

Probabilistic Commonsense Knowledge Evaluation

Xiang Lorraine Li

Assistant Professor at SCI Pitt

UMassAmherst

Robert and Donna Manning
College of Information
& Computer Sciences



Impressive Progress in AI



TECHNOLOGY

The AI That Has Nothing to Learn From Humans

DeepMind's new self-taught Go player has defeated other players designed by humans.

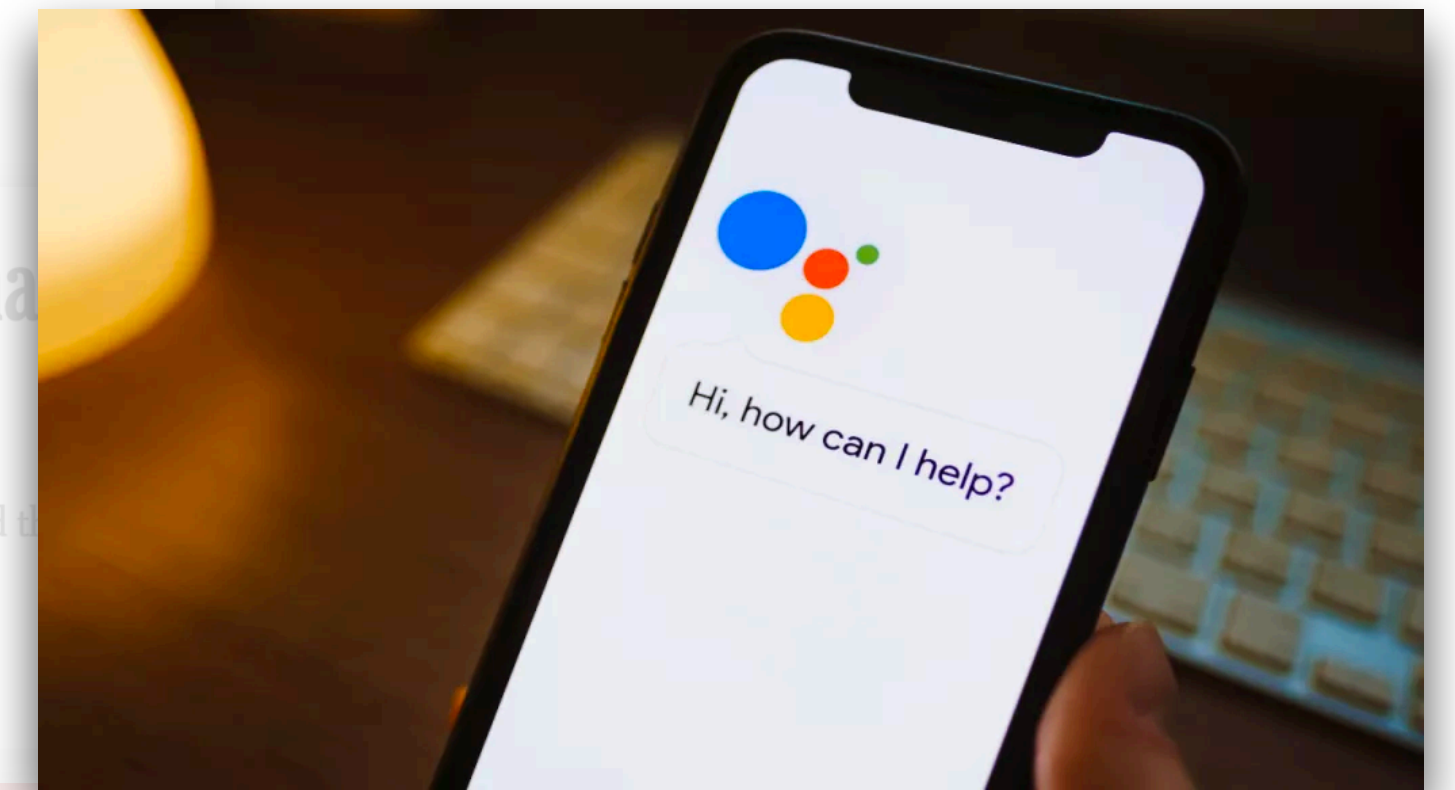
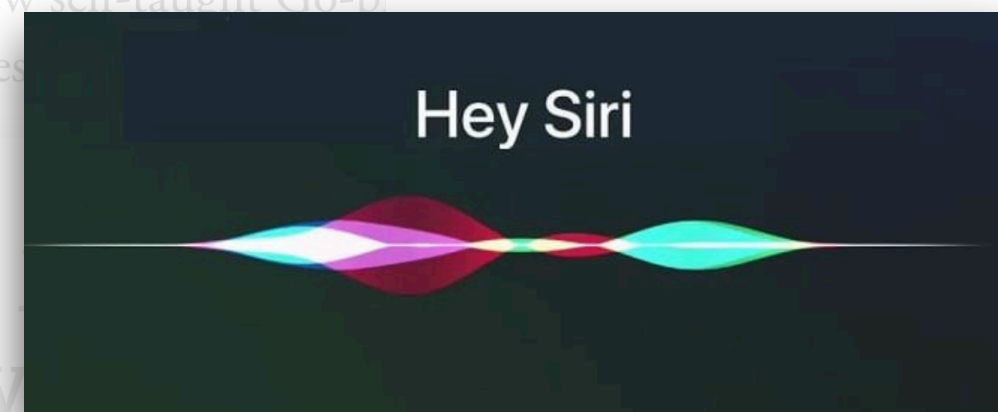
2,764 views

How a Human Can Do This?

ability to write, and the ability to be really good at it.

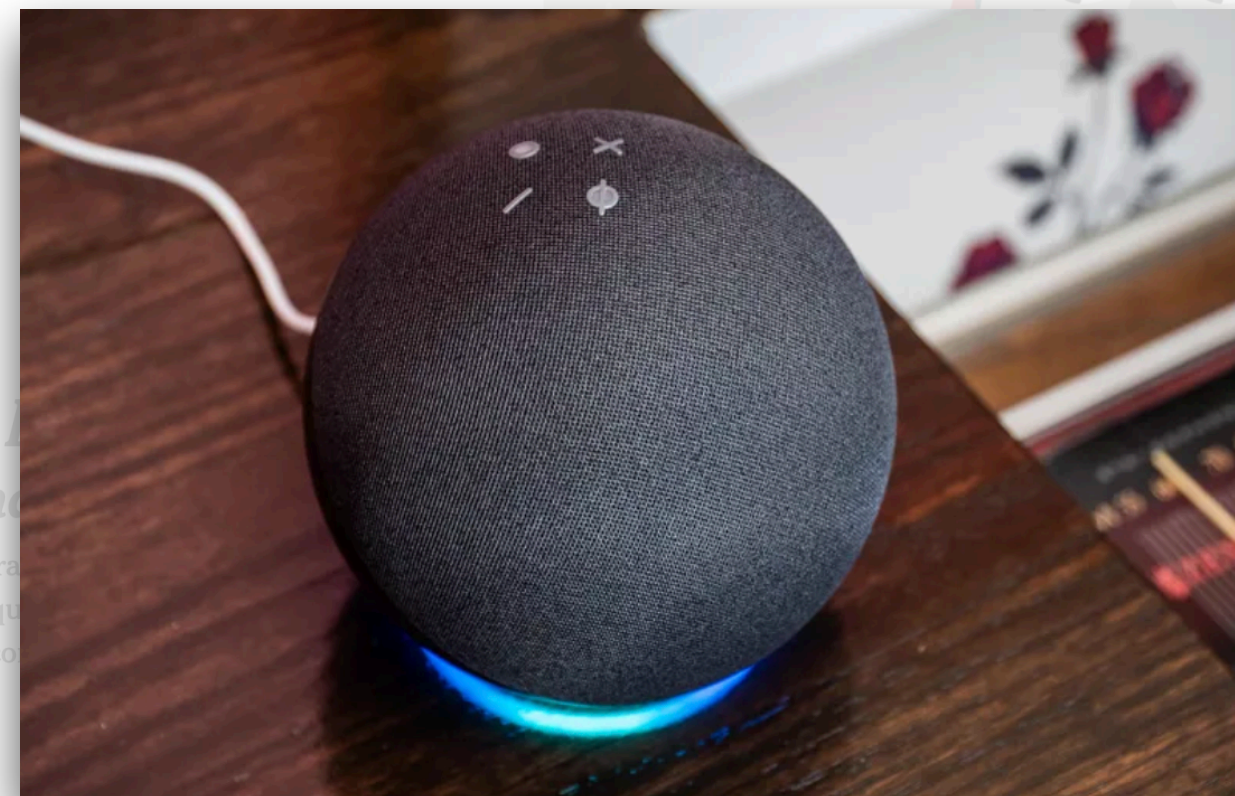
July 29, 2020

The New York Times

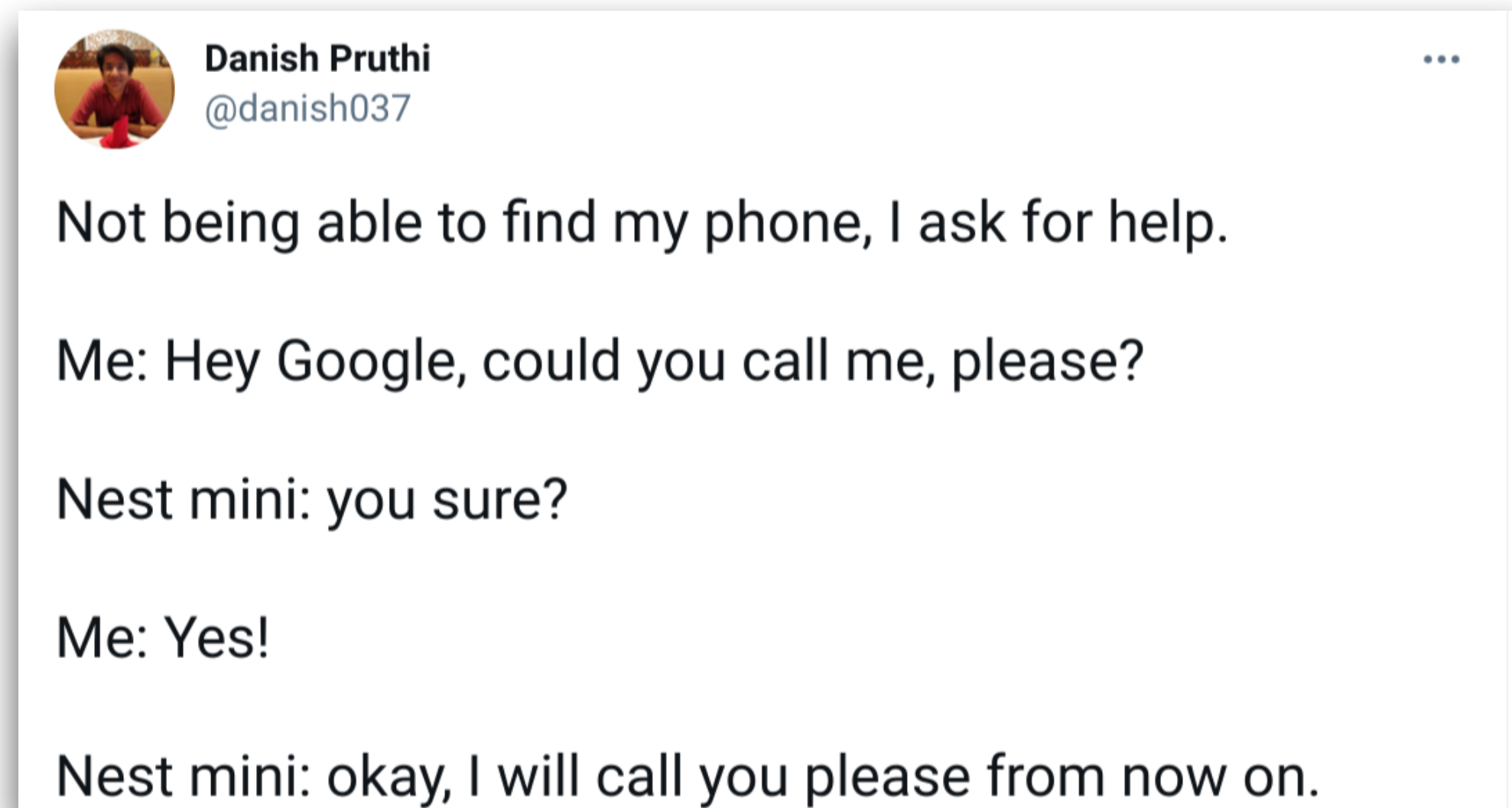


Meet GPT-3. It Has
Code (and Blog and

The latest natural-language system generates text that summarizes emails, answers trivia questions, translates languages and even writes its own code.



