Recent Breakthroughs and Uphill Battles in Modern Natural Language Processing

Daniel Khashabi

About me

- Assistant Professor,
  Johns Hopkins University, Baltimore
- Working on Artificial Intelligence & Natural Language Processing
- PhD from University of Pennsylvania, 2019
- BSc from Tehran Polytechnic, 2012

Popular Media: Language is solved!!

State-of-the-art AI solutions: (1) Google BERT, an AI model that understands language better than humans

AIN Dev Team Follow
Jan 31 · 8 min read

VICE

Algorithms Have Nearly Mastered Human Language. Why Can't They Stop Being Sexist?

It turns out that data-fueled algorithms are no better than humans—and ... Even AI researchers who work with machine learning models—like neural nets, which ...

Sep 18, 2019

Digital Journal

Researchers shut down AI that invented its own language

An artificial intelligence system being developed at Facebook has created its own language. It developed a system of code words to make ...

Jul 21, 2017

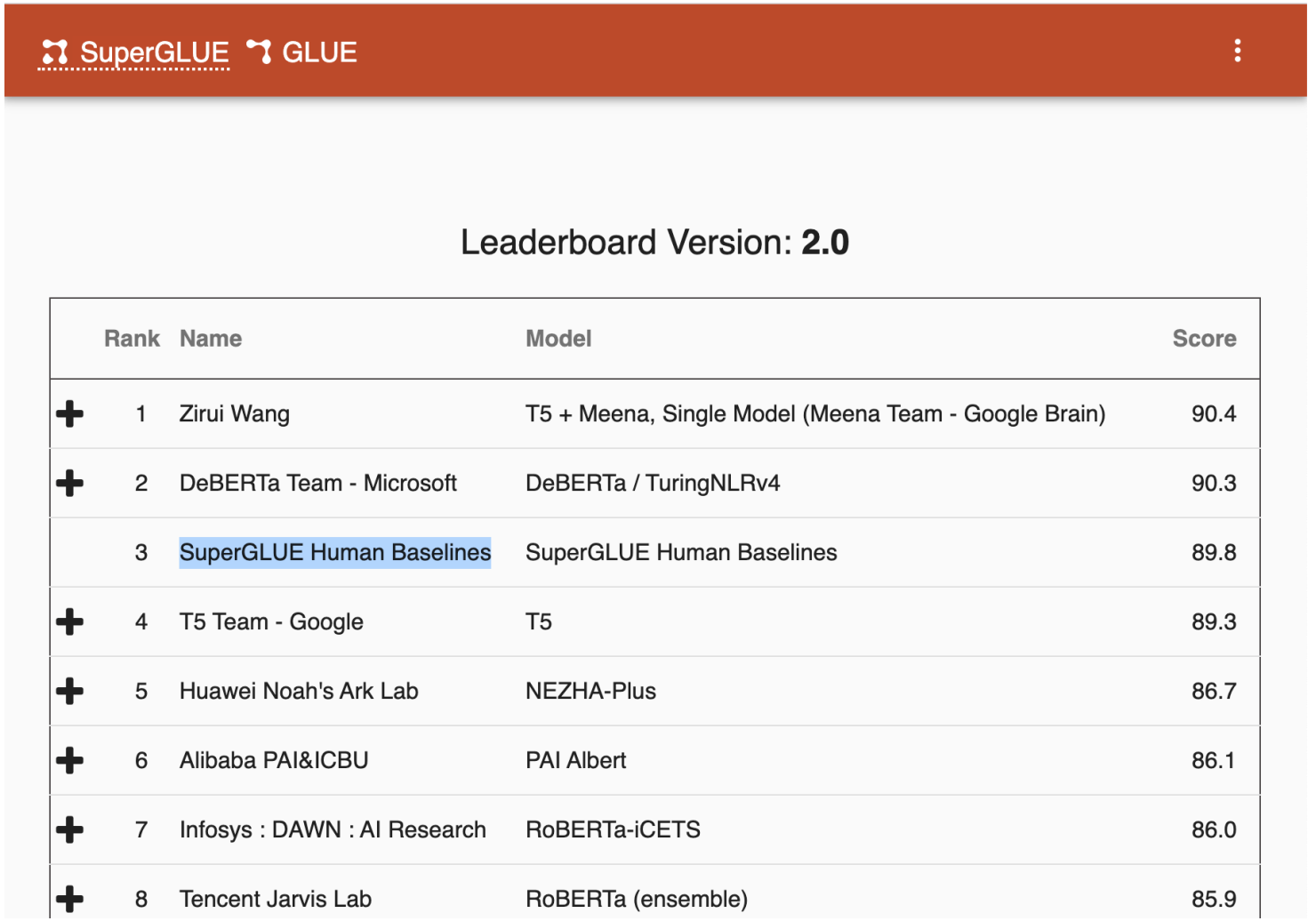
Alizila

Alibaba AI Beats Humans in Reading-Comprehension...

Alibaba Group's machine-learning technology is better at reading comprehension than humans, according to a well-known test built for the industry by Microsoft.

Jul 9, 2019

Leaderboards: NLP is on-par w/ humans



	Rank	Name	Model	Score
+	1	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)	90.4
+	2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	90.3
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines	89.8
+	4	T5 Team - Google	T5	89.3
+	5	Huawei Noah's Ark Lab	NEZHA-Plus	86.7
+	6	Alibaba PAI&ICBU	PAI Albert	86.1
+	7	Infosys : DAWN : AI Research	RoBERTa-iCETS	86.0
+	8	Tencent Jarvis Lab	RoBERTa (ensemble)	85.9

https://super.gluebenchmark.com/leaderboard

Self-Supervised Models

Self-Supervised Models

Self-Supervision

[Slide from Colin Raffel]

Self-Supervision

[Slide from Colin Raffel]

Self-Supervision

[Slide from Colin Raffel]

Self-Supervision

Dataset of natural images

[Slide from Colin Raffel]

Self-Supervision

Dataset of natural images

Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide from Colin Raffel]

Self-Supervision

Dataset of natural images

Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide from Colin Raffel]

Self-Supervision

Dataset of natural images

Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide from Colin Raffel]

Self-Supervision

Dataset of natural images

Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide from Colin Raffel]

Self-Supervision

== treaty of paris (1763)

the treaty of paris, also known as the treaty of 1763, was signed on 10 february 1763 by the kingdoms of great britain, france and spain, with portugal in agreement, after great britain's victory over france and spain during the seven years' war.

the signing of the treaty formally ended the seven years' war, known as the french and indian war in the north american theatre, ...

[Slide from Colin Raffel]

Self-Supervision

== wheelbarrow

A wheelbarrow is a small hand-propelled vehicle, usually with just one wheel, designed to be pushed and guided by a single person using two handles at the rear, or by a sail to push the ancient wheelbarrow by wind. The term "wheelbarrow" is made of two words: "wheel" and "barrow." "Barrow" is a derivation of the Old English "barew" which was a device used for carrying loads. The wheelbarrow is designed to distribute the weight of its load between the wheel …

[Slide from Colin Raffel]

Self-Supervision

WIKIPEDIA
The Free Encyclopedia

== lemon

The lemon (Citrus limon) is a species of small evergreen trees in the flowering plant family Rutaceae, native to Asia, primarily Northeast India (Assam), Northern Myanmar or China.[2] The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.[2] The pulp and rind are also used in cooking and baking. The juice of the lemon is about 5% to 6% citric acid, with a pH ....

== wh

A wh
vehic
desig
single
rear,
whee
"whe
"whe
deriv
which
loads
distri
the wheel ...

== tr

the t
treat
1763
franc
agre
over
years

the s
the s
and i
theatre, ...

Dataset of Wikipedia articles

[Slide from Colin Raffel]

Self-Supervision

WIKIPEDIA
The Free Encyclopedia

== tr

the t
treat
1763
franc
agre
over
years

the s
the s
and i
theatre, ...

== wh

A wh
vehic
desig
single
rear,
whee
"whe
"whe
deriv
which
loads
distri
the wheel ...

== lemon

The lemon (Citrus limon) is a species of small evergreen trees in the flowering plant family Rutaceae, native to Asia, primarily Northeast India (Assam), Northern Myanmar or China.[2] The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.[2] The pulp and rind are also used in cooking and baking. The juice of the lemon is about 5% to 6% citric acid, with a pH ....

Dataset of Wikipedia articles

== wings over kansas

wings over kansas is the second studio album by jason ammons, john bolster and mo rosato. the album debuted at number one on the billboard 200, selling 35,000 copies in it first week at the time. it was the second highest selling album to debut at the billboard top 50 and the third highest selling album to debut at the top heatseekers, with 26,000 copies sold. this is the supremes album earning the nickname nitty gritty but their other two singles by the band in ...

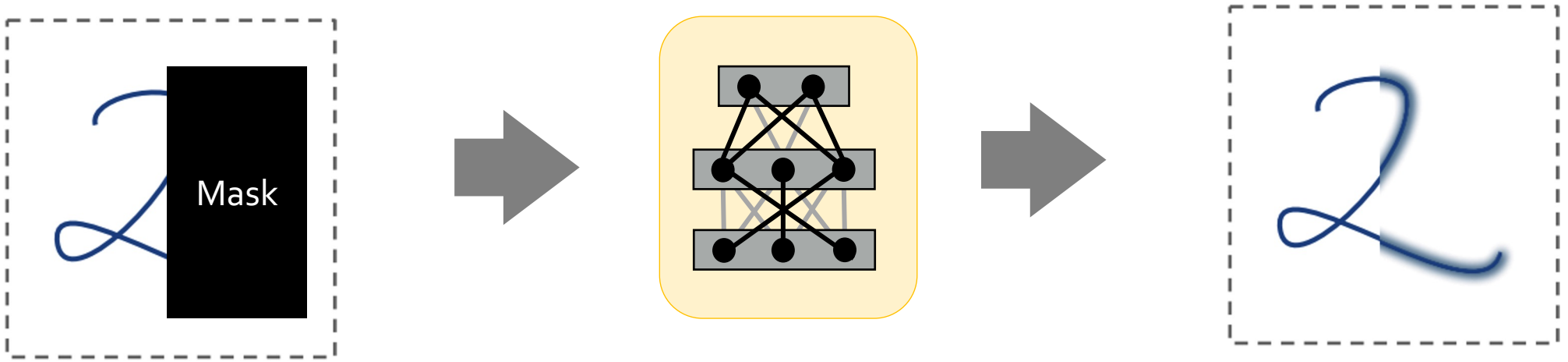
[Slide from Colin Raffel]

Self-Supervised Models

Self-Supervised Models

Self-Supervised **Models**

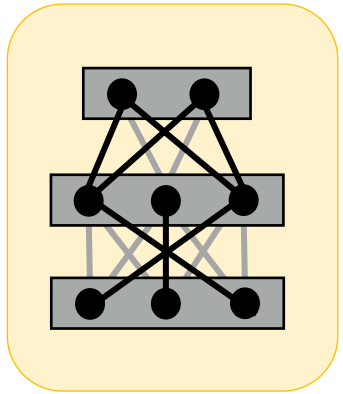
Self-Supervised **Models**



[Bengio et al. 2004, Hinton et al. 2006, Peters et al. 2018, …]

Self-Supervised **Models**

"Wings over Kansas is [MASK]"



"Wings over Kansas is an aviation website founded in 1998 by Carl Chance owned by Chance Communications, Inc."

