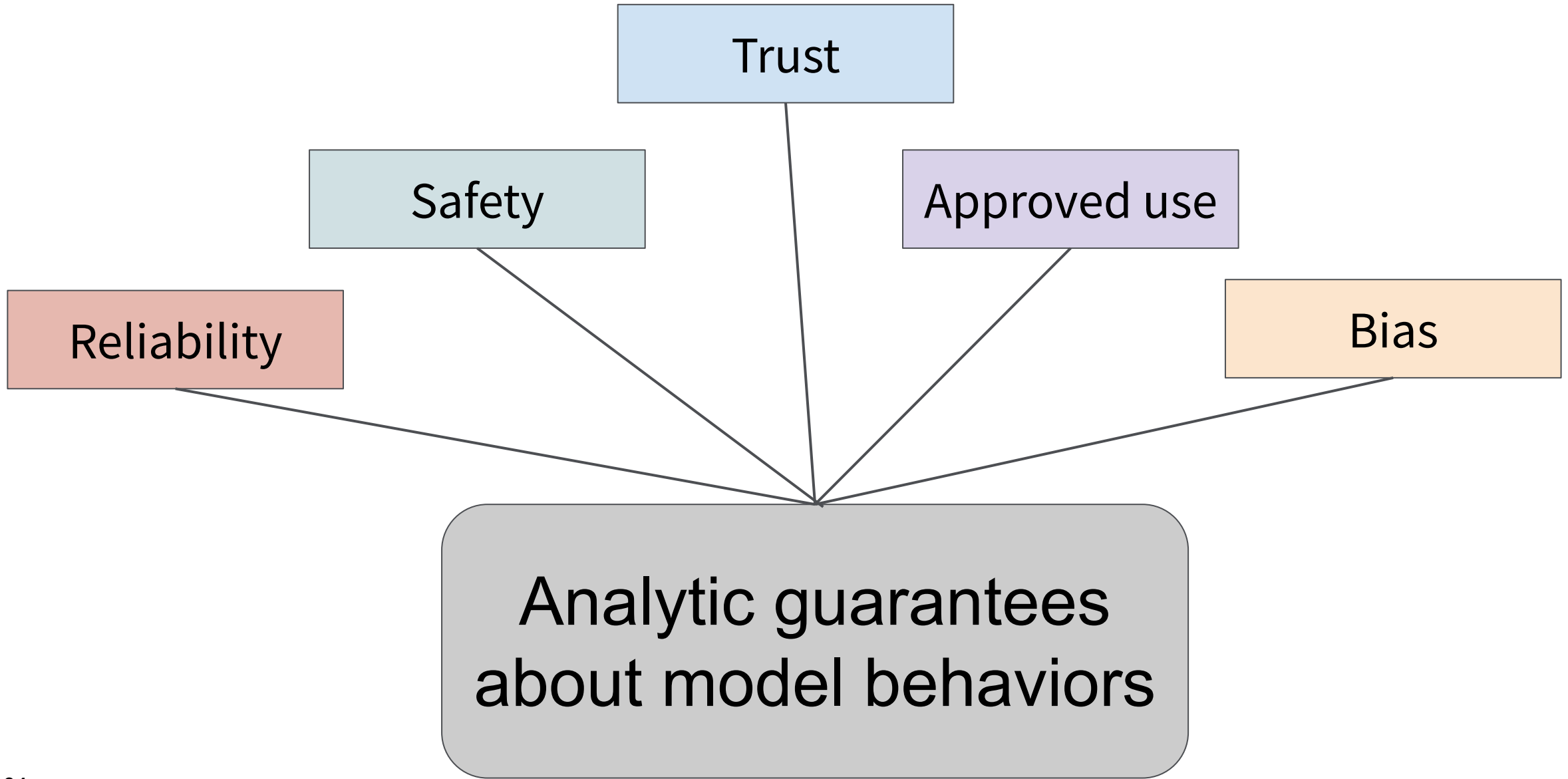
Assessments tomorrow

- High-dimensional and fluid
- Highly sensitive to context (use-case)
- Terms set by the stakeholders
- Judgments ultimately made by users
- Tailored to human tasks

Course theme

Faithful, human-interpretable explanations of models





Standards for explanation

Human interpretable

We can give mechanistic, mathematical explanations of how models work that are perfectly faithful and accurate.