Impressive Progress in AI



The key aspect of **successful, clear** and **effective** interaction is handling implicit information, the information that is unstated in those situations — **Common Sense**.

Machines Need Common Sense!

What is Common Sense?

They boiled the water.

What is Common Sense?

Shared

They boiled the **water**.

What is Common Sense?

Shared

Water is liquid.

Water can be used for cleaning.

Water can be found in river.

Water can be used to wash clothes.

Humans drink water.

Water evaporates.

Water is wet.

They boiled the **water**.

Water needs to be held in a container.

What is Common Sense?

Shared

Water is liquid.

Water can be used for cleaning.

Water can be found in river.

Water can be used to wash clothes.

Humans drink water.

Water evaporates.

Water is wet.

They **boiled** the **water**.

Water needs to be held in a container.

What is Common Sense?

Shared

Water is liquid.

Water can be used for cleaning.

Water can be found in river.

Water can be used to wash clothes.

Humans drink water.

Water evaporates.

Water is wet.

They **boiled** the **water**.

Water needs to be held in a container.

Boiled water is too hot to drink.

Heat is needed to boil water.

Boiled water can cook food.

Burner can provide heat.

What is Common Sense?

Shared

Implicit

Everyday Matters

Water is liquid.

Water can be used for cleaning.

Water can be found in river.

Water can be used to wash clothes.

Humans drink water.

Water evaporates.

Water is wet.

They **boiled** the **water**.

Water needs to be held in a container.

Boiled water is too hot to drink.

Heat is needed to boil water.

Boiled water can cook food.

Burner can provide heat.

Why is Common Sense Challenging?

Water is liquid.

Water can be used for cleaning.

Water can be found in river.

Water can be used to wash clothes.

Humans drink water.

Water evaporates.

Water is wet.

They **boiled** the **water**.

Water needs to be held in a container.

Boiled water is too hot to drink.

Heat is needed to boil water.

Boiled water can cook food.

Burner can provide heat.

Why is Common Sense Challenging?

Massive

Humans drink water. Boiled water can cook food. Water can be found in river. Water can be used for cleaning. Sugar can melt in water. Boiled water can cook food. Water needs to be held in a container.

Water is liquid. **Open the jelly jar.** Heat is needed to boil water. There are usually waiter helping you order food. Heat is needed to boil water. When it's cloudy, sometimes there is no sunset. People are walking along the river bank.

can provide heat. Human needs water to live. People who wants to lose weight usually avoid peanut butter. Sweet water tastes good. Sunset time is usually in the afternoon. Sunset can be beautiful.

Humans drink water. Water needs to be held in a container. Sugar water is wet. There is water in the river. River water is not directly drinkable. Water can be used for

Heat is needed to boil water. Water can be used for cleaning. Water needs to be held in a container. Human feel satisfied after having sweet stuff. **They walked along the river at sunset time.** Water needs to be h

Heat is needed to boil water. Water can be used to wash clothes. Most bread is not sweet. Sugar water is also liquid. People are walking a

Boiled water can cook food. Boiled water is too hot to drink. Water needs to be held in a container. Water needs to be held in a container. Humans drink water.

Water can be used for cleaning. Water can be used to wash clothes. Water can be used for cleaning.

Sweet water tastes good. Boiled water can cook food. Water needs to be held in a container. Water can be used for cleaning.

Water is wet. Human feel satisfied after having sweet stuff. **They boiled the water.** Water is wet. Humans drink water. There are usually waiter helping you order food.

People needs tools to put peanut butter on the bread. Opening a jar needs tool. Person can open jar, but not dogs. Sometimes the ordering is automatic too. Order food means choosing dishes on the menu.

A knife with peanut butter could be the tool. Human can put peanut butter on the bread. **Spread the peanut butter on the bread.** Heat is needed to boil water. She walked into a restaurant and started ordering. Boiled water can cook food. Sometimes the ordering is automatic.

Some people love sugar. Peanut butter on the bread is usually breakfast. Human feel satisfied after having sweet stuff. People walk into restaurant through door. Person can open jar, but not cats.

People who wants to lose weight usually avoid peanut butter. Water can be used for cleaning. Ordering food needs menu. Restaurant serves food.

Sweet water tastes good. Peanut butter is high calorie food. Water is wet. Boiled water can cook food. Walking into a restaurant usually at breakfast/lunch/dinner time.

Human feel satisfied after having sweet stuff. Some people hate sugar. Most bread is not sweet. Water needs to be held in a container. People walk into restaurant through door. Boiled water can cook food.

Peanut butter can be spread. Some people are allergic to peanut butter. There is water in the river. Peanut butter is sweet. **Tom asked me how to get to the library.** Ordering food needs menu. Water is wet. Boiled water can cook food.

Some people hate peanut butter. The kind of bread that can add peanut butter is flat. A knife with peanut butter could be the tool. Walking into a restaurant usually at breakfast/lunch/dinner time.

Allergy reactions can be very serious, life-threatening. Bread with peanut butter can be satisfying. Water needs to be held in a container.

Why is Common Sense Challenging?

Massive

Food Chemistry
Volume 303, 15 January 2020, 125385

Melatonin treatment maintains nutraceutical properties of pomegranate fruits during cold storage

Morteza Soleimani Aghdam, Zisheng Luo, Li Li, Abbasali Jannatizadeh, Javad Rezapour Fard, Farhad Pirzad

Highlights

- Sufficient supply of intracellular NADPH may be due to the combined activities provided by G6PDH and 6PGDH.

Article Talk

COVID-19 pandemic

From Wikipedia, the free encyclopedia

The **COVID-19 pandemic**, also known as the **coronavirus pandemic**, is an ongoing global pandemic of **severe acute respiratory syndrome coronavirus 2** (SARS-CoV-2). The novel virus emerged in Wuhan, China, in late 2019 and spread across the world in early 2020, becoming the most contagious and deadly. Severe illness is more likely in older people, particularly indoors in poorly ventilated areas. Transmission rarely occurs via contaminated surfaces, but is highly contagious for 10 days, often beginning before or without symptoms. Mutations have produced many variants, and widely distributed in various countries since December 2020. Other recommended preventive measures include wearing a face mask and avoiding contact with those who have been exposed or are symptomatic.

What do scientists think this week?

[Nutraceutical properties of lycopene]

[Article in Spanish]
Krzysztof N Waliszewski, Gabriela Blasco

Affiliations + expand
PMID: 20485889 DOI: 10.1590/s0036-36342010000300010

Abstract
In recent years, dietary recommendations have suggested an increase in the consumption of foods that contain phytochemicals that provide benefits to human health and play an important role in preventing chronic diseases. Lycopene -the carotenoid responsible for the red color of tomatoes- has attracted attention because of its physicochemical and biological properties in the prevention of chronic diseases in which oxidative stress is a major etiological factor, such as cancer, cardiovascular and neurodegenerative diseases, and hypertension, among others. Antioxidants, including lycopene, interact with reactive oxygen species, can mitigate their damaging effects and play a significant role in preventing these diseases. This article presents a review of some epidemiological studies published in recent years on beneficial effects of lycopene in human health.

Out of Asia: mitochondrial evolutionary history of the globally introduced supralittoral isopod *Ligia exotica*

Luis A. Hurtado, Mariana Mateos, Chang Wang, Carlos A. Santamaria, Jongwoo Jung, Valiallah Khalaji-Pirbaltouty and Won Kim

Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX, United States of America
Department of Biology, New York University, New York City, NY, United States of America
Biology Faculty, College of Science and Mathematics, University of South Florida, Sarasota, FL, United States of America
Department of Science Education, Esha Women's University, Seoul, South Korea
Department of Biology, Shahrood University, Shahrood, Iran
School of Biological Sciences, Seoul National University, Seoul, South Korea

Abstract
Principal component analysis is a widely used method for the of a given data set in a high-dimensional Euclidean space. Here, we study the Stiefel tropical linear space of fit the data points in the tropical projective torus; in the other

Amazon's Alexa Just Gave A Lethal Challenge To A 10-Year-Old

Amazon Alexa reportedly told a child to do a potentially lethal challenge. Fortunately, the kid is safe, and the company has fixed the glitch.

BY KISHALAYA KUNDU
PUBLISHED 5 DAYS AGO

Donald Trump business, Jan. 3, 2022

The New York Times business children

The involvement of Trump, Trump generation his child



COP26 is seen as crucial if climate change is to be brought under control

As the COP26 climate summit enters its second week, negotiations in Glasgow have hit a critical phase.

The conference is seen as crucial if climate change is to be brought under control. So we asked more than a dozen climate scientists, negotiators and economists from around the world what they wanted to see agreed this week.

Cut emissions now

The scientists all wanted to see more countries commit to net zero by 2050 at the latest. Yet many said changes in the next decade would be the most impactful.

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, ... Demis Hassabis + Show authors

Nature 596, 583–589 (2021) | Cite this article

502k Accesses | 568 Citations | 2962 Altmetric | Metrics

Abstract
Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort, the structures of around 100,000 unique proteins have been determined, but this represents a small fraction of the billions of known protein sequences. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the ‘protein folding problem’—has been an important open research problem for more than 50 years. Despite recent progress, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14), demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a

ARTICLE INFO

Article history
Received 10 June 2016
Received in revised form 13 February 2017
Accepted 28 February 2017
Available online 14 March 2017

Keywords
MERS-CoV
DNA vaccine
Spike protein

1. Introduction
Middle East respiratory syndrome (MERS)-coronavirus (MERS-CoV) is a novel human coronavirus that causes human respiratory disease. The development of a detection method for this virus that can lead to rapid and accurate diagnosis would be significant. In this study, we established a nucleic acid visualization technique that combines the reverse transcription loop-mediated isothermal amplification technique and a vertical flow visualization strip (RT-LAMP-VF) to detect the N gene of MERS-CoV. The RT-LAMP-VF assay was performed in a constant temperature water bath for 30 min, and the result was visible by the naked eye within 5 min. The RT-LAMP-VF assay was capable of detecting 2 × 10³ copies/μl of synthesized RNA transcript and 1 × 10³ copies/μl of MERS-CoV RNA. The method exhibits no cross-reactivities with multiple CoVs including SARS-related (SARS)-CoV, HKU1, HKU1, OC43 and 229E, and thus exhibits high specificity. Compared to the real-time RT-PCR (RT-PCR) method recommended by the World Health Organization (WHO), the RT-LAMP-VF assay is easy to handle, does not require expensive equipment and can rapidly complete detection within 35 min.