[Bengio et al. 2004, Hinton et al. 2006, Peters et al. 2018, …]

Self-Supervised **Models**: A History

- Shannon (1950) the entropy of English
  — i.e., its predictive difficulty.

Prediction and Entropy of Printed English

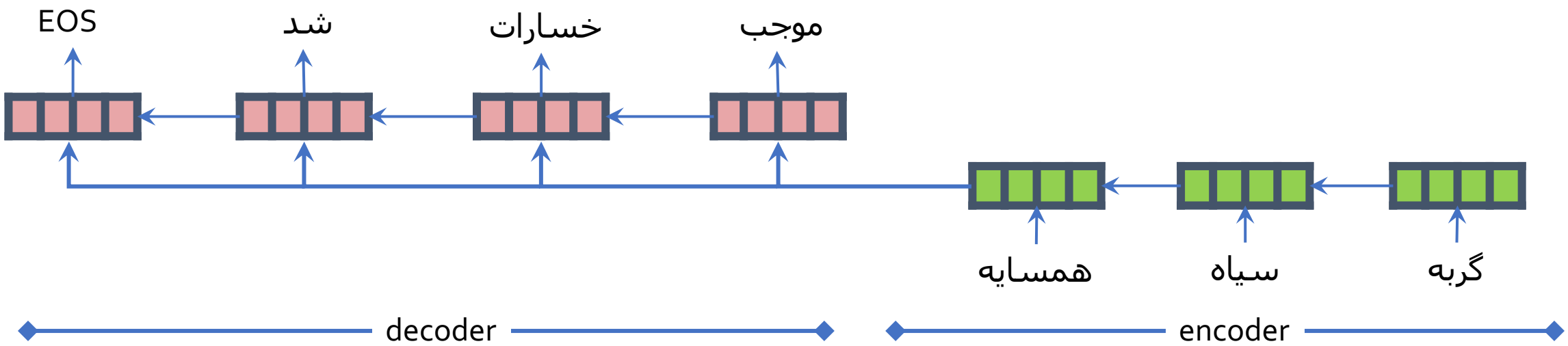
By C. E. SHANNON

(*ManuscriptReceived Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

✓ **Self-supervised models**
are predictive models of the world.

Encoder Decoder Models



- Translation (e.g., Google Translate)
- Grammatical Error Correction (Grammarly, etc.)
- Email response suggestion (Gmail, Outlook, etc.)

Encoder-Decoders: Example

فصل تابستان را توصیف کنید. ابستان و گرما این فصل باعث می‌شود افراد برای فرار از شلوغی و مشغله روزانه به پارک و فضاهای سرسبز و سرسبز و آرام پناه ببرند. این پارک‌ها بسیار هیجان انگیز و سرسبز هستند و با داشتن انواع وسایل بازی برای کودکان لحظات خوشی را ایجاد می‌کنند. اگر تابستان در یکی از این پارک‌ها اقامت داشته باشید می‌توانید با خانواده و دوستان به تفریح بپردازید. اگر می‌خواهید در این پارک‌ها به گشت و گذار بپردازید باید بدانید که این پارک‌ها نیاز به یک مکان مناسب برای اقامت دارند تا بتوانید ساعات خوشی را در آن‌ها سپری کنید. از دیگر مکان‌های مناسب برای سفر تابستان پارک‌های بازی کودکان هستند. در این پارک‌ها به علت اینکه کودکان بیشتر در آن‌ها سرگرم می‌شوند از وسایل بازی متنوع‌تری برخوردار هستند. از جمله وسایل بازی موجود در این پارک می‌توان به کلبه‌های چوبی، تونل‌های بازی و غیره اشاره کرد. پارک‌های آبی و مراکز خرید در این فصل بسیار پرطرفدار هستند. اگر تابستان را برای سفر انتخاب کنید ممکن است به فکر خرید و گشت و گذار در این مراکز تفریحی باشید. این مراکز تفریحی در فصل تابستان بسیار شلوغ هستند و مسافران بیشتری به آنجا می‌روند. در ادامه این مطلب می‌توانید بهترین

https://huggingface.co/flax-community/gpt2-medium-persian

Encoder-Decoders: Example

Here is a paragraph describing my summer experience.

In the past three summers, I've worked with a variety of clients, all of whom had the same basic project in mind: a beautiful residential garden. I've designed and built gardens for the most simple of clients, to the most sophisticated, and for the most well-known and well-established gardeners.

This summer, I was lucky enough to work with the owner of the garden I was designing and building, a very well-known, well-known garden. I would have to say that this was one of the most incredible projects I've ever been a part of.

https://6b.eleuther.ai/

✓ **Language models** are convenient frameworks for **text-in text-out** tasks.

Self-Supervised **Models**

For What End Though?

Many **tasks we care about*** are tightly connected to **predictive models** of the world!

*** Thus far**

Self-Supervised Models and End Tasks

- **Goal:** Answering questions

Question: "Where is the birthplace of the American national anthem?"

✍️

"The birthplace of the American national anthem [MASK]"

"The birthplace of the American national anthem, "The Star-Spangled Banner," lies in Baltimore, Maryland."

Produced with GPT-J: https://6b.eleuther.ai/

Self-Supervised Models and End Tasks

- **Goal:** Sentiment classification

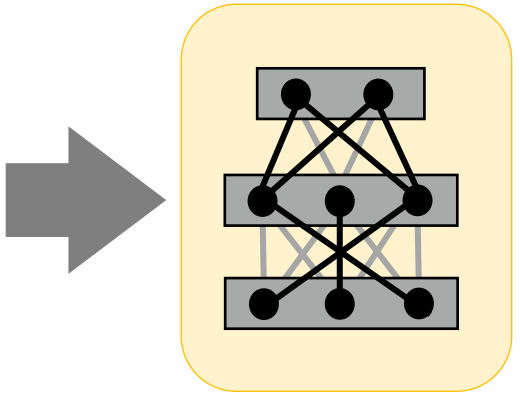
Review: "While this restaurant is popular on Google, I absolutely disliked it."

✍️

"We want to decide whether the sentiment of the review is "positive" or "negative".

Review: "While this restaurant is popular on Google, I absolutely disliked it".

The sentiment of this review is [MASK]"



"...is negative."

Produced with GPT-J: https://6b.eleuther.ai/

✓ Self-supervised models

- *The resulting representations are closely tied to the end tasks*
  - ➢ *Traditionally addressed by **supervised** learning*
- *Learning from unlabeled data*
  - ➢ *A special case of **unsupervised** learning*

In-context Learning w/ **Self-Supervised Models**

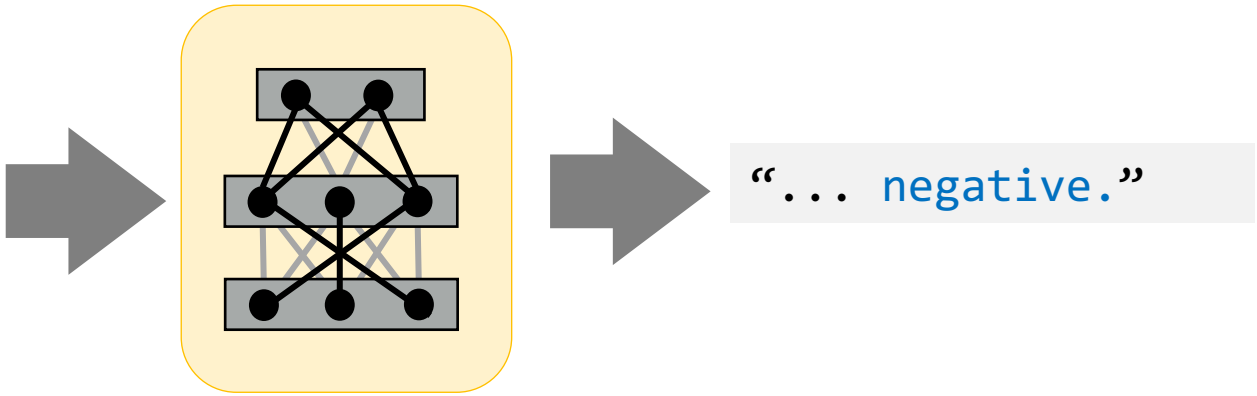
- **Goal:** Sentiment classification

```
Tweet: “I hate when my phone battery dies”
Sentiment: Negative.

Tweet: “My day has been fine.”
Sentiment: Positive.

Tweet: “This phone is useless!”
Sentiment: Negative.

Review: "I absolutely disliked it".
Sentiment:
```



Produced with GPT-J: https://6b.eleuther.ai/

In-context Learning w/ **Self-Supervised Models**

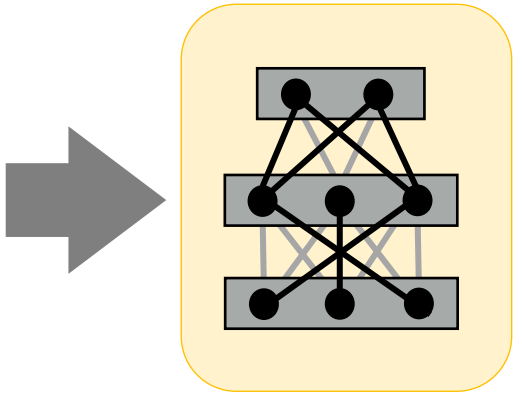
- **Goal:** Sentiment classification

جمله: این رستوران خیلی بده!
احساس: منفی

جمله: غذاش خیلی خوب بود!
احساس: مثبت

جمله: فضاش بد نبود اما در کل لذت نبردم.
احساس: منفی

جمله: بهترین رستورانی بود که امتحان کرده بودم.
احساس:



“مثبت”

Produced with GPT-J: https://huggingface.co/bolbolzaban/gpt2-persian

✓ **In-context prompting** of language models is a convenient way to make them solve tasks, **without additional training**.