However, these explanations fail to illuminate the concepts we care about.

Faithful

We can give human interpretable explanations that are not true to how our models actually work.

These can seem satisfying, but if we can't guarantee that they are faithful to how the models actually work, we are simply confusing ourselves.

Goal: Concept-level of the causal effects

Explanation methods for NLP models

Train/test evaluations cannot provide guarantees about behavior on new examples.

Probing methods illuminate internal representations but do not support causal inferences.

Attribution methods illuminate the causal dynamics of models but don't characterize their internal representations

Active manipulations of model internal states provide *causal insights* and *rich characterizations* of those states.

Interchange intervention training: train models to conform to the structure of high-level symbolic models.

Looking ahead

1980

```

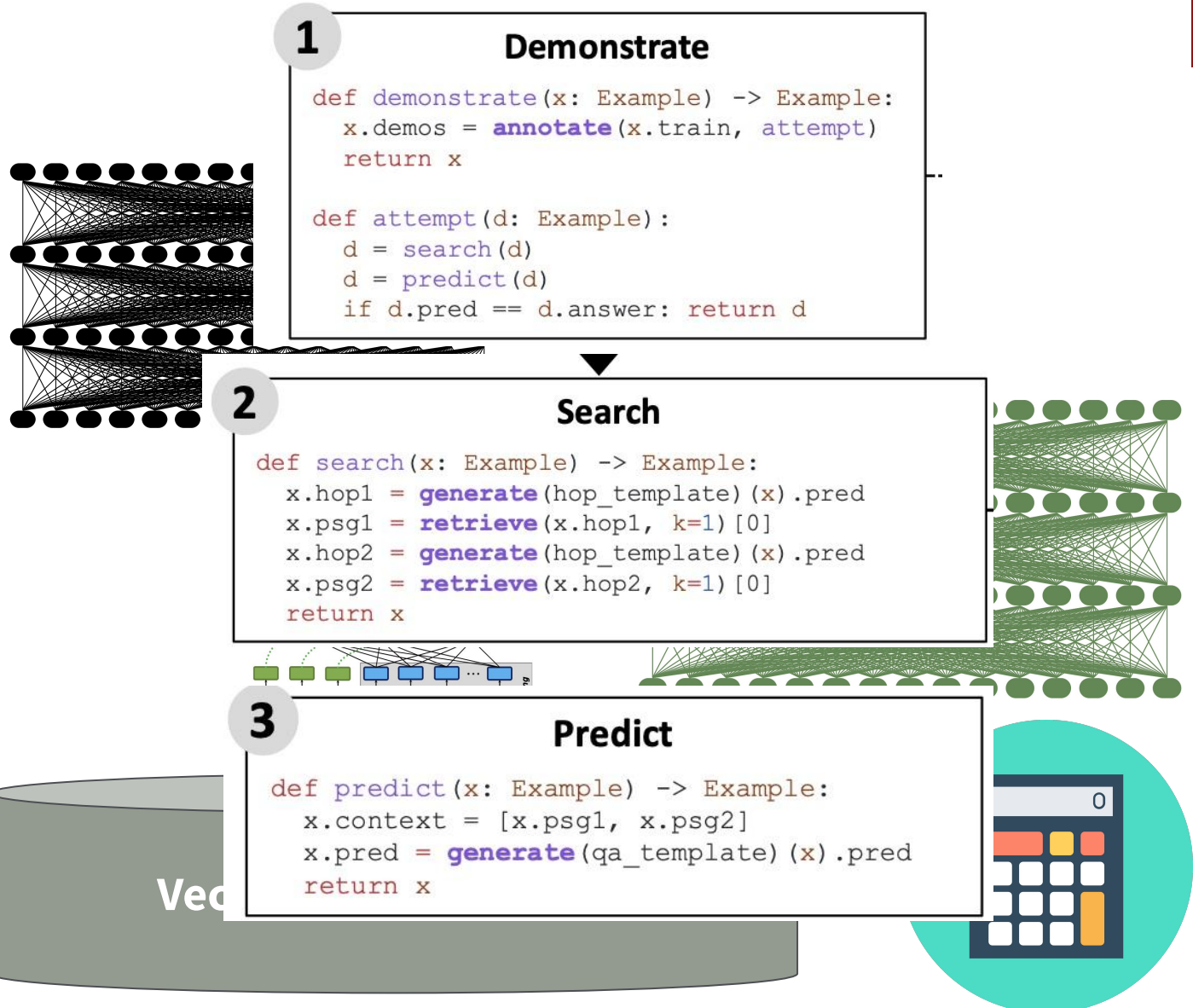
/* Sentences */
sentence(S) --> declarative(S), terminator(.) .
sentence(S) --> wh_question(S), terminator(?) .
sentence(S) --> yn_question(S), terminator(?) .
sentence(S) --> imperative(S), terminator(!) .

/* Noun Phrase */
np(np(Agmt, Pronoun, []), Agmt, NPCase, def, _, Set, Nil) -->
  {is_pp(Set)},
  pers_pron(Pronoun, Agmt, Case),
  {empty(Nil), role(Case, decl, NPCase)}.

/* Prepositional Phrase */
pp(pp(Prep, Arg), Case, Set, Mask) -->
  prep(Prep),
  {prep_case(NPCase)},
  np(Arg, _, NPCase, _, Case, Set, Mask) .