A Rapid and Specific Assay for the Detection of MERS-CoV

Pai Huang, Hualie Wang, Zengqun Cao, Hongli Jin, Hang Chi, Jincun Zhao, Baohai Wu, Fuhai Yan, Xiangping He, Fangfang Wu, Guohua Chen, Pengfei He, Shengnan Xu, Yongkun Zhao, Ma Feng, Jianzhong Wang, Weliang Sun, Tiecheng Wang, Yawei Gao, Songtao Yang, and Xianzhu Xia



AI-powered voice assistants like Amazon Alexa can be of great help, but as a recent case shows, they can also pose a grave danger to children sometimes. Alongside Google Assistant and Apple's Siri, Alexa is one of the leading voice-based digital assistants that debuted on the company's Echo smart speakers back in 2014. It is now supported by a whole host of gadgets and smart home devices, including phones, tablets, TVs, media boxes, wearables, headphones, and more.

Alexa comes with many capabilities with over 100,000 available 'skills.' It can pair with a range of home automation devices, including smart bulbs, doorbells, microwave ovens, etc. Users can also order take-outs using Alexa, stream music on a plethora of music streaming services, and even order groceries through the Amazon Fresh app. However, for all its advantages, it

0025-7125/02 \$15.00 + .00

TICK FEVER
Richard Klasco, MD

malaria Journal
Open Access
CrossMark

enegal
diaye, Denis Malvy

Advent of malaria by an increasing often have an and concurrent

su region presenting Nile (WNV), dengue fever virus (DENV) and chikungunya virus (CHIKV) analysis of single or

Why is Common Sense Challenging?

Massive

Water is liquid.

Water can be used for cleaning.

Water can be found in ocean.

Water can be used to wash clothes.

Humans drink water.

Water evaporates.

Water is wet.

They boiled the water.

Water needs to be held in a container.

Boiled water is too hot to drink.

Heat is needed to boil water.

Boiled water can cook food.

Burner can provide heat.

Why is Common Sense Challenging?

Massive

They boiled the water.

In what?

Using what?

Why is Common Sense Challenging?

Massive

They boiled the water.

In what?

Kettle

Pot

Glass

Beaker

Etc.

Using what?

Stove

Microwave

Bunsen burner

Etc.

Why is Common Sense Challenging?

Massive

They boiled the water.

In what?

Kettle

Pot

Glass

Beaker

Etc.

Using what?

Stove

Microwave

Bunsen burner

Etc.

Why is Common Sense Challenging?

Massive

Probabilistic

They boiled the water.

In what?

Kettle

Pot

Glass

Beaker

Etc.

Using what?

Stove

Microwave

Bunsen burner

Etc.

Why is Common Sense Challenging?

Massive

Probabilistic

They boiled the water
and added spaghetti.

In what?

Kettle

Pot

Glass

Beaker

Etc.

Using what?

Stove

Microwave

Bunsen burner

Etc.

Why is Common Sense Challenging?

Massive

Probabilistic

They boiled the water
and added spaghetti.

In what?

Pot

Glass

Etc.

Using what?

Stove

Etc.

Why is Common Sense Challenging?

Massive

Probabilistic

Contextual

They boiled the water
and added spaghetti.

In what?

Pot

Glass

Etc.

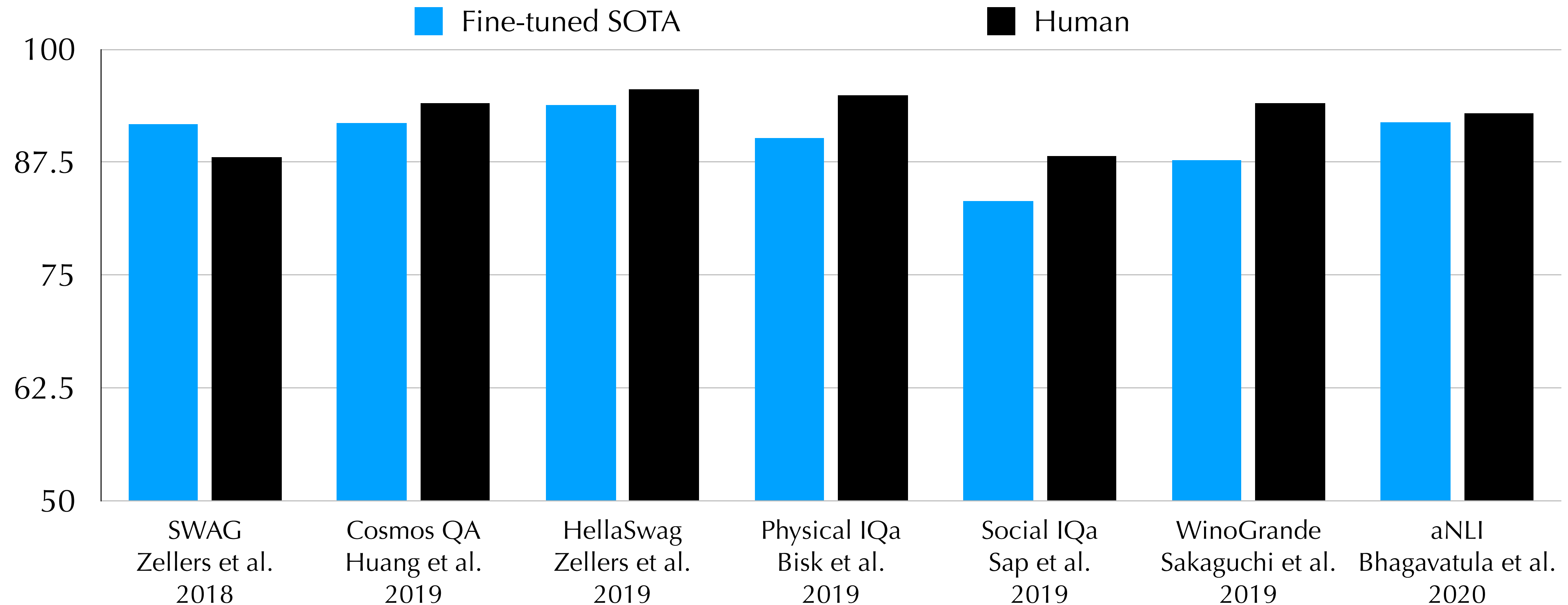
Using what?

Stove

Etc.

Common Sense in Language Model

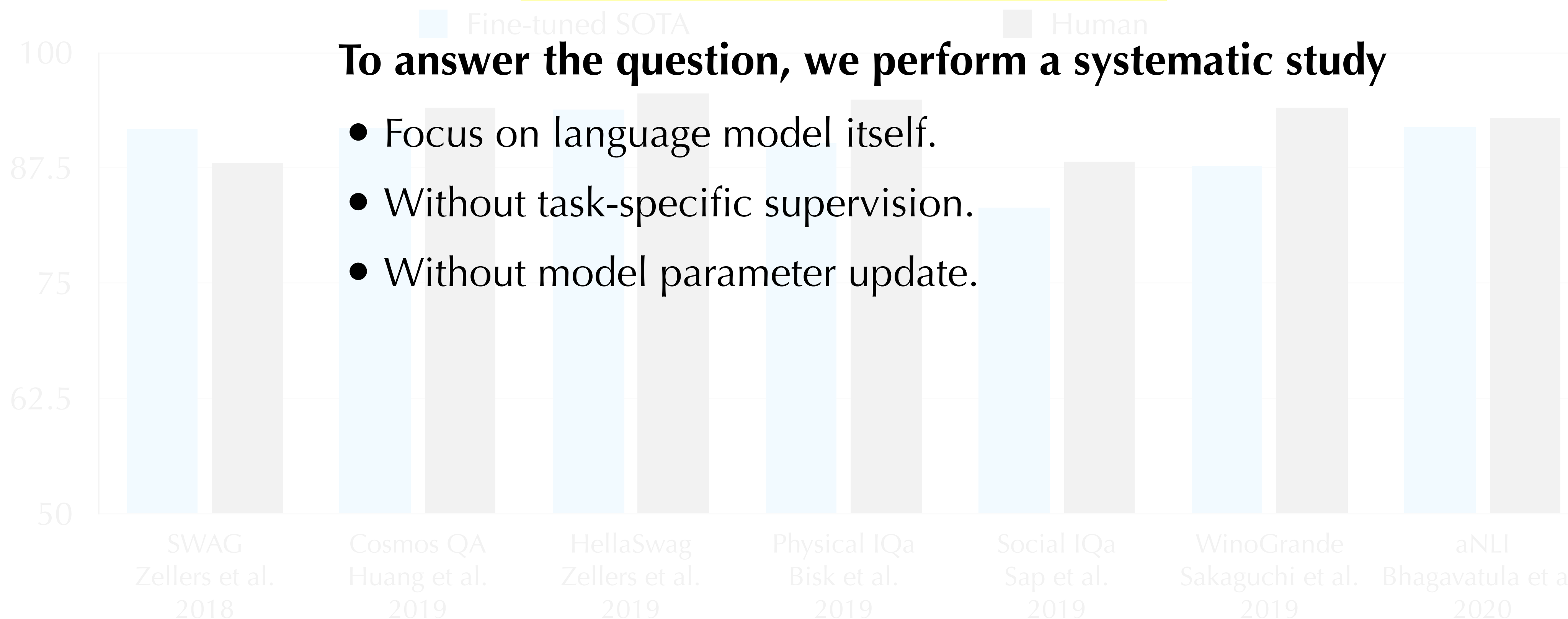
Models based on large language models show impressive performance on many **commonsense question answering** tasks.



Do language models learn common sense?

Models based on large language models show impressive performance on many commonsense question answering tasks.

Zero-shot evaluation on language models



To answer the question, we perform a systematic study

- Focus on language model itself.
- Without task-specific supervision.
- Without model parameter update.

Do language models learn common sense?