New Age of NLP

Pre-1993: Rule based NLP

1993 - 2020: Statistical NLP

2020: Language Model based NLP

The manner in which we have started to do NLP is qualitatively different from before

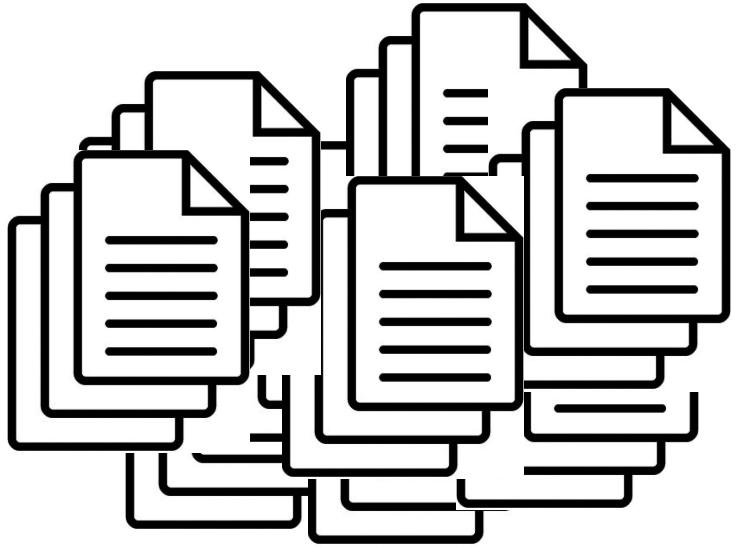
Scaling Self-Supervised Models

- Larger pre-training datasets

cheap freely-available text

1GB

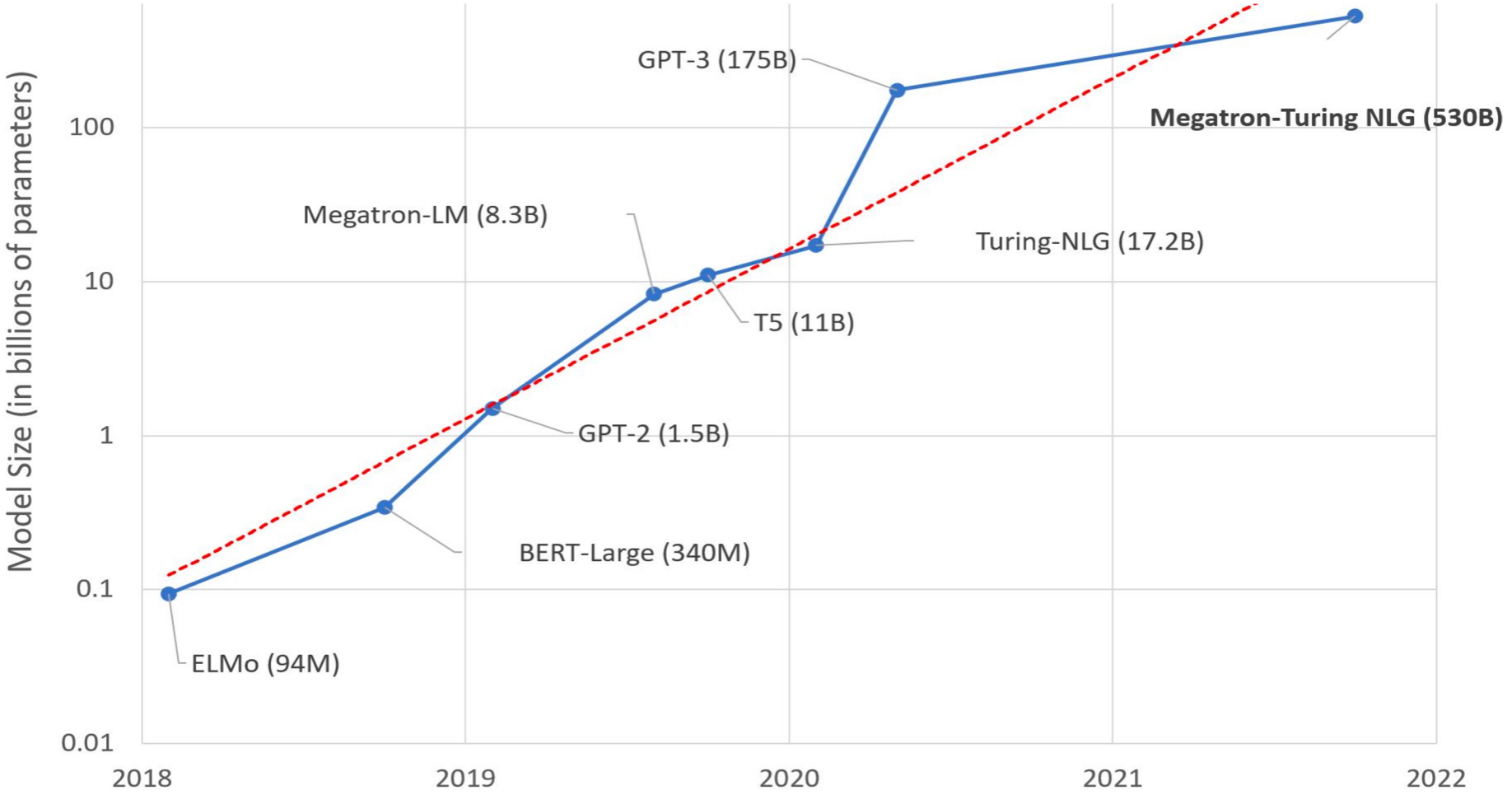
10GB



10TB

Scaling Self-Supervised Models

- Larger pre-training datasets
- Larger models



Scaling Self-Supervised Models

- Larger pre-training datasets
- Larger models

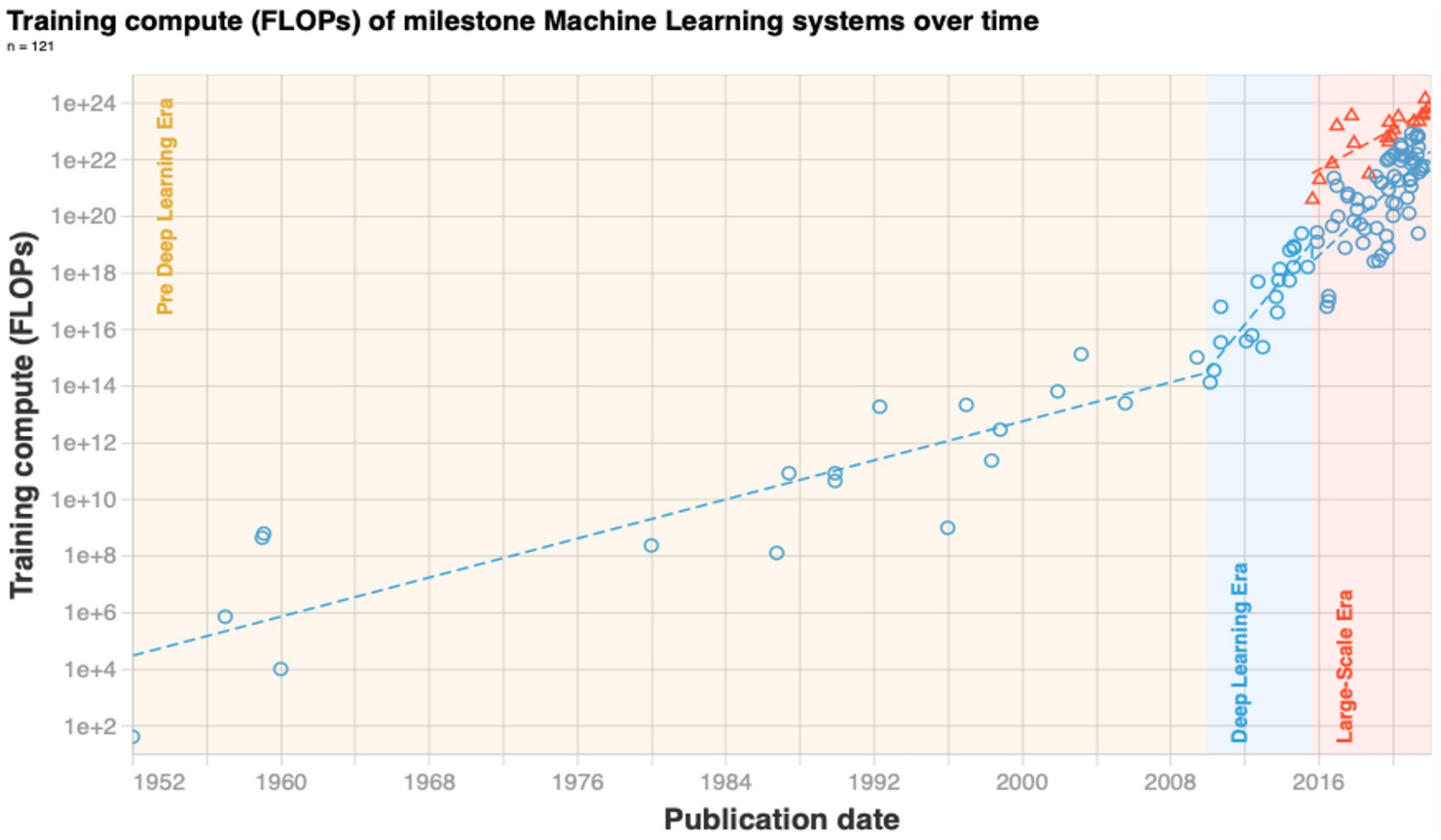
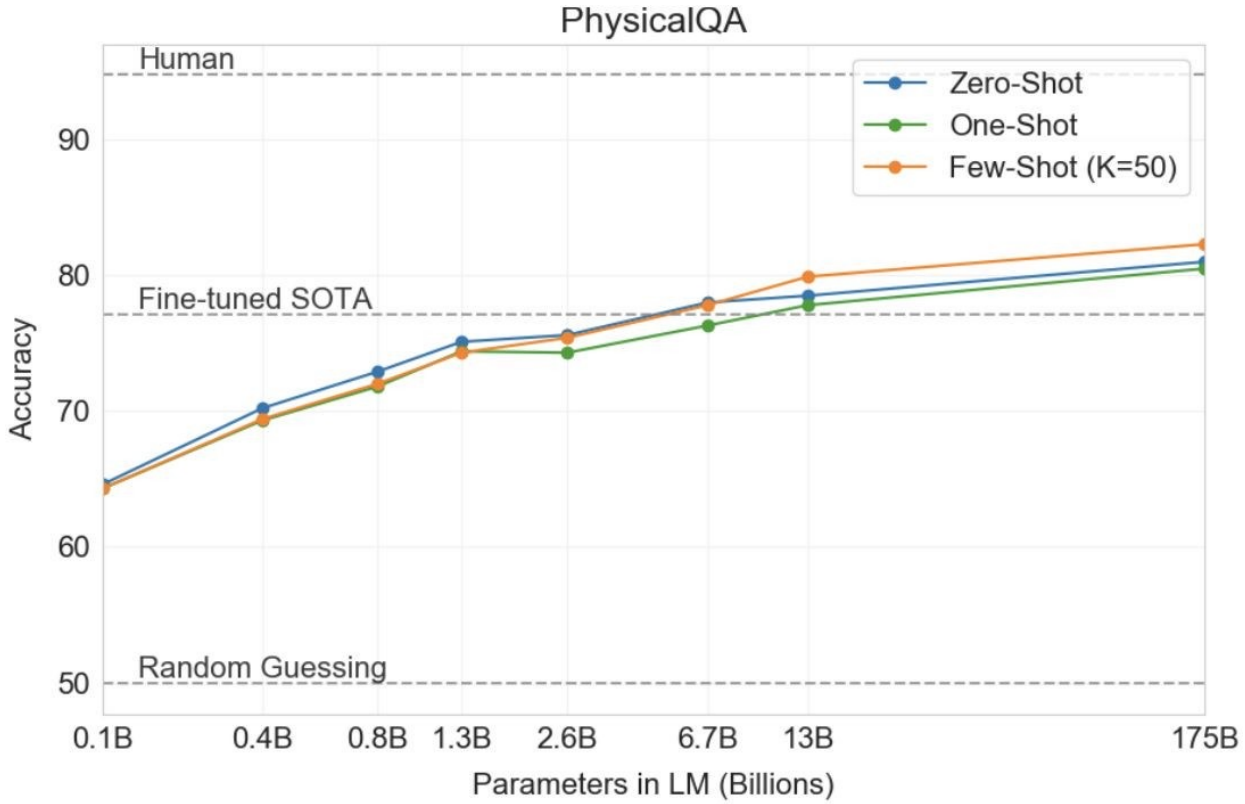
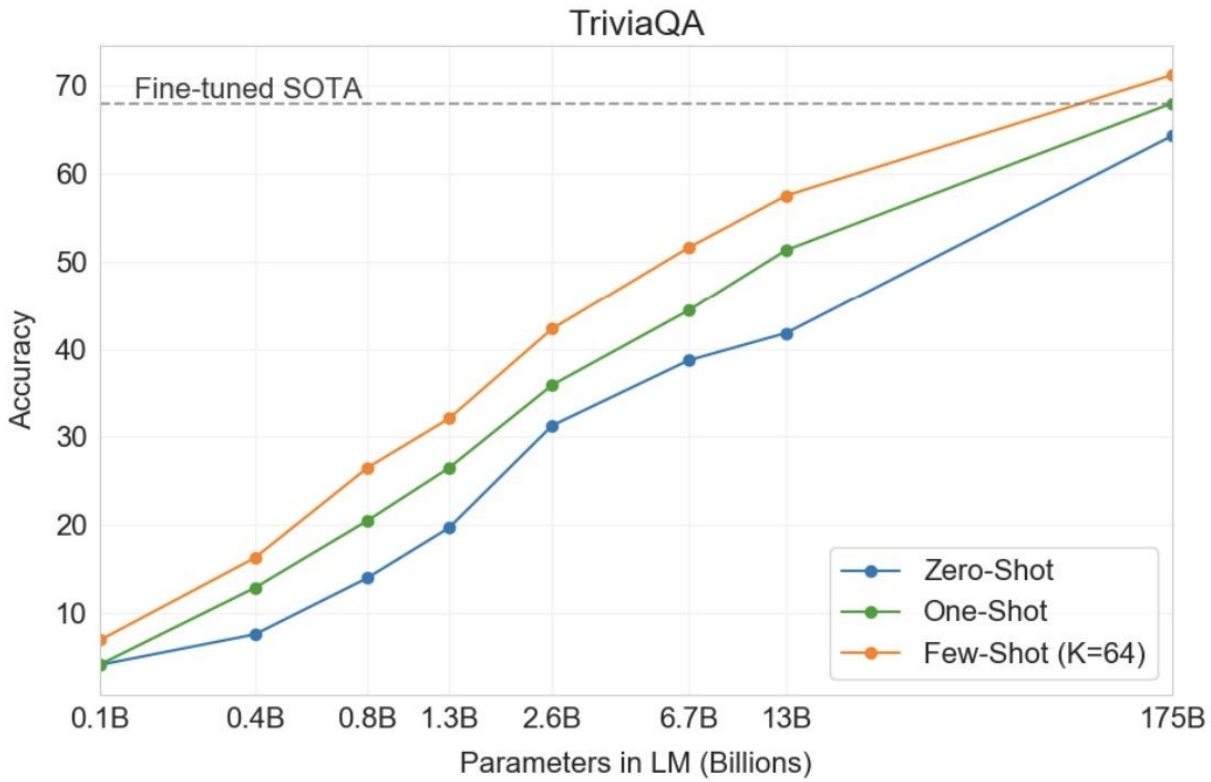