```

2023



Course mechanics

Core course components

Quizzes	15%
Homeworks and bakeoffs	35%
Literature review	10%
Experiment protocol	10%
Final project paper	30%

Fully asynchronous

- All lectures are recorded, and attendance is not required.
- Attending lectures is a great way to participate in shaping the course and build connections with the teaching team.
- Office hours are offered in person and on Zoom; details to come.
- Continuous evaluation: three assignments, four online quizzes, and three components to the project work.

A note on grading original systems

All the homeworks culminate in an “original system” question that becomes your bakeoff entry. Here are the basic guidelines we will adopt for grading this work:

1. We want to emphasize that this needs to be an *original* system. It doesn't suffice to download code from the Web, retrain, and submit, even if this leads to an outstanding bakeoff score. You can build on others' code, but you have to do something new and meaningful with it.
2. Systems that are very creative and well-motivated will be given full credit even if they do not perform well on the bakeoff data. We want to encourage creative exploration!
3. Other systems will receive less than full credit, based on the judgment of the teaching team. The specific criteria will vary based on the nature of the assignment. Point deductions will be justified in feedback.

Project work

1. The second half of the course is devoted to projects.
2. The associated lectures, notebooks, and readings are focused on methods, metrics, and best practices.
3. The assignments are all project-related; details are available at the course website
4. [Exceptional final projects from past years](#) (access restricted)
5. [Lots of guidance on projects](#)

Crucial course links

- [Website](#)
- [Code repository](#)
- [Discussion forum](#)
- [Gradescope](#)
- Teaching team: cs224u-spr2223-staff@lists.stanford.edu

Quizzes

1. Quiz 0 is on course requirements and related details. The sole purpose of the quiz is to create a clear incentive for you to study the website and understand your rights and obligations.
2. Quizzes 1–4 create a course-related incentive for individual students to study the material beyond what is required for the more free-form and collaborative assignments.
3. All quizzes are open notes, open book, open ChatGPT, etc., but no collaboration is permitted.

For next time

1. Get set up using `setup.ipynb` in the course repo.
2. Make sure you're in the discussion forum. If not, follow the link
3. given at the homepage for our course Canvas.
4. Consider doing Quiz 0 as a way of getting to know your rights and obligations for this course.
5. Check out `hw_sentiment.ipynb`. If this material is new to you or you need a refresher, check out the [background materials](#).

Computing resources

1. We expect to get you AWS credits.
2. Consider getting a [Colab Pro](#) account; at \$9.99/month, a three-month subscription is cheaper than even the cheapest textbooks.
3. Sign up for [SageMaker Studio Lab](#) for additional free GPU support.
4. Sign up for [Cohere](#) for (for now) free access to outstanding language models (and [OpenAI](#) still offers \$5 in credits for new accounts).

Core goals (repeated from above)

- Hands-on experience with a wide range of challenging NLU problems.
- A mentor from the teaching team will guide you through the project assignments – there are many examples of these projects becoming important publications.
- **Central goal:** to make you the best – most insightful and responsible – NLU researcher and practitioner wherever you go next.