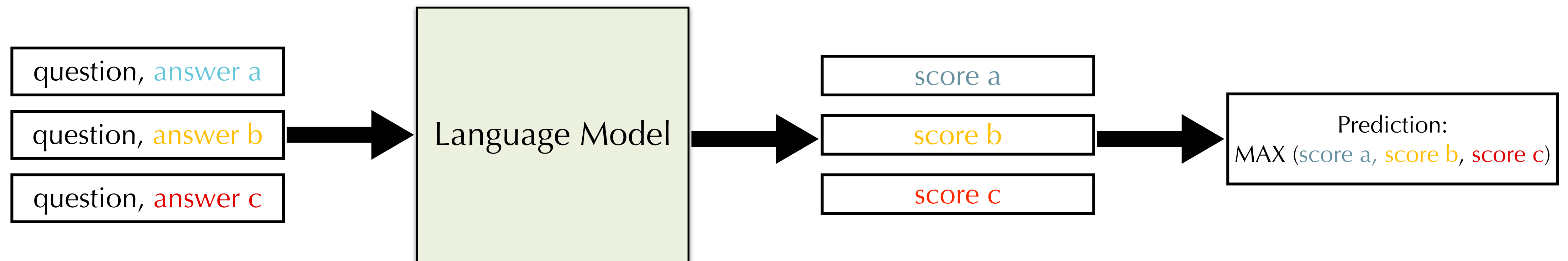
Dataset	Example	Number of Choices	Reasoning Type
Physical IQa (Bisk et al. 2019)	Question: To apply eyeshadow without a brush, should I use a cotton swag or a toothpick? Answer: Cotton swab.	2	Physical
Social IQa (Sap et al. 2019)	Question: Tracy had accidentally pressed upon Austin in the small elevator and it was awkward. Why did Tracy do this? Answer: Squeeze into the elevator	3	Social
WinoGrande (Sakaguchi et al. 2019)	Question: The trophy didn't fit the suitcase, because it is too big. What does it refers to? Answer: The trophy	2	Physical, Social etc
HellaSwag (Zellers et al. 2019)	Question: Four sentence short story. Answer: the possible ending.	4	Temporal, Physical etc

Four multiple choice selection QA datasets.

Do language models learn common sense?

Question: Tracy had accidentally pressed upon Austin in the small elevator and it was awkward. Why did Tracy do this?

- **Answer a:** get very close to Austin.
- **Answer b:** squeeze into the elevator.
- **Answer c:** get flirty with Austin.



Zero-shot Performance: random baseline

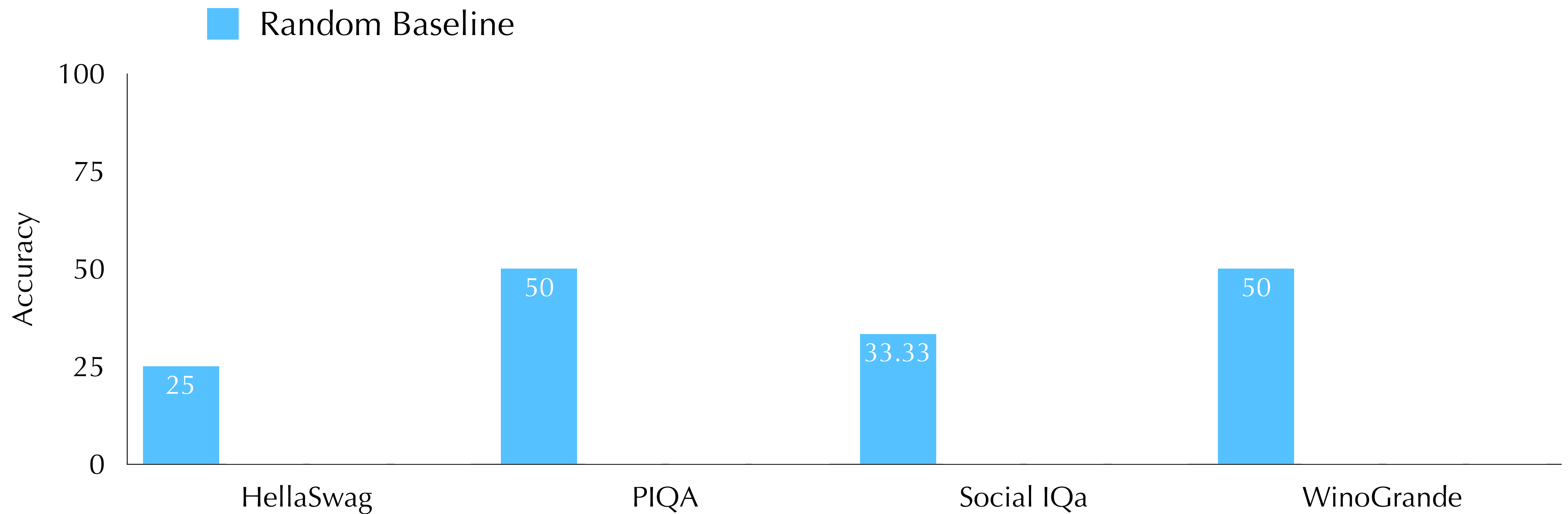


Figure: the dev accuracy for each dataset evaluated on Gopher.

Zero-Shot is not bad, especially for HellaSwag and PIQA

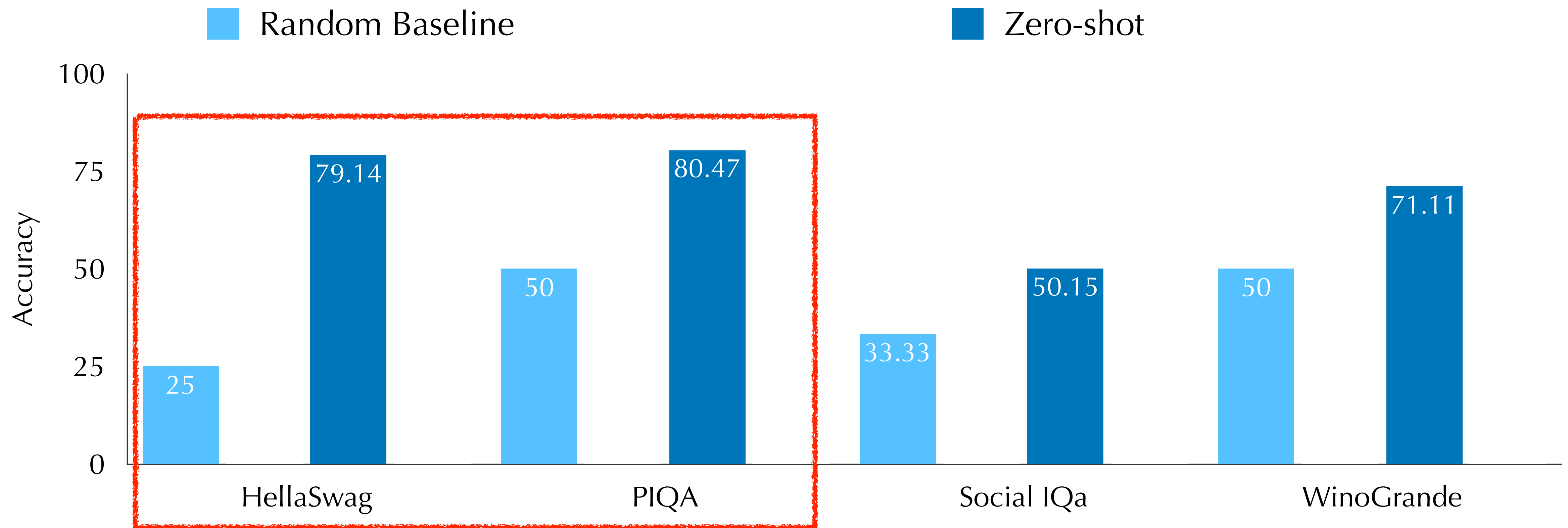


Figure: the dev accuracy for each dataset evaluated on Gopher.

How much of the performance comes **only** from **answers**?

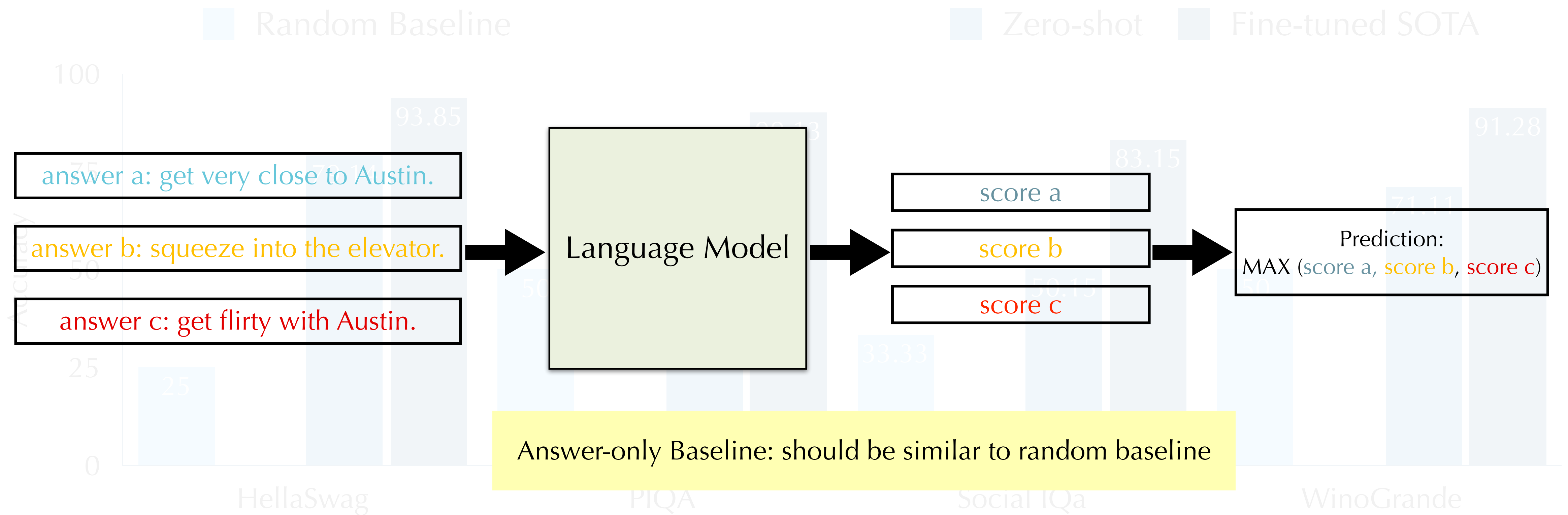


Figure: the dev accuracy for each dataset evaluated on Gopher.

Models pick the correct answer without seeing the question

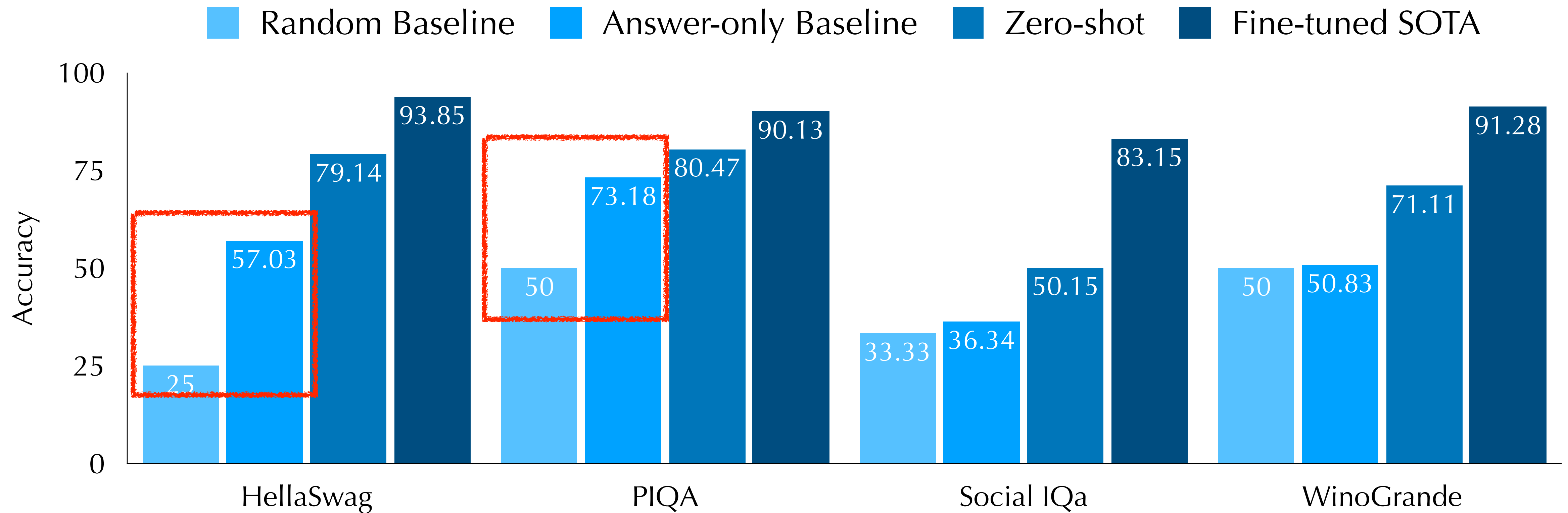


Figure: the dev accuracy for each dataset evaluated on Gopher.