Figure 1: Trends in $n = 121$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

Sevilla, Jaime, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. "Compute trends across three eras of machine learning." *arXiv preprint arXiv:2202.05924* (2022).

Scaling Self-Supervised Models



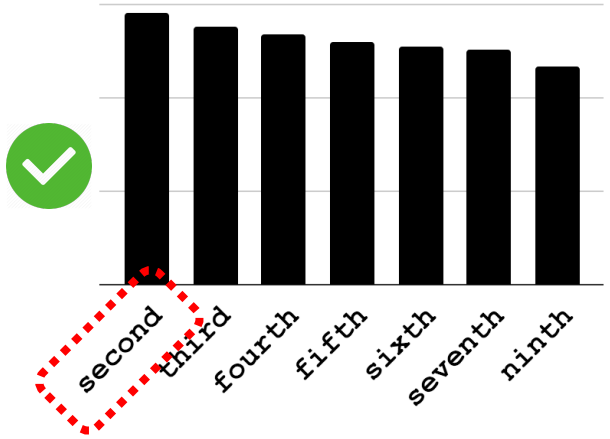


Brown et al. 2020. “Language Models are Few-Shot Learners”

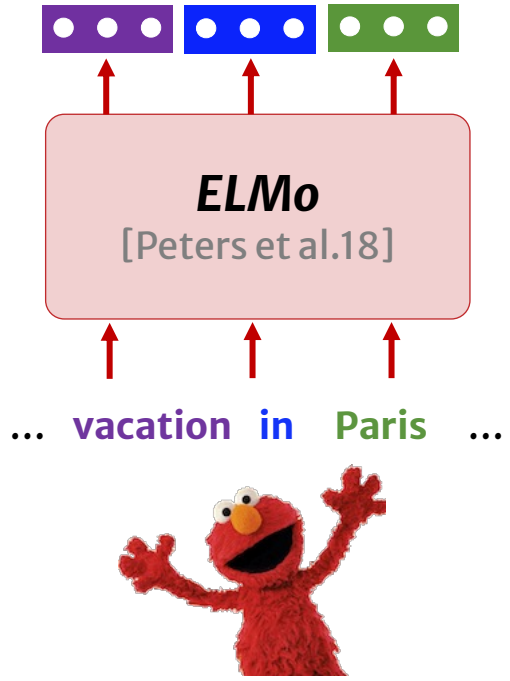
✓ **Scaling language models** consistently lead to stronger models.

Language Models: Means to Access Knowledge

- They let you "query" for knowledge:



Pittsburgh is the _____ -largest populated city in Pennsylvania.

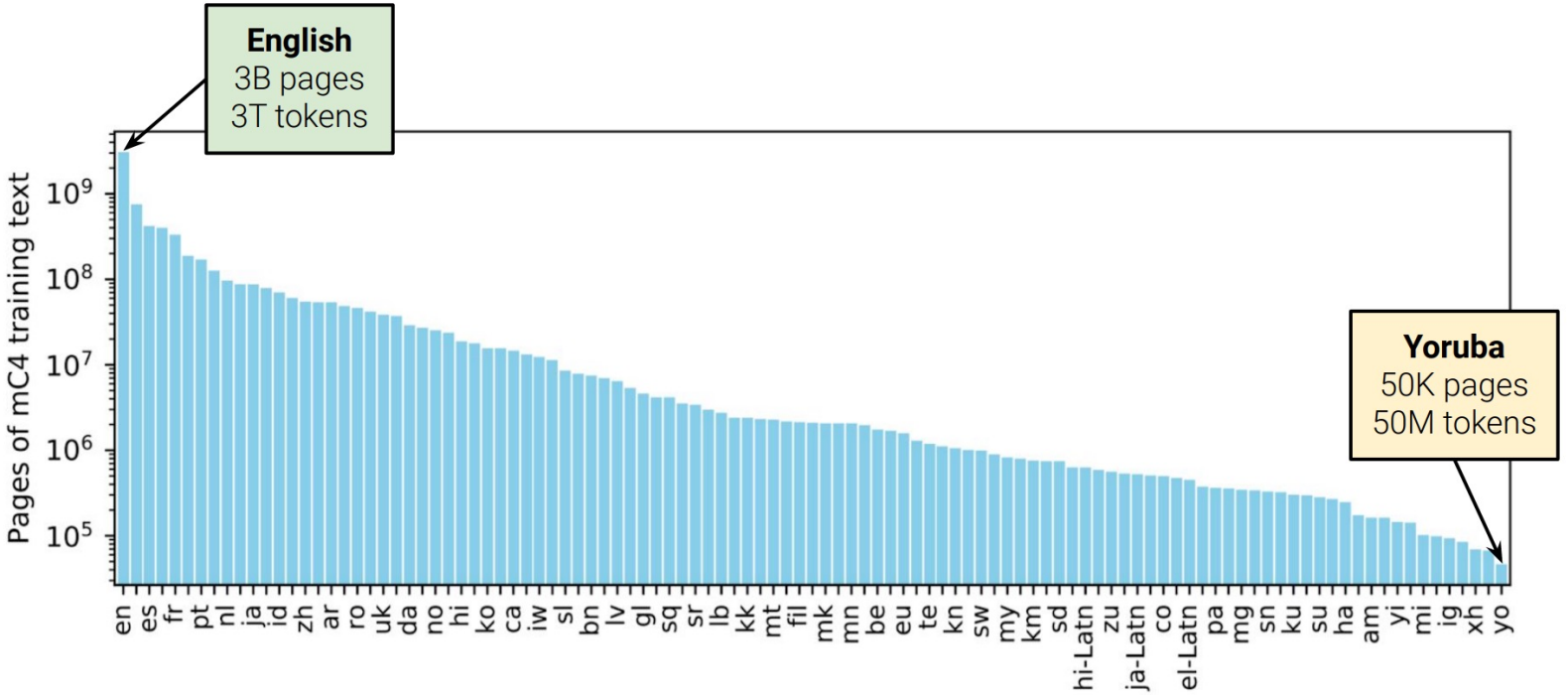


Other works:
[Devlin et al.18; Radford et al.18]

✓ **Language models** serve as efficient mechanisms for knowledge retrieval.

Multilinguality

- Pre-train on a many languages



mT5: https://github.com/google-research/multilingual-t5
multilingual bert: https://huggingface.co/bert-base-multilingual-cased

Benchmarking LMs in Persian

- A Persian language understanding benchmark.

https://github.com/persiannlp/parsinlu

TACL'21

PARSINLU: A Suite of Language Understanding Challenges for Persian

Daniel Khashabi[1] **Arman Cohan**[1] **Siamak Shakeri**[2] **Pedram Hosseini**[3] **Pouya Pezeshkpour**[4]
Malihe Alikhani[5] **Moin Aminnaseri**[6] **Marzieh Bitaab**[7] **Faeze Brahman**[8]
Sarik Ghazarian[9] **Mozhdeh Gheini**[9] **Arman Kabiri**[10] **Rabeeh Karimi Mahabadi**[11]
Omid Memarrast[12] **Ahmadreza Mosallanezhad**[7] **Erfan Noury**[13] **Shahab Raji**[14]
Mohammad Sadegh Rasooli[15] **Sepideh Sadeghi**[2] **Erfan Sadeqi Azer**[2] **Niloofar Safi Samghabadi**[16]
Mahsa Shafaei[17] **Saber Sheybani**[18] **Ali Tazarv**[4] **Yadollah Yaghoobzadeh**[19]
[1]Allen Institute for AI, [2]Google, [3]George Washington U., [4]UC Irvine, [5]U. of Pittsburgh, [6]TaskRabbit, [7]Arizona State U., [8]UC Santa Cruz
[9]U. of Southern California, [10]IMRSV Data Labs, [11]EPFL, [12]U. of Illinois - Chicago, [13]U. of Maryland Baltimore County
[14]Rutgers U., [15]U. of Pennsylvania, [16]Expedia Inc., [17]U. of Houston, [18]Indiana U. - Bloomington, [19]Microsoft

ParsiNLU Tasks

1. Reading Comprehension

سوال: نهاوند جزو کدام استان است؟

Question: Nahavand is part of which province?

پاراگراف: نَهاوند شهری در غرب ایران است. این شهر در جنوب غربی استان همدان قرار گرفته است. نهاوند دارای جمعیت ...

Paragraph: Nahavand (Navan) is a city in western Iran. This city is located in the southern part of Hamedan province and it is the capital of Nahavand. Nahavand has a population of ...

پاسخ: همدان، استان همدان

Answer: Hamedan; Hamedan province

ParsiNLU Tasks

1. Reading Comprehension
2. Multiple-Choice Question Answering

بزرگترین قاره‌ی جهان کدام است؟

(۱ ✓آسیا ۲) اروپا ۳) آمریکا ۴) آفریقا

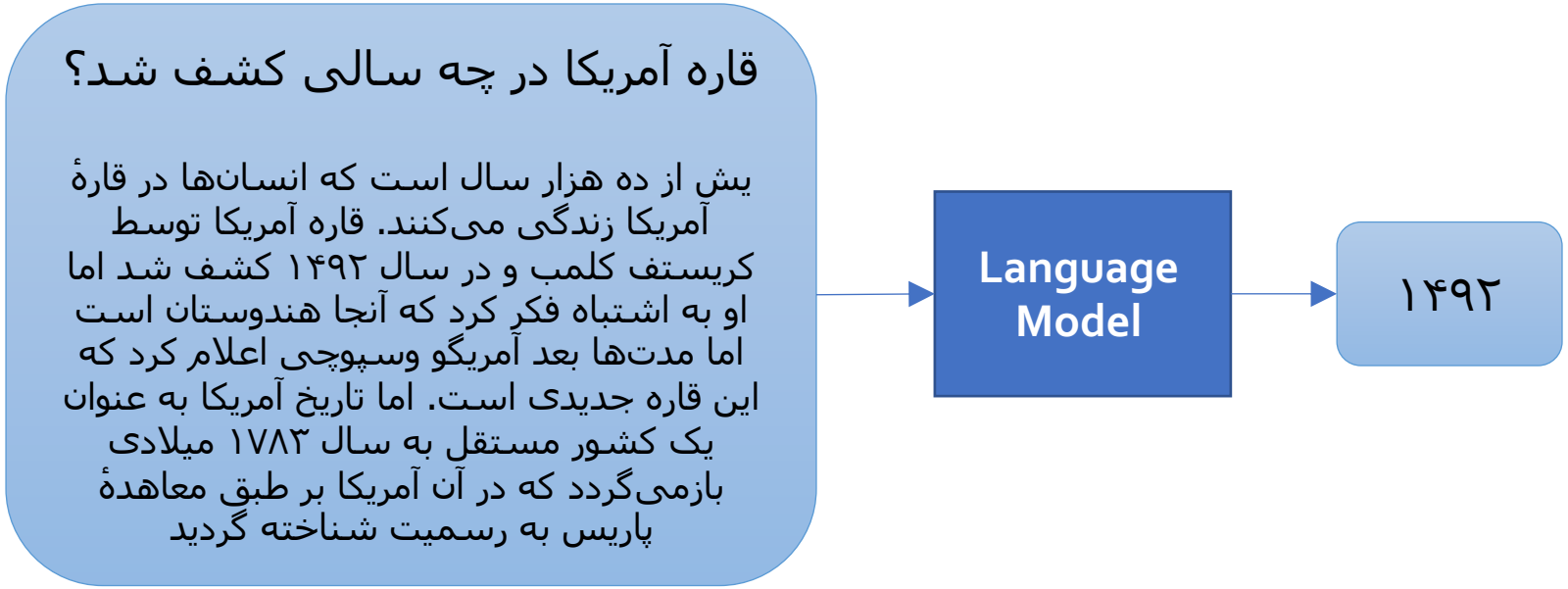
What is the largest continent in the world?

✓ 1) Asia 2) Europe 3) Americas 4) Africa

ParsiNLU Tasks

1. Reading Comprehension
2. Multiple-Choice Question Answering
3. Sentiment Analyses
4. Textual Entailment
5. Paraphrasing
6. Machine Translation

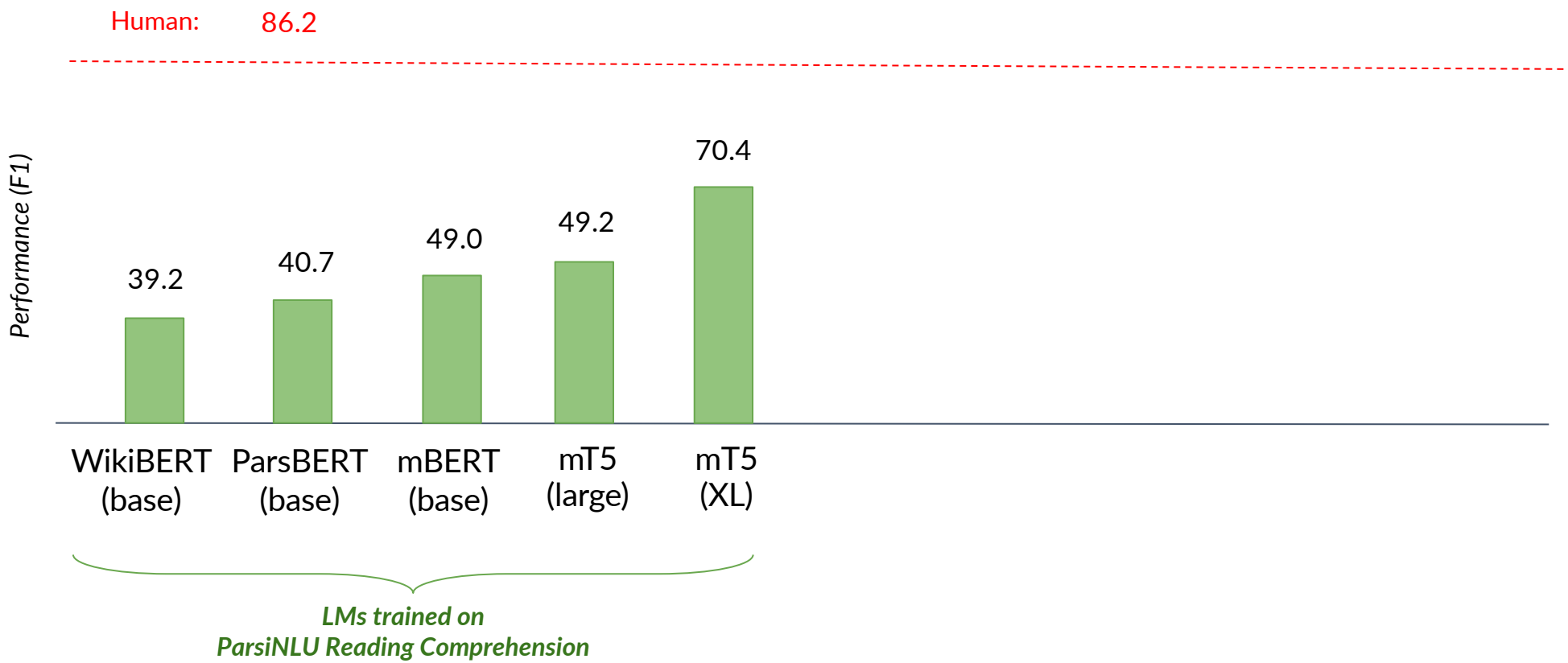
Benchmarking LMs in Persian



https://github.com/persiannlp/parsinlu
https://huggingface.co/persiannlp

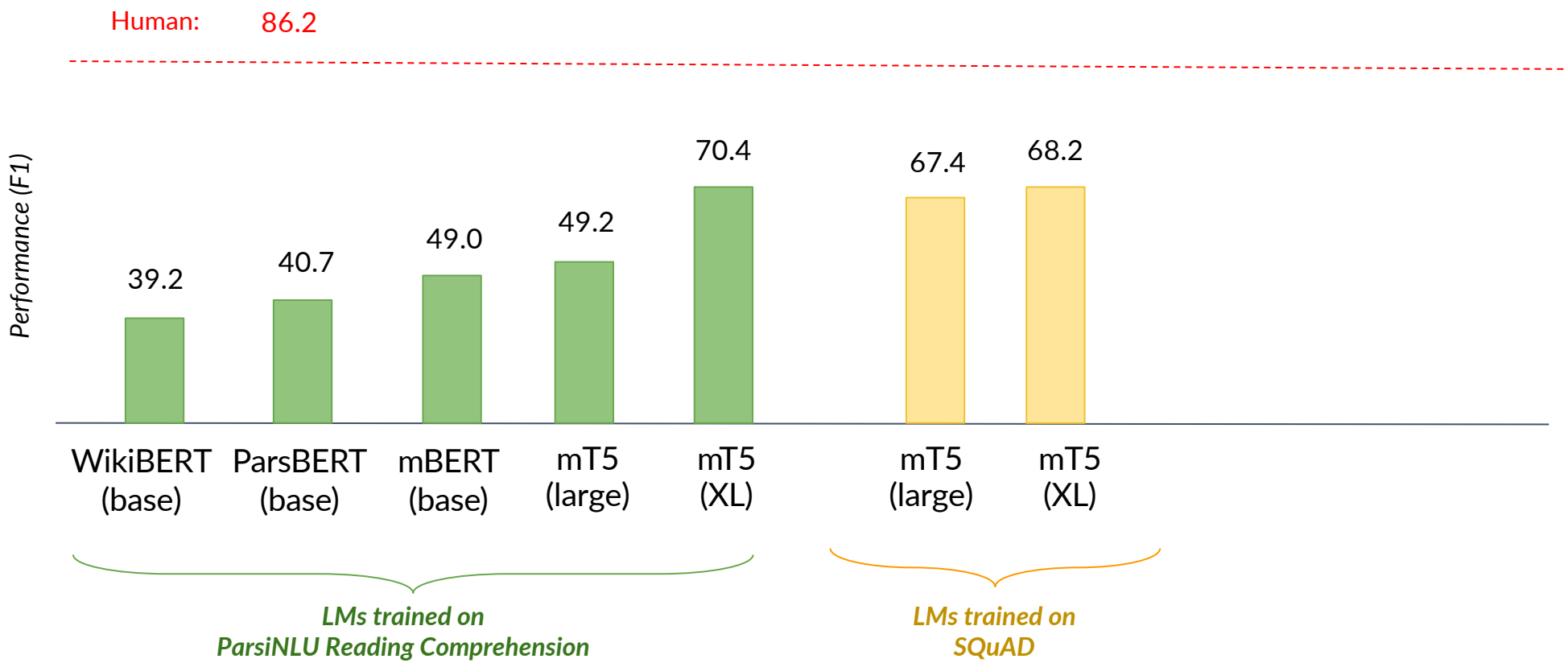
Experimental Findings (1)

- **Finding 1:** The proposed dataset(s) is decent quality.