We need better commonsense evaluation!

Dataset Bias!

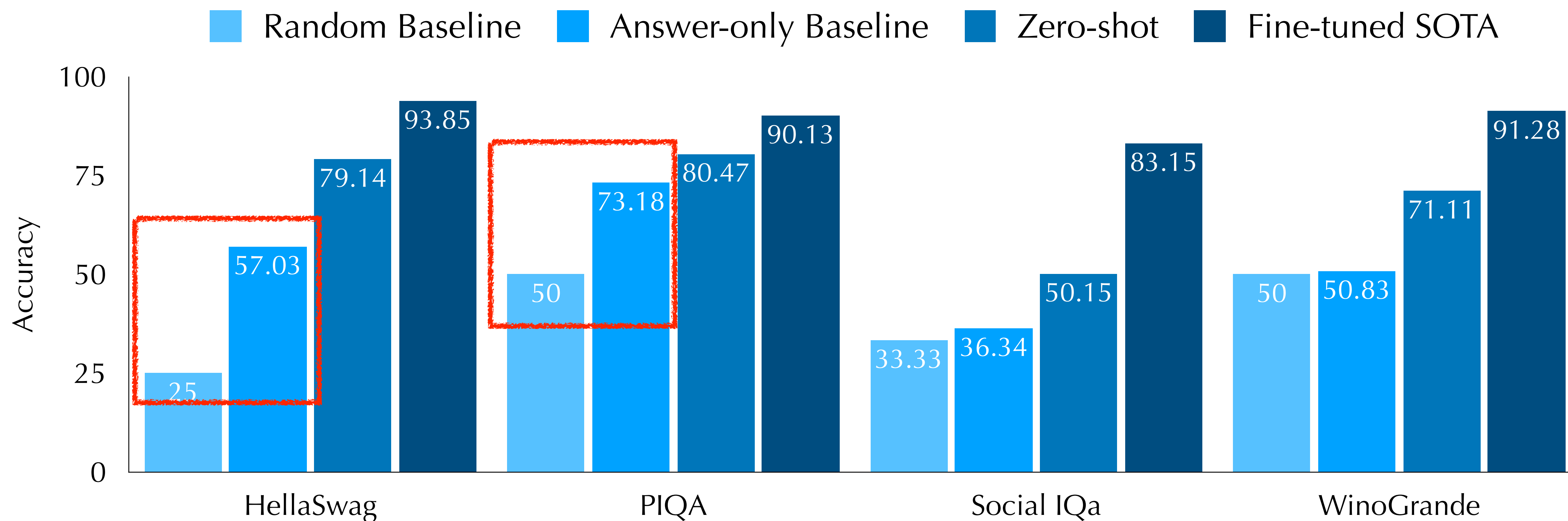


Figure: the dev accuracy for each dataset evaluated on Gopher.

Outline

Benchmark: **Probabilistic** Evaluation for Common Sense Question with **Multiple-answers**

- Every Answer Matters: Evaluating Commonsense with Probabilistic Measures. [ACL 2024]

Benchmark: **Long-tail Question:** Commonsense Reasoning Evaluation

- UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations. [NAACL 2024]

Analysis: Using Common Sense to Reason about **Complex Problems**

- Faith and Fate: Limits of Transformers on Compositionality. [NeurIPS 2023 Spotlight]

Probabilistic Evaluation of Commonsense

They boiled the water.



In what?

Kettle

Pot

Glass

Beaker

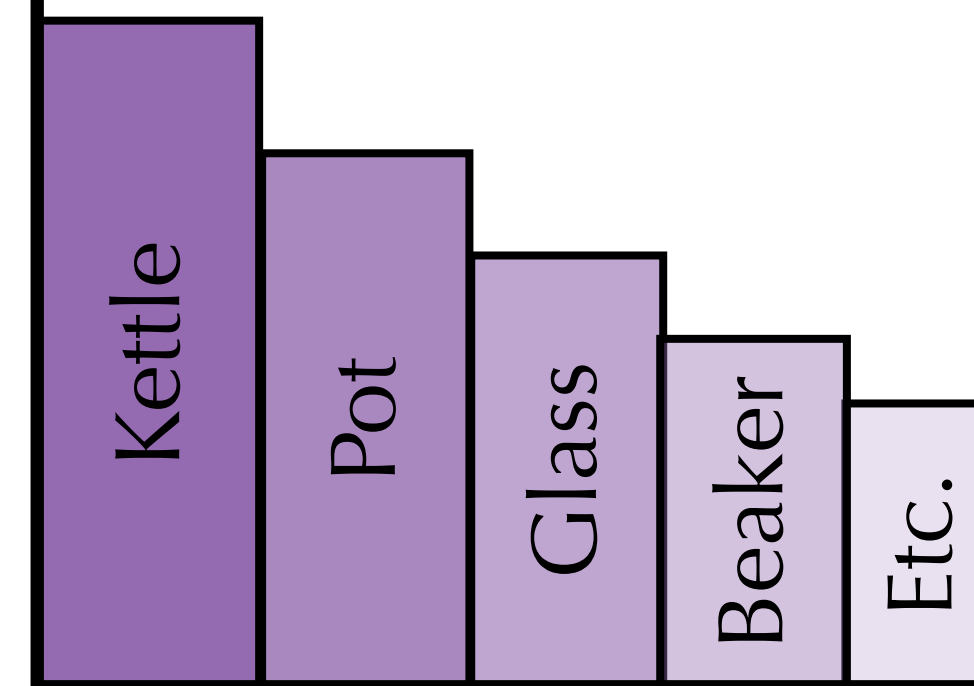
Etc.

Probabilistic Evaluation of Commonsense

They boiled the water.

In what?

Probability



Answers

Question Answering

Dialogue

Any language tasks!

CFC Data Collection

We crowd-source high-quality evaluation data

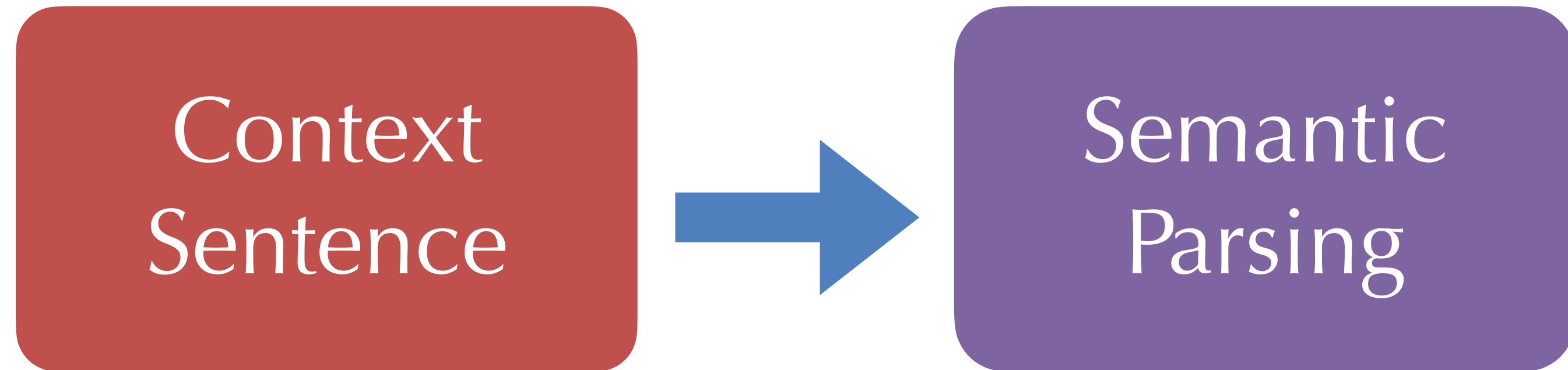
Context
Sentence

“Dog catches the
thrown frisbee.”

CommonGen (Image Captions)

CFC Data Collection

We crowd-source high-quality evaluation data



“Dog catches the thrown frisbee.”



CommonGen (Image Captions)

AMR Parsing

CFC Data Collection

We crowd-source high-quality evaluation data



“Dog catches the thrown frisbee.”



“Who throws the frisbee?”

CommonGen (Image Captions)