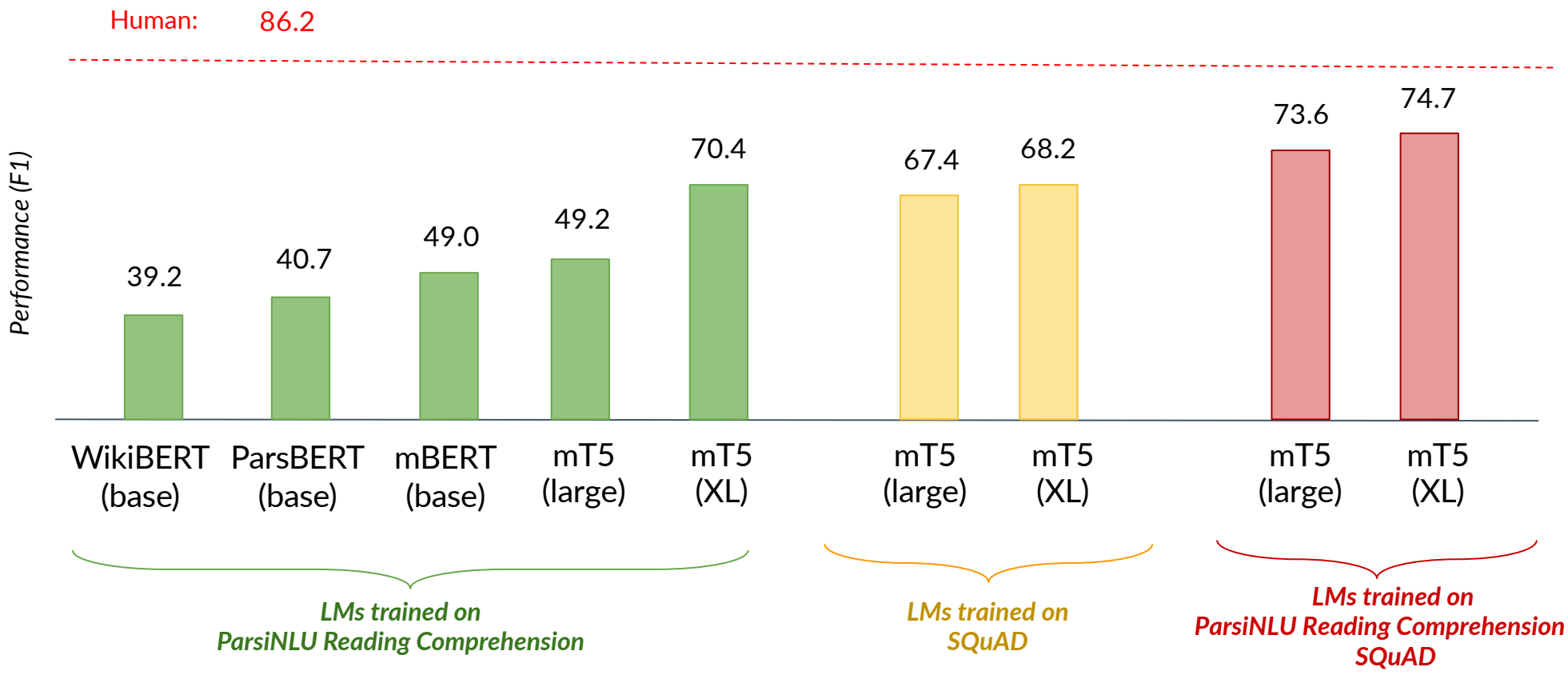
Experimental Findings (2)

- **Finding 1:** The proposed dataset(s) is decent quality.
- **Finding 2:** English models successfully transfer to Persian.



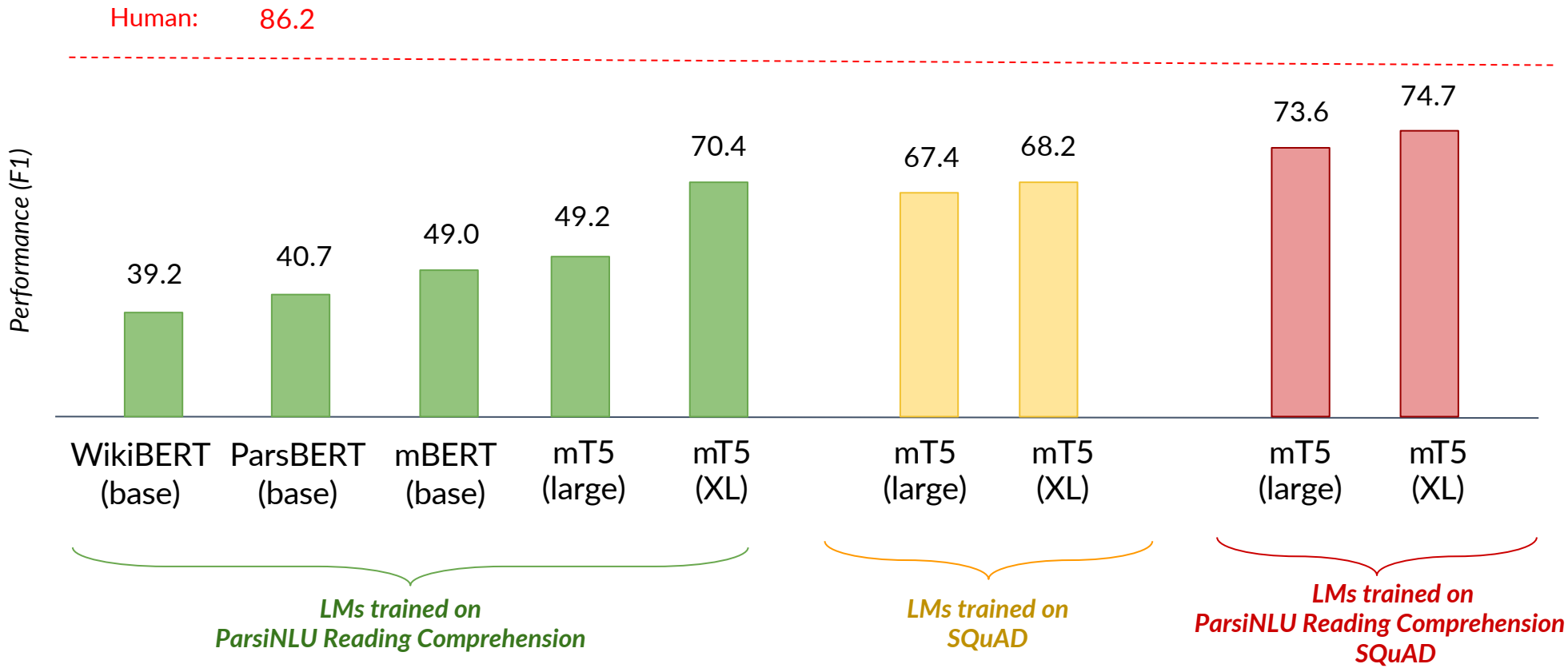
Experimental Findings (3)

- **Finding 1:** The proposed dataset(s) is decent quality.
- **Finding 2:** English models successfully transfer to Persian. Joint training on **English and Persian** helps.



Experimental Findings (4)

- **Finding 1:** The proposed dataset(s) is decent quality.
- **Finding 2:** English models successfully transfer to Persian. Joint training on **English and Persian** helps.
- **Finding 3:** ParsiNLU has room for progress.



✓ **Multi-lingual LMs** have been successful in transfer across languages.

Image-Text Models

an invisible man, wearing glasses and sitting at a desk in front of a computer"

https://openai.com/dall-e-2/

one piece of fruit that's apple on the outside, orange texture on the inside, cut in half

https://imagen.research.google/

✓ **Multi-modal** have shown successful transfer across different modalities.

Is NLP solved? Are we done? 🤔

Humans Generalize from Few Examples

The cat eats.

The cat drinks.

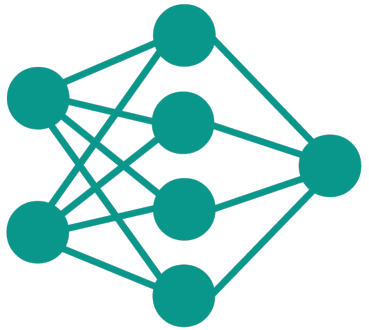
The cat sleeps.

The cat eats.

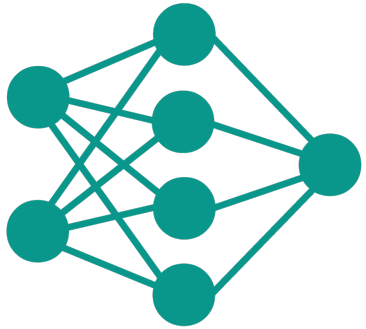
Machines Need Many Examples

Training

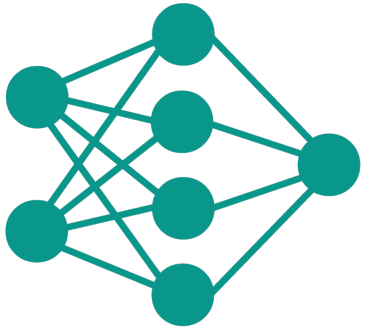
The cat eats.



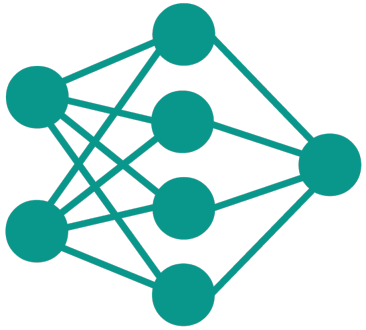
Inference



???



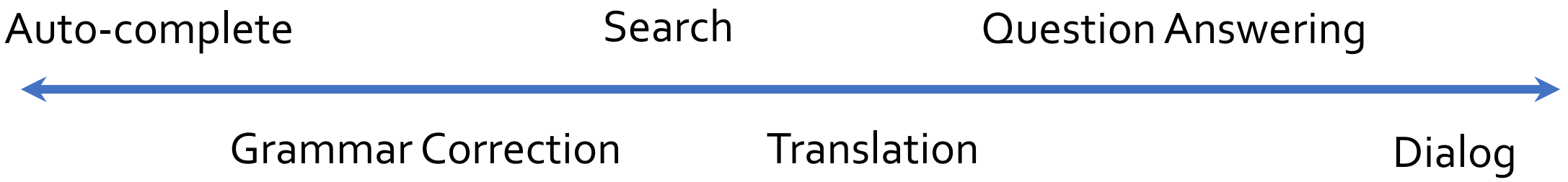
???



???

x Our computational models are **data-hungry**!

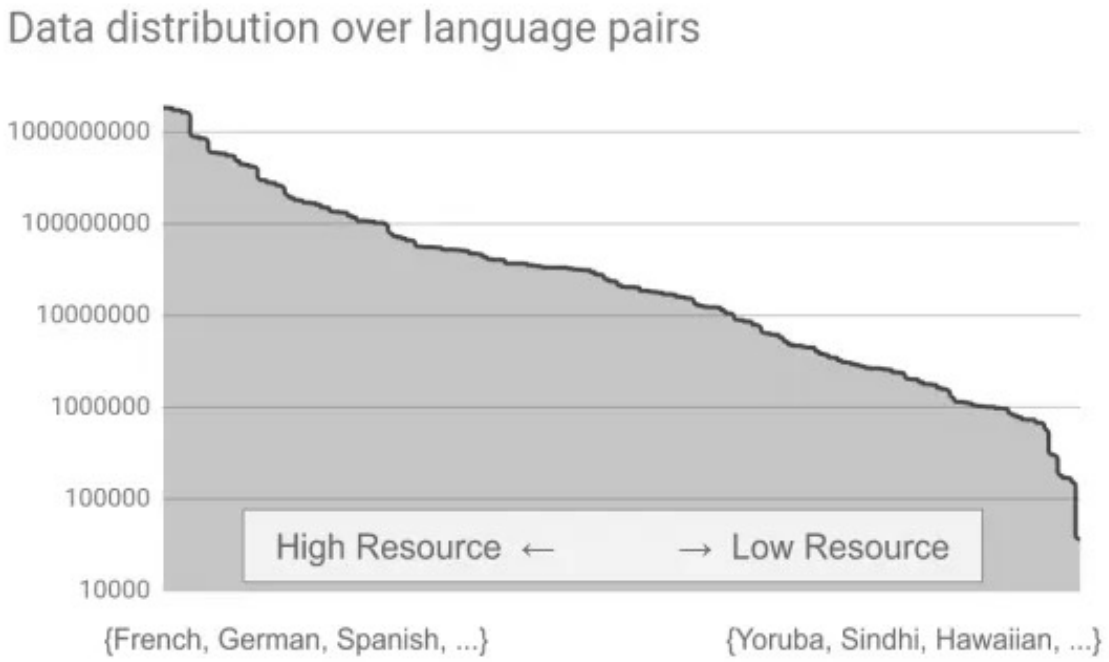
NLP in Production



They work well in domains where data is available.

Data scarcity in other languages

- **Example:** Google Translate



"The number of parallel sentences [...] ranges from around tens of thousands to almost 2 billion."

Massively multilingual neural machine translation in the wild: Findings and challenges. Arivazhagan, Naveen, et al. arXiv 2019.

Brittleness with respect to small changes

Context: *In the United States especially, several high-profile cases such as Debra LaFave, Pamela Rogers, and Mary Kay Letourneau have caused increased scrutiny on teacher misconduct.*

Question: *What has been the result of this publicity?*

What's

→ *"increased scrutiny on teacher misconduct"*

→ *"teacher misconduct"*

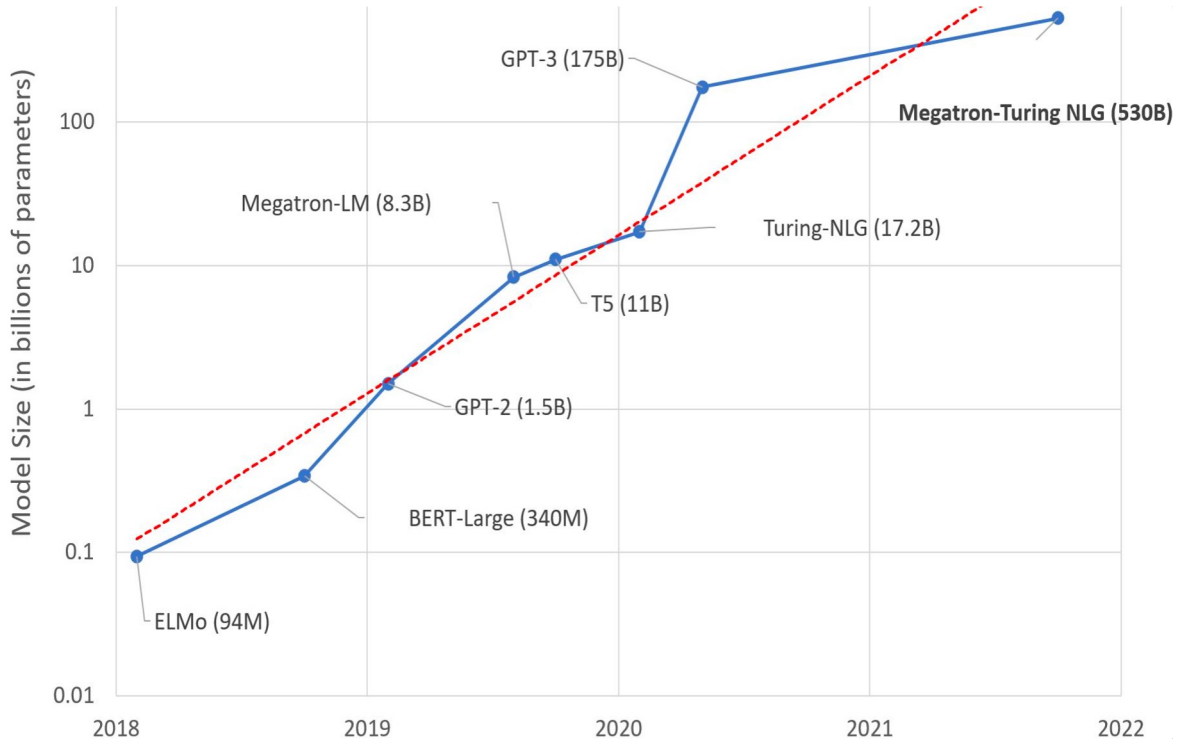
[K at al. 2016; Jia et al. 2017; Gardner et al. 20; others]

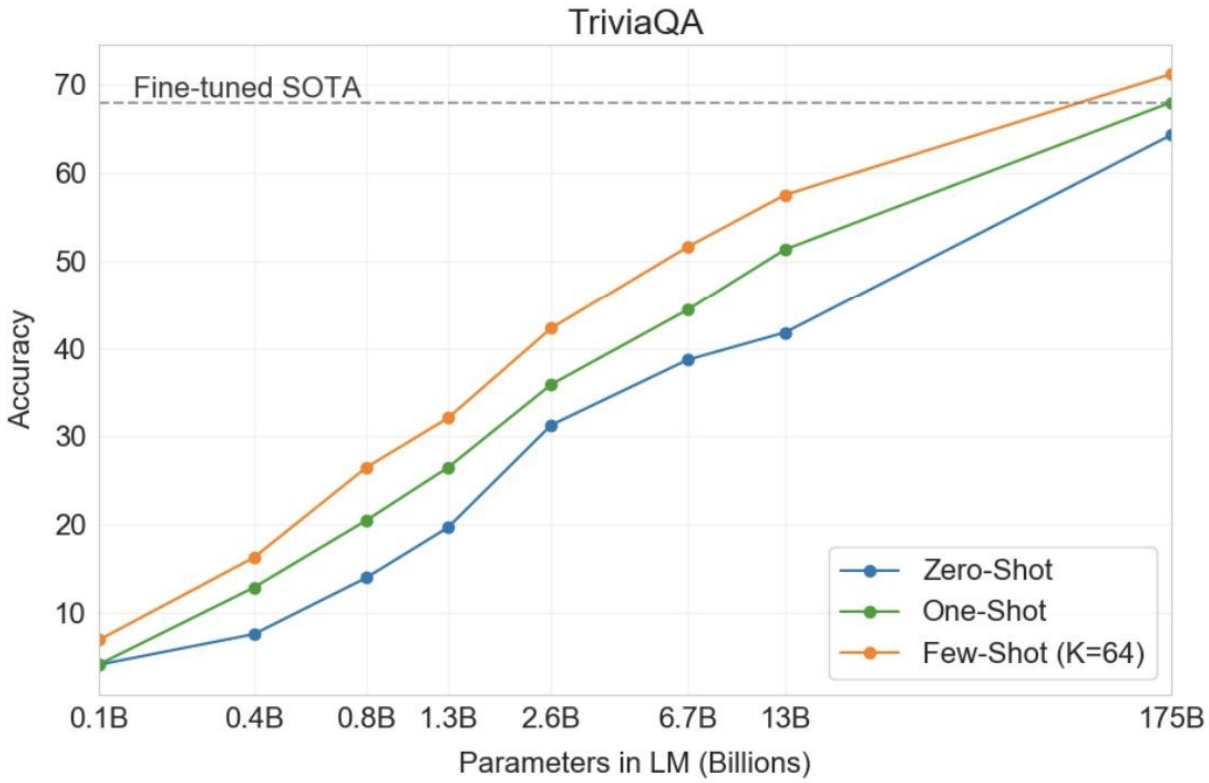
Brittleness with respect to small changes

"Tesla's Autopilot system confusing horse-drawn carriage for truck"

x Limited data lead to brittle models!

Problem: Diminishing Returns of Scaling





x Scaling models is costly.
x There are diminishing returns.

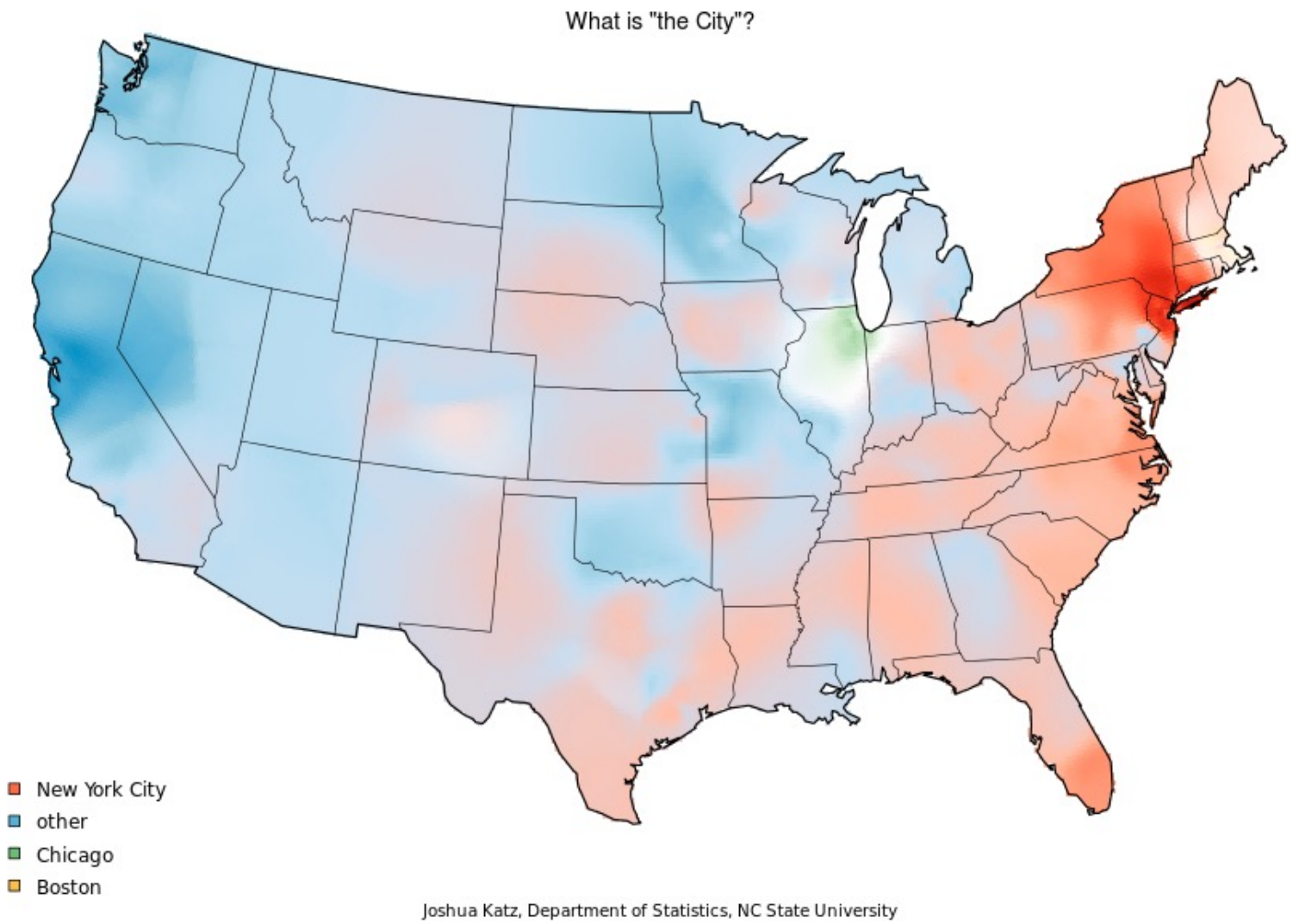
Open question: how far can we scale?