AMR Parsing

AMR-unknown

CFC Data Collection

We crowd-sourced high-quality **101 questions (manual filtering)**



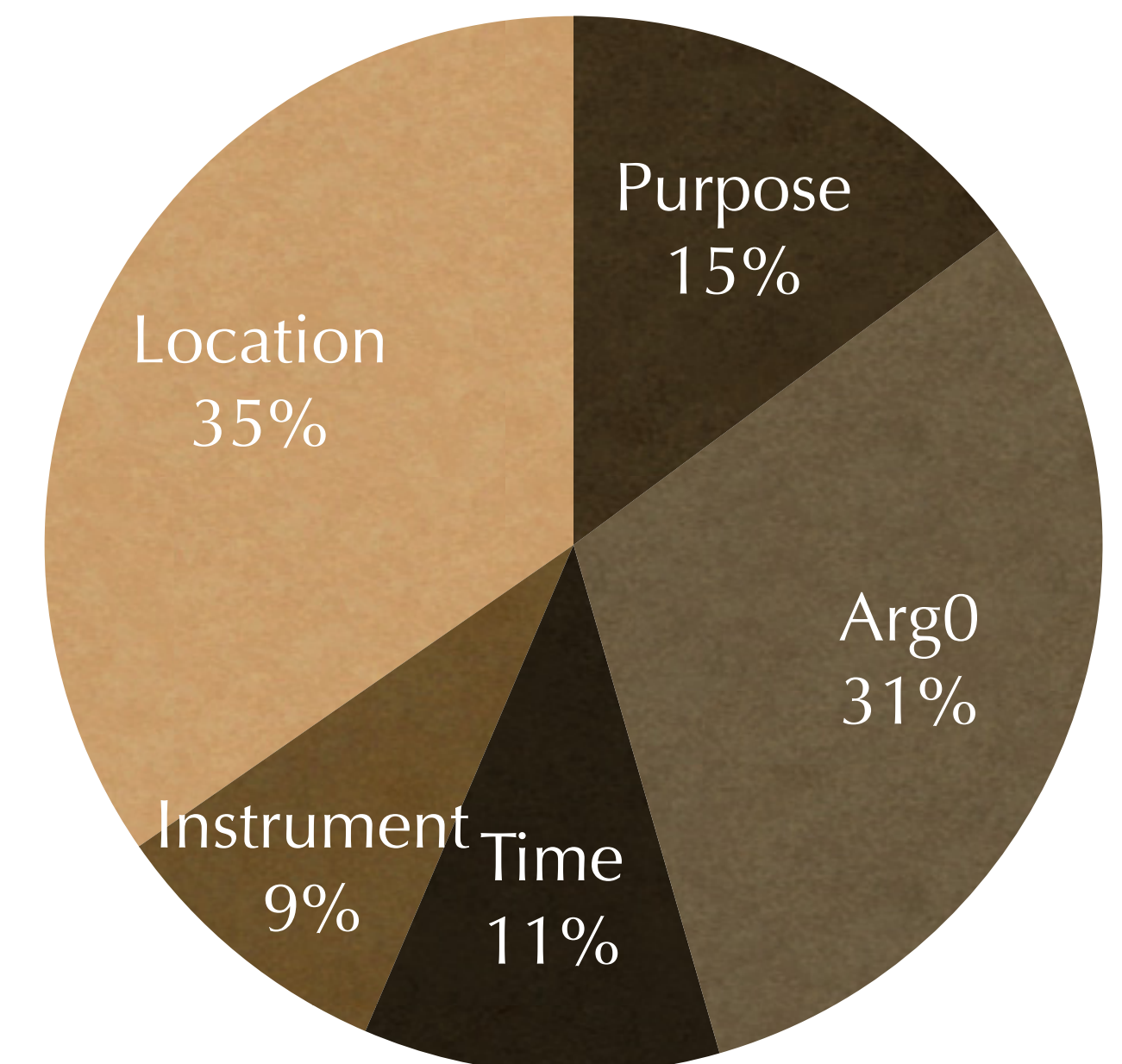
Missing Slot	Definition	Examples
Arg0	Who/what does the event?	Sentence: putting cheese on the pizza. Arg0? Answers: person, cook
Purpose	What is the goal for doing the event?	Sentence: putting cheese on the pizza. Purpose? Answers: get nutrition, stop being hungry
Instrument	What kind of tools are used to accomplish the event?	Sentence: putting cheese on the pizza. Instrument? Answers: hands, spoon
Time	What is a particular time (time of day, season, etc.) for doing the event?	Sentence: putting cheese on the pizza. Time? Answers: lunch time, dinner time
Location	Where would the event usually happen?	Sentence: putting cheese on the pizza. Location? Answers: kitchen, restaurant

CFC Data Collection

We crowd-sourced high-quality **101 questions (manual filtering)**



Missing Slot	Definition	Examples
Arg0	Who/what does the event?	Sentence: putting cheese on the pizza. Arg0? Answers: person, cook
Purpose	What is the goal for doing the event?	Sentence: putting cheese on the pizza. Purpose? Answers: get nutrition, stop being hungry
Instrument	What kind of tools are used to accomplish the event?	Sentence: putting cheese on the pizza. Instrument? Answers: hands, spoon
Time	What is a particular time (time of day, season, etc.) for doing the event?	Sentence: putting cheese on the pizza. Time? Answers: lunch time, dinner time
Location	Where would the event usually happen?	Sentence: putting cheese on the pizza. Location? Answers: kitchen, restaurant



CFC Data Collection

“They boiled the water” Purpose?

cooking clean
cook make tea disinfect disinfecting
for making tea making dinner cleaning
to cook cook food for a hot drink cleaning tools
cooking spaghetti making pasta kill bacteria
steaming vegetables purify purification
boiling potatoes make safe to drink
boiling chicken sterilization for an experiment

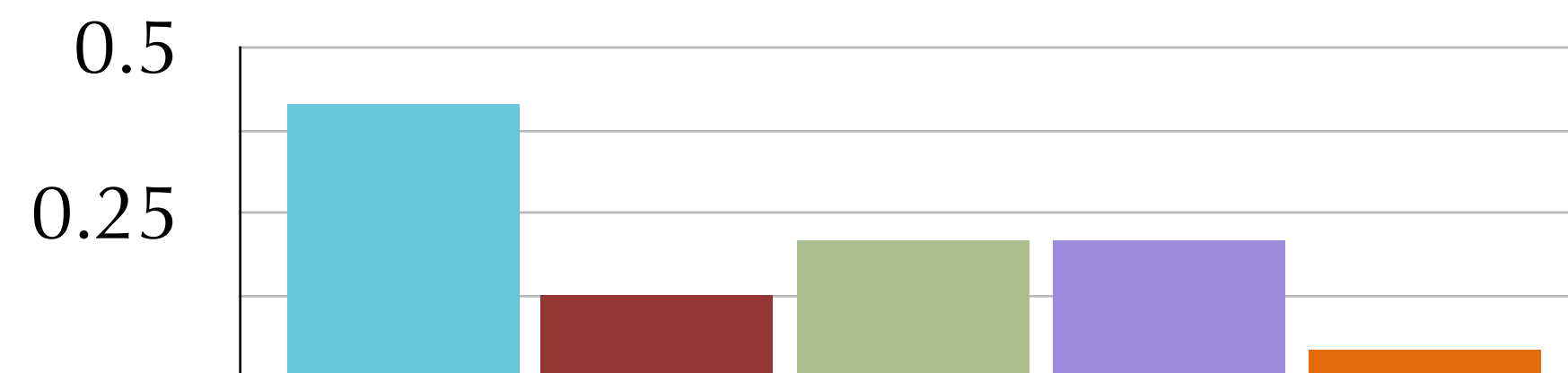


CFC Data Collection

“They boiled the water” Purpose?

for making tea
for a hot drink
clean disinfect
make tea
disinfecting
cleaning cleaning tools
cooking spaghetti making pasta
to cook steaming vegetables
boiling potatoes
kill bacteria
purify purification
cook boiling chicken
make safe to drink
cook food making dinner
sterilization
cooking for an experiment

How many answers are enough to approximate the true human answer distribution?



CFC Data Collection

How many answers are enough to approximate the true human answer distribution?

- Classic problem in statistics.
 - KL divergence between [Neyman-Pearson lemma]
➔ true distribution f and empirical sample distribution g .
 - The approximated error rate is bounded by [1]

$$\rightarrow \mathbb{P}(D_{KL}(g_{n,k}||f) \geq \epsilon) \leq e^{-n\epsilon} \left[\frac{3c_1}{c_2} \sum_{i=0}^{k-2} K_{i-1} \left(\frac{e\sqrt{n}}{2\pi} \right)^i \right]$$

CFC Data Collection

- Classic problem in statistics.
 - The approximated error rate is bounded by [1]

$$\rightarrow \mathbb{P}(D_{KL}(g_{n,k} || f) \geq \epsilon) \leq e^{-n\epsilon} \left[\frac{3c_1}{c_2} \sum_{i=0}^{k-2} K_{i-1} \left(\frac{e\sqrt{n}}{2\pi} \right)^i \right]$$

- n : number of samples
- k : number of category in the categorical distribution
- ϵ : KL error rate

How many answers are enough to approximate the true human answer distribution?

CFC Data Collection

- Classic problem in statistics.
 - The approximated error rate is bounded by [1]

$$\rightarrow \mathbb{P}(D_{KL}(g_{n,k} \| f) \geq \epsilon) \leq e^{-n\epsilon} \left[\frac{3c_1}{c_2} \sum_{i=0}^{k-2} K_{i-1} \left(\frac{e\sqrt{n}}{2\pi} \right)^i \right]$$

- n : number of samples
- k : number of category in the categorical distribution = 8
- ϵ : KL error rate = 0.2

How many answers are enough to approximate the true human answer distribution?

CFC Data Collection

How many answers are enough to approximate the true human answer distribution?

- Classic problem in statistics.

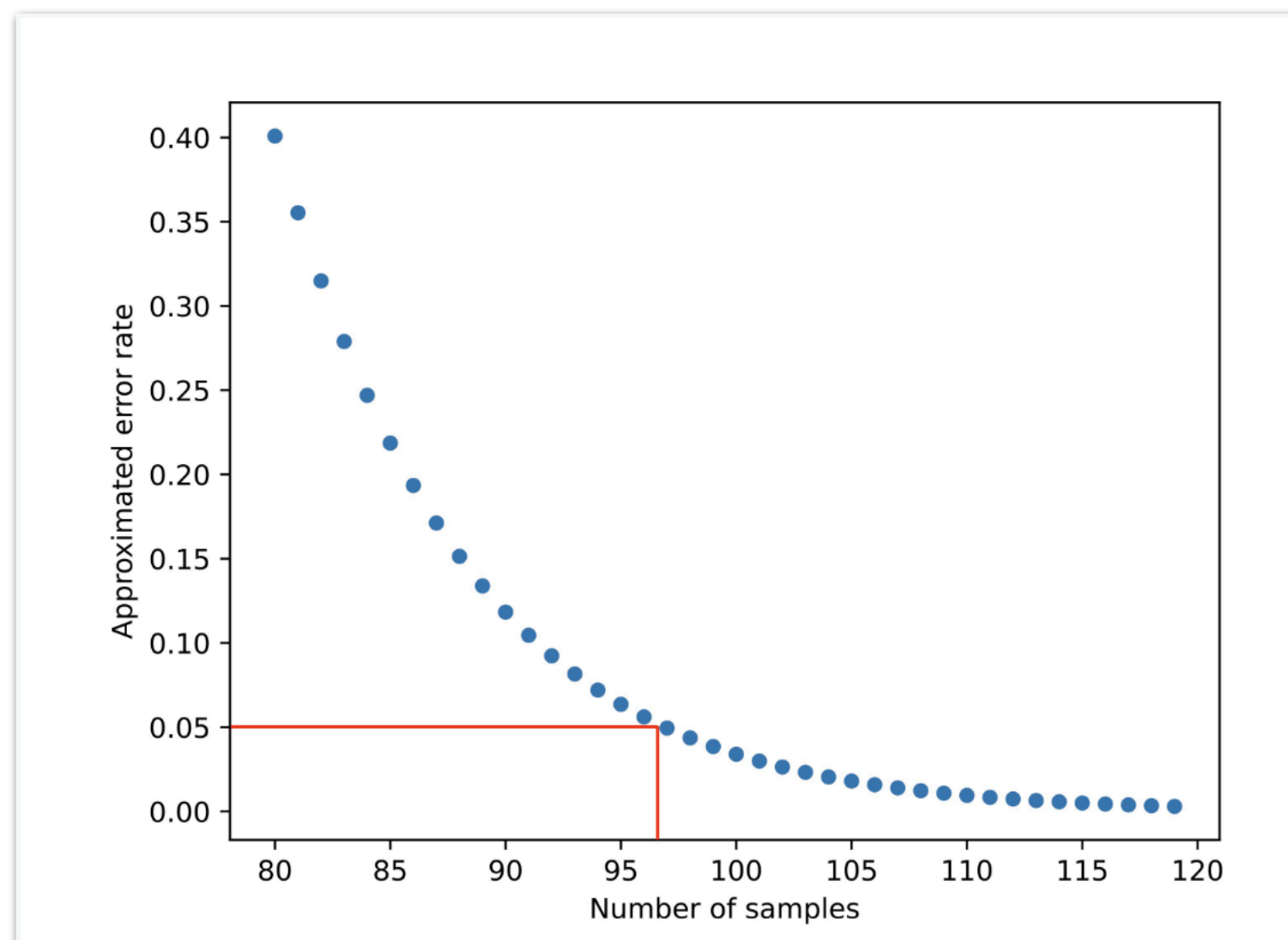
- The approximated error rate is bounded by [1]

$$\Rightarrow \mathbb{P}(D_{KL}(g_{n,k} || f) \geq \epsilon) \leq e^{-n\epsilon} \left[\frac{3c_1}{c_2} \sum_{i=0}^{k-2} K_{i-1} \left(\frac{e\sqrt{n}}{2\pi} \right)^i \right]$$

- n : number of samples

- k : number of category in the categorical distribution = 8

- ϵ : KL error rate = 0.2



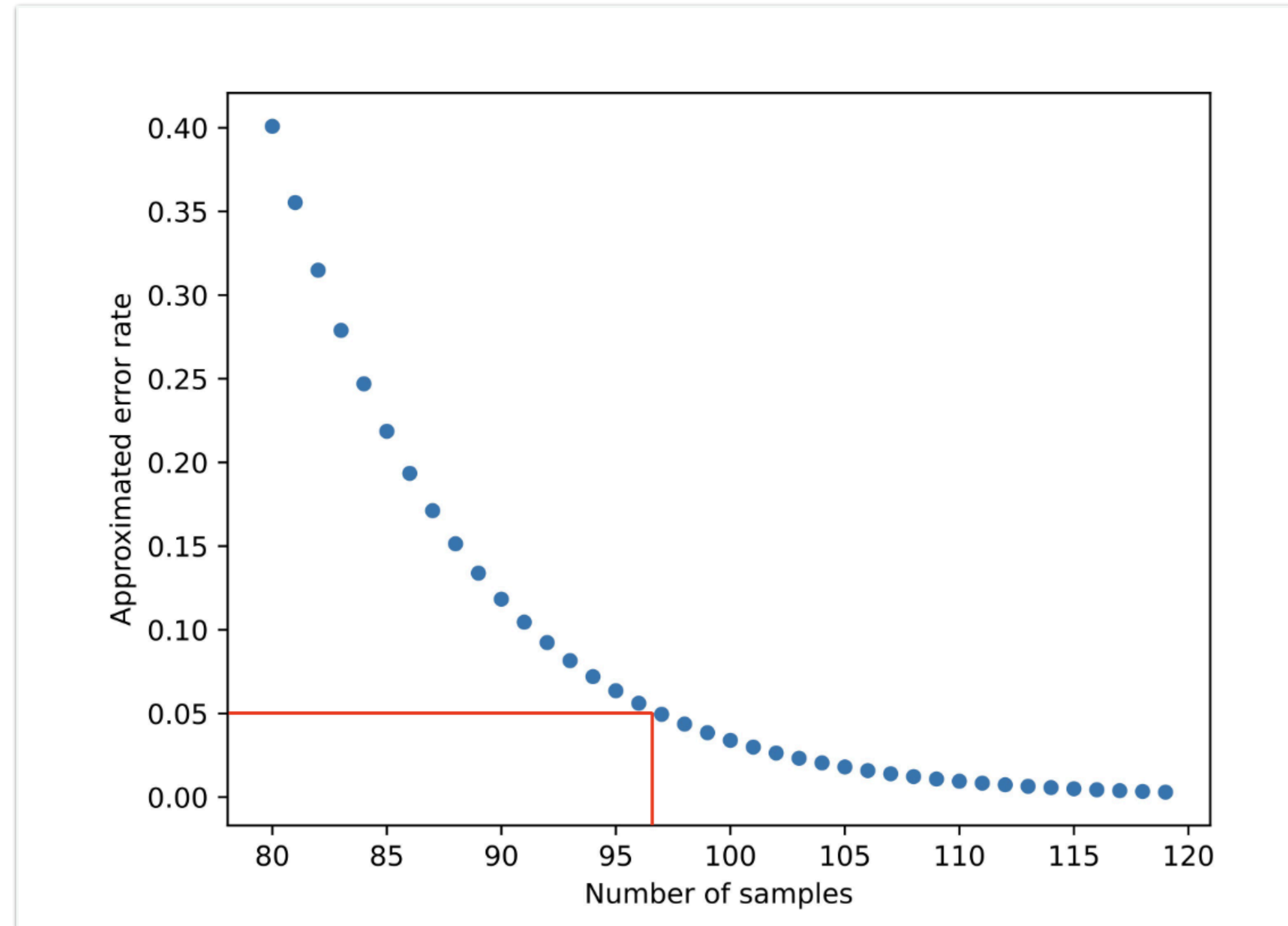
[1] Mardia, Jay, Jiantao Jiao, Ervin Tánčzos, Robert D. Nowak, and Tsachy Weissman. "Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types." *Information and Inference: A Journal of the IMA* 9, no. 4 (2020): 813-850.

Qi, Boratko, Yelugam, O’Gorman, Singh, McCallum, Li. "Every Answer Matters: Evaluating Commonsense with Probabilistic Measures" ACL 2024

CFC Data Collection

How many answers are enough to approximate the true human answer distribution?

~97. we collect 100 answers for each question.



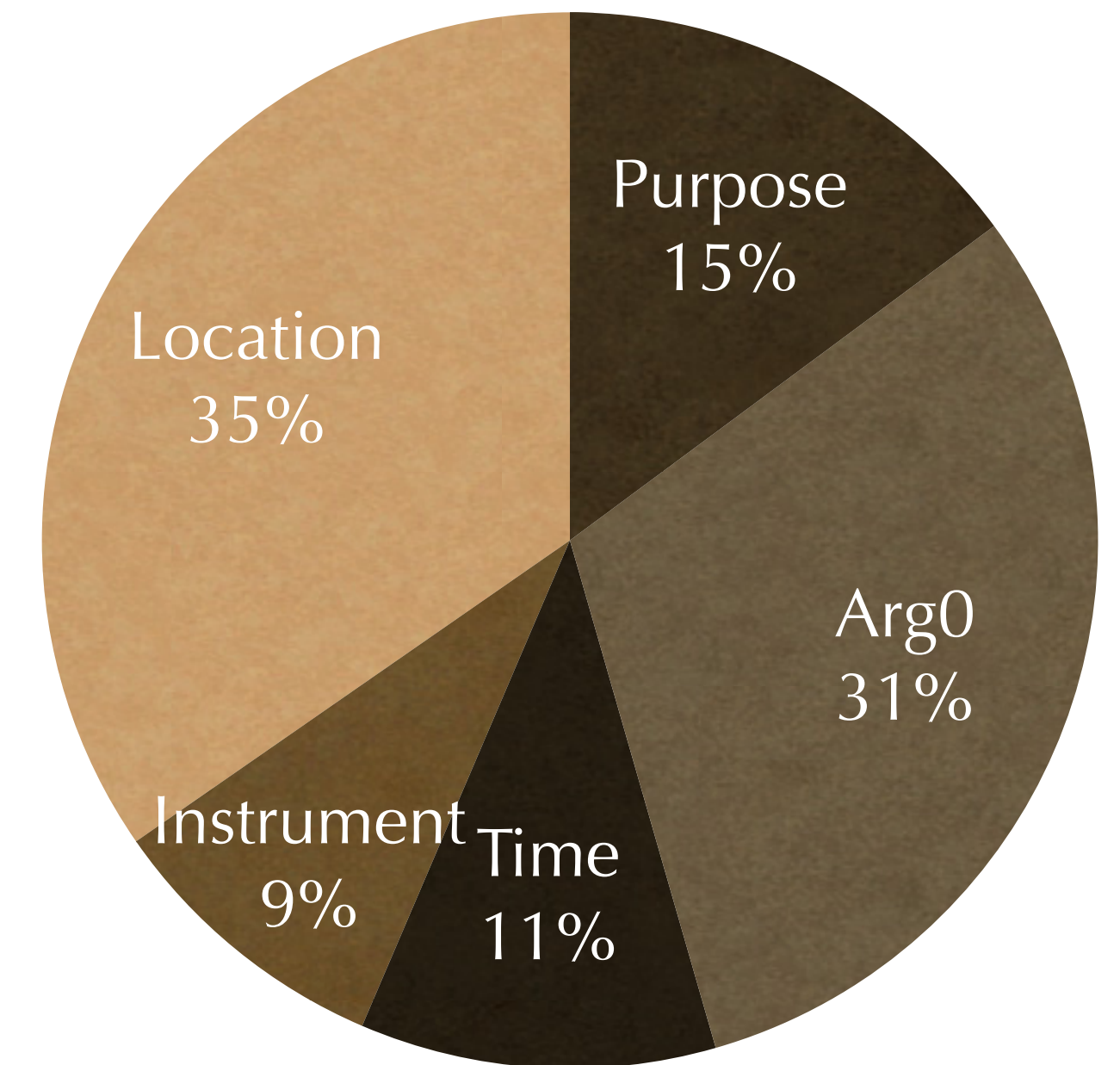
CFC Data Statistics

We crowd-sourced high-quality **101 questions (manual filtering)**

- 55 Dev Questions
- 46 Test Questions

Each question have 100 answers to **accurately** approximate human distribution.

- **Questions:** They boiled the water. Purpose?
- **Answers:**

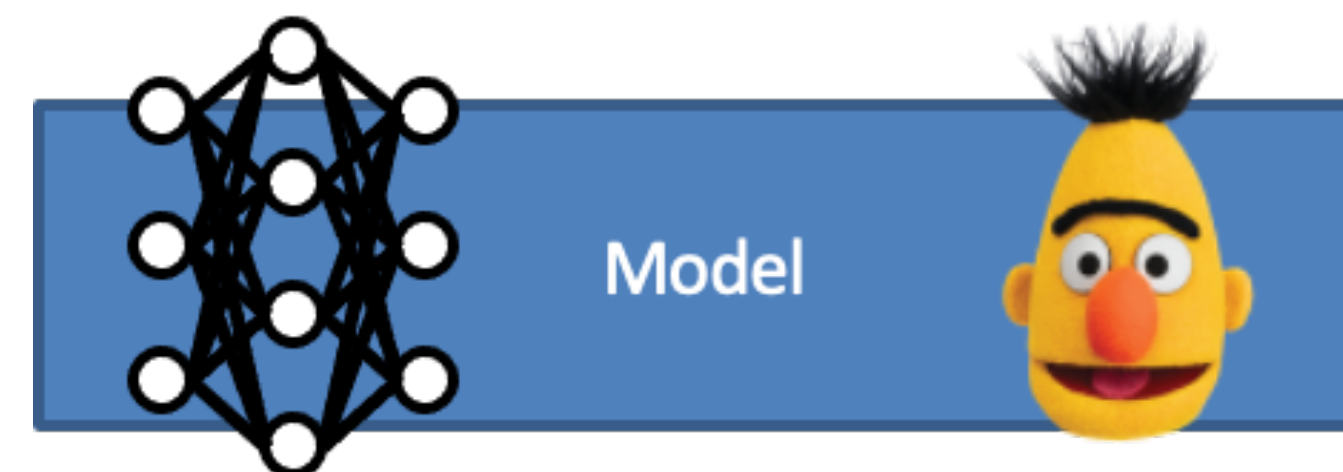


Question Slot Type

cook, cook noodles, cook pasta, bake cake, boil eggs, pasta, make pasta, cook meal, to make tea, coffee, make coffee, to make it safe to drink, to sterilize it, to remove germs and make it safe to drink ...