Language Interpretation Is Hard

	What I say	What I mean
Implicit meaning	*I didn't eat anything since this morning*	*I am hungry*
Non-literal meaning	*I can eat a horse*	*I'm really hungry (even though can't actually eat a horse)*
Pragmatics	*Do you have some food here?*	*Please give me some*
Common background	*There is a new restaurant in the city*	*… in the city we're are in now.*

But what is “context”?

- ”Context” is everything
  - Seeing
  - Touching
  - Hearing
  - Reading



x Computers have
a **narrow** access to the world!

LMs are not Consistent

They do not know what they know!

How many birds?	**A:** 1
Is there 1 bird?	**A:** no
Are there 2 birds?	**A:** yes
Are there any birds?	**A:** no

(a) Input image from the **VQA dataset**.

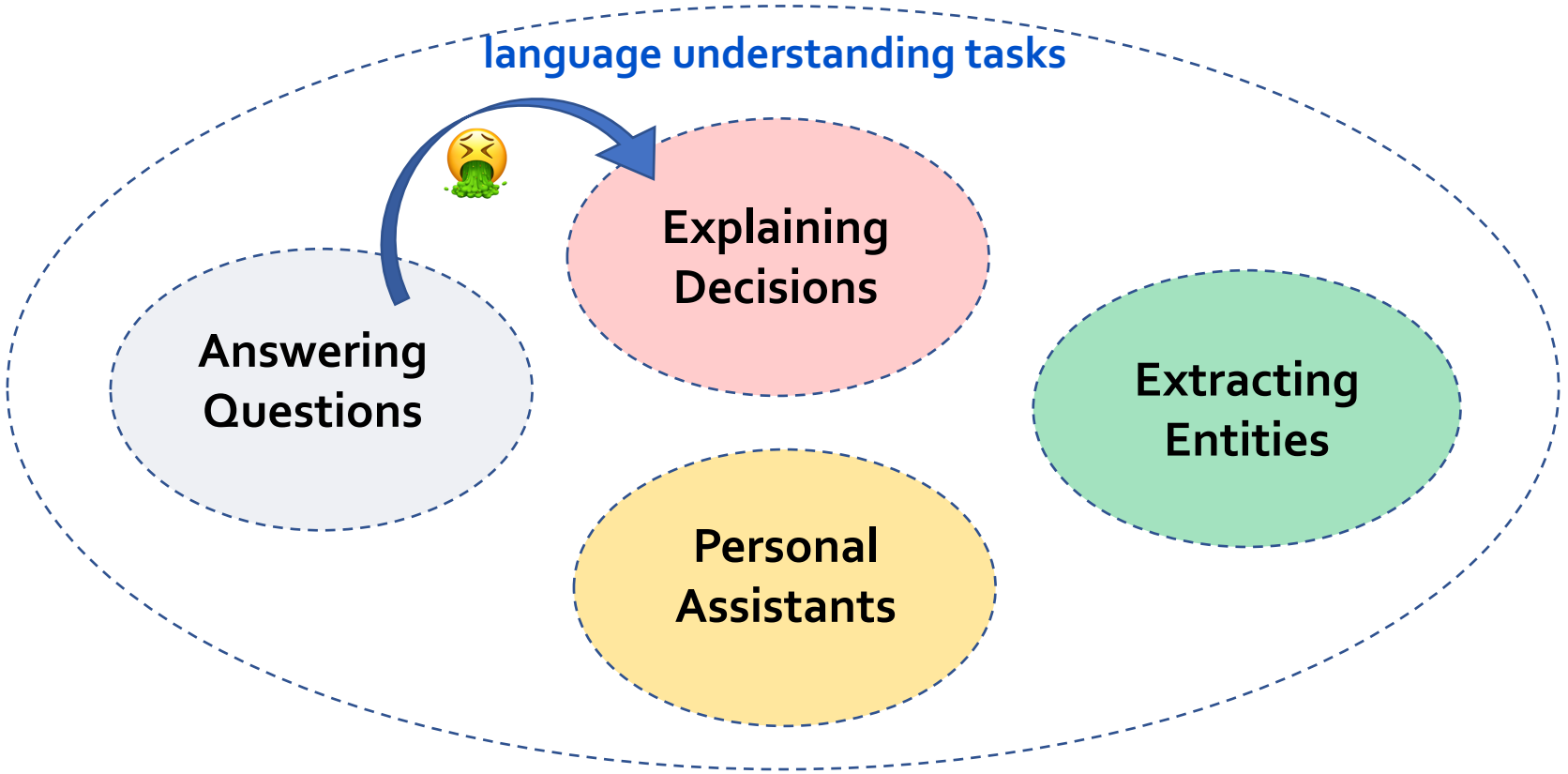
(b) Model (Zhang et al., 2018) provides inconsistent answers.

Ribeiro et al, 2019

x LMs do not know if they [don't] know!

Lack of Generality

- Successes in NLP are focused on niche domains



x Current Successes of AI have limited generality!

Interactive Semantics

Single-shot evaluation

input problem

evaluation

Learning from interactions

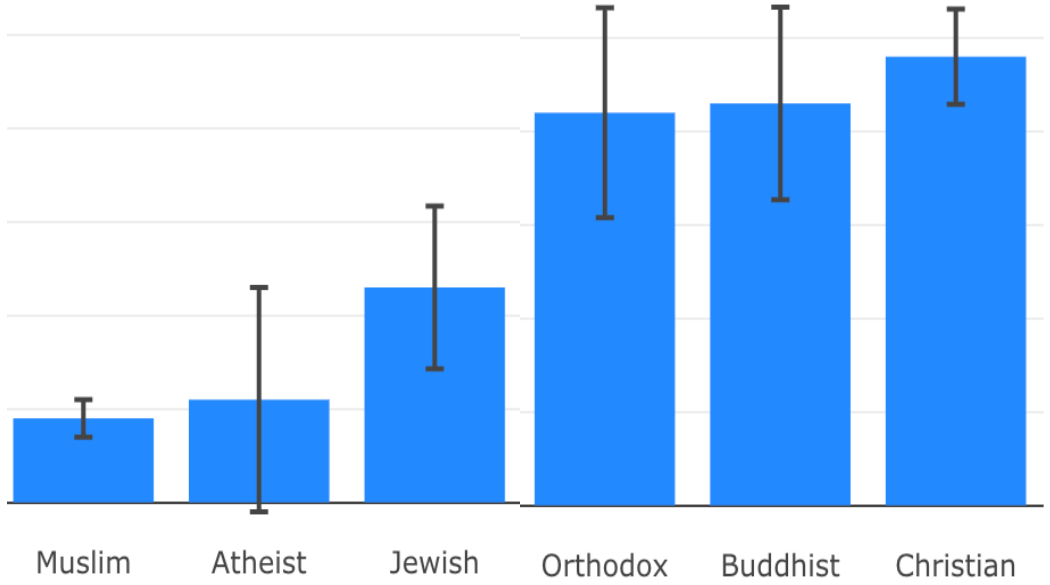
x How can we incorporate "interactivity" in AI/NLP?

Language Models: Biases

- What does this mean for the NLP systems built out of such systems?
- Discovery:
  - How can we automate the discovery of issues?
- Mitigation:
  - How can we resolve the such biases?

Social Biases of QA Models

Social Biases in QA Models [Li et al. EMNLP-Findings 2020]



Association of ethylic/religious groups with negative stereotypes

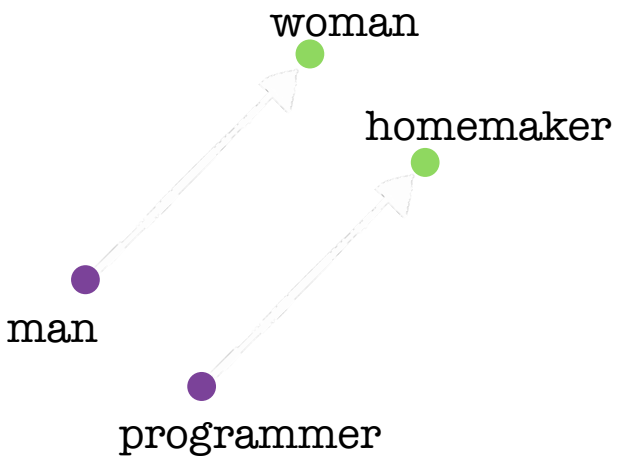
https://unqover.apps.allenai.org/

What could go wrong?

1

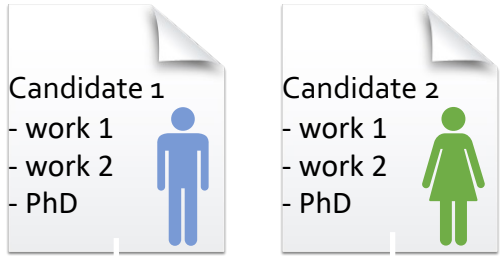
Biased Supervision and/or underrepresented groups in the training set

2



Biased Input Representations

3



CV filtering model

Uninterpretable Features

x Models perpetuate social biases!

AI Checklist: Building models that …

- ✓ seamlessly learn input-output mappings.
- ✓ don’t need to be trained from scratch every time.
- ✓ are context-sensitive.
- ✓ use fewer training examples.
- x understand the emergent ability of language models.
- x know what they know.
- x capture accurate common sense knowledge.
- x are interpretable.
- x don’t perpetuate social biases.

And finally ….

Are we climbing a tree to reach to the moon?

That's it!