CFC Probabilistic Evaluation

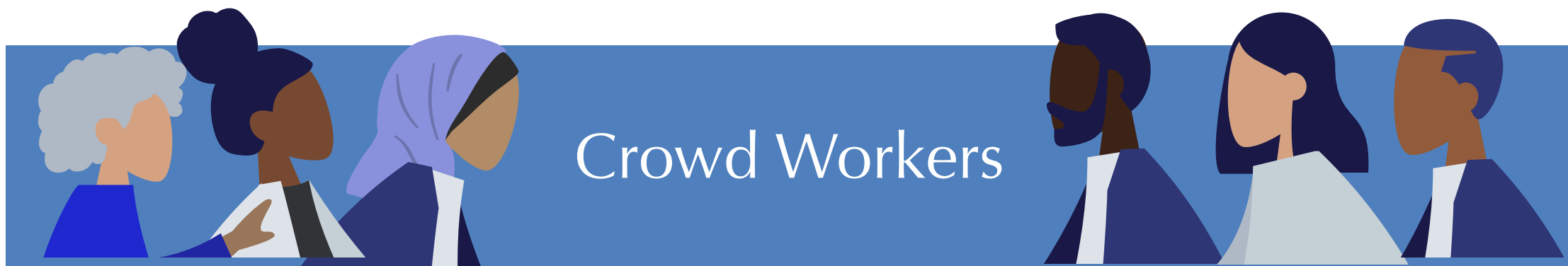
“They boiled the water” Purpose?



CFC Probabilistic Evaluation

“They boiled the water” Purpose?

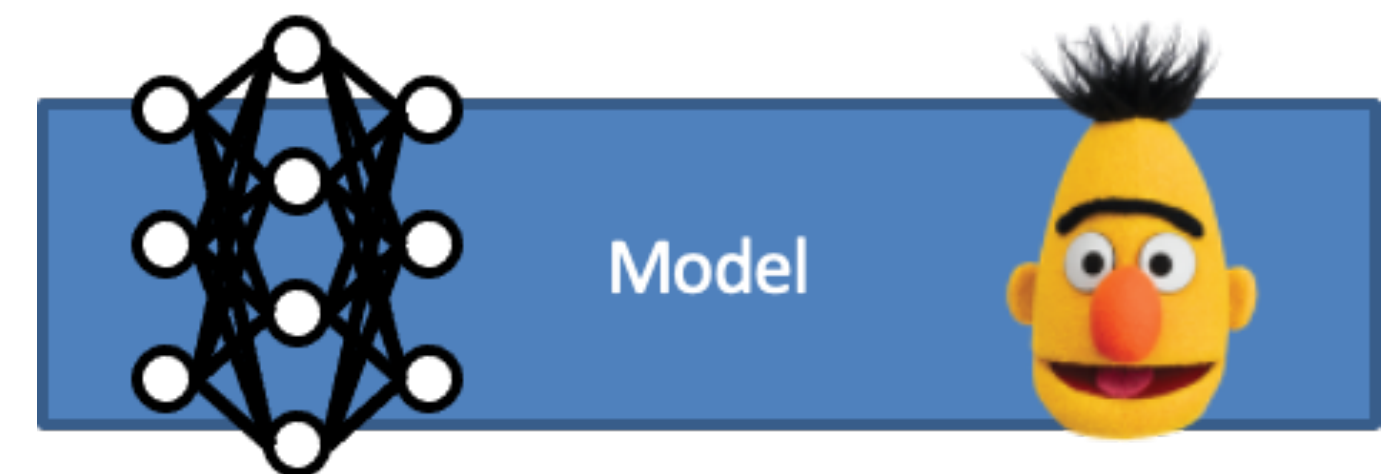
cooking clean
cook make tea disinfect disinfecting
for making tea making dinner cleaning
to cook cook food for a hot drink cleaning tools
cooking spaghetti making pasta kill bacteria
steaming vegetables purify purification
boiling potatoes make safe to drink
boiling chicken sterilization for an experiment



CFC Probabilistic Evaluation

“They boiled the water” Purpose?

cooking clean make a cup of tea
disinfect disinfecting
cook make tea making dinner
for making tea cleaning
to cook cook food for a hot drink cleaning tools
cooking spaghetti making pasta kill bacteria
steaming vegetables purify purification
boiling potatoes to make hard boiled eggs
boiling chicken make safe to drink
sterilization for an experiment



CFC Probabilistic Evaluation

“They boiled the water” Cause?

cooking clean make a cup of tea
disinfect disinfecting
cook make tea making dinner for tea making coffee
for making tea cleaning making coffee
to cook cook food for a hot drink cleaning
cooking spaghetti making pasta cleaning tools to sanitize
steaming vegetables purify kill bacteria cook dinner kill parasites
boiling potatoes purification to make hard boiled eggs
boiling chicken make safe to drink making food steriliza instruments
sterilization for an experiment



CFC Probabilistic Evaluation

“They boiled the water” Purpose?

for making tea
for a hot drink
make tea
clean disinfect
disinfecting
cleaning cleaning tools
make a cup of tea
for tea making coffee
cooking
cleaning
to cook steaming vegetables
kill bacteria
purify purification
cook dinner
to sanitize
boiling potatoes
make safe to drink
sterilization
kill parasites
cook boiling chicken
cook food making dinner
to make hard boiled eggs
cooking for an experiment
making food steriliza instruments



CFC Probabilistic Evaluation

“They boiled the water” Purpose?

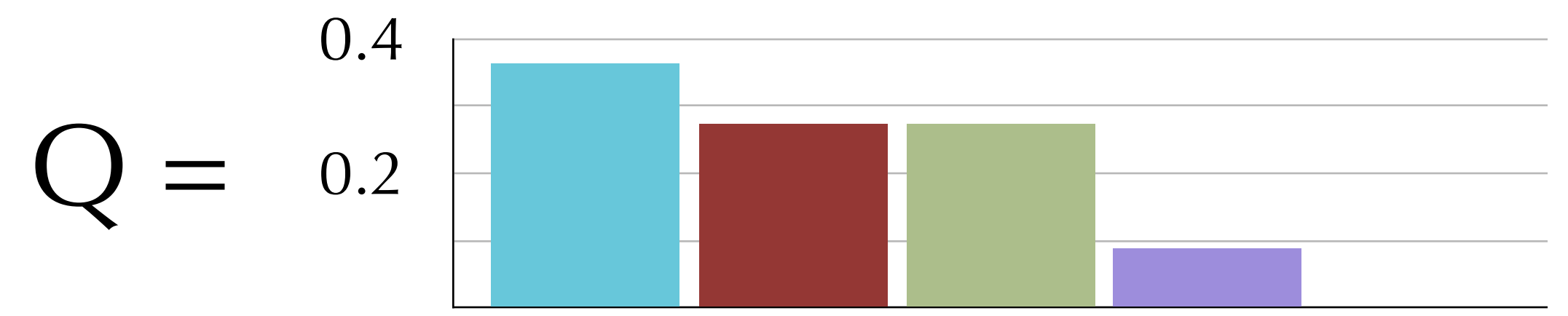
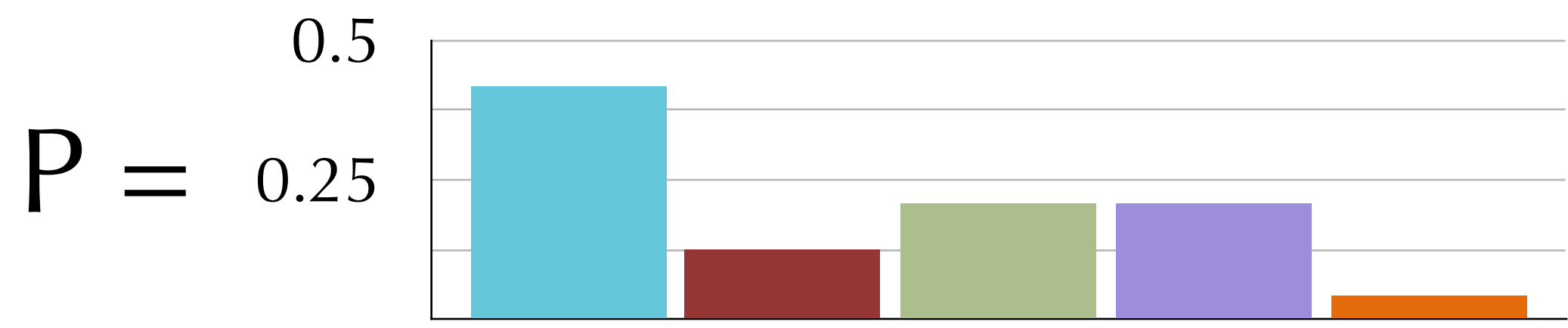
for making tea
for a hot drink
make tea
clean
disinfect
disinfecting
cleaning
cleaning tools
make a cup of tea
for tea
making coffee
cooking spaghetti
making pasta
to cook
steaming vegetables
boiling potatoes
kill bacteria
purify
purification
cooking
to make hard boiled eggs
cook
boiling chicken
make safe to drink
sterilization
making food
cook dinner
kill parasites
cook food
making dinner
cooking
for an experiment
steriliza instruments
cleaning
to sanitize



CFC Probabilistic Evaluation

“They boiled the water” Purpose?

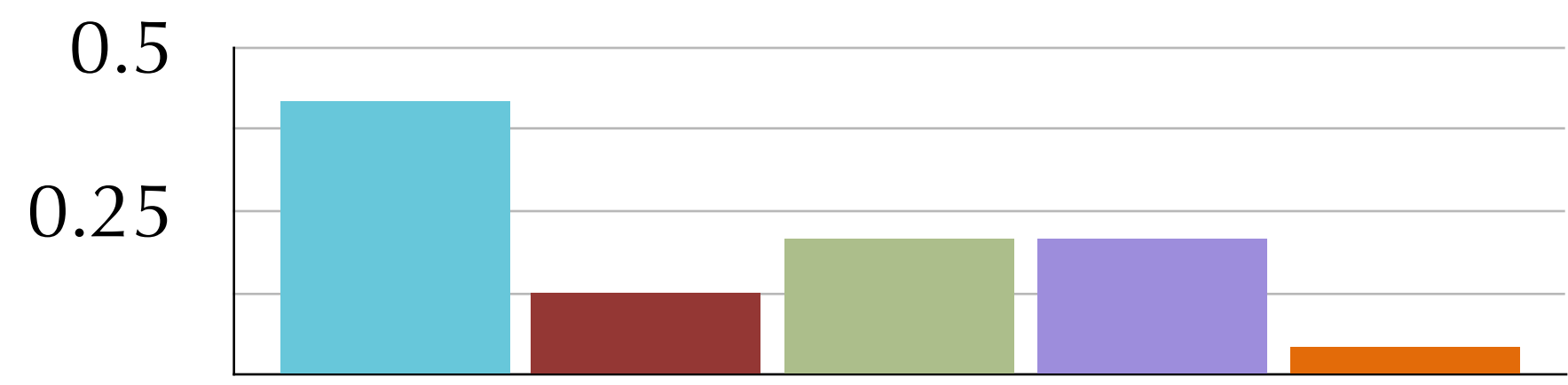
for making tea
for a hot drink
make tea
clean
disinfect
disinfecting
cleaning
cleaning tools
make a cup of tea
for tea
making coffee
cooking spaghetti
making pasta
to cook
steaming vegetables
kill bacteria
purify
purification
make safe to drink
sterilization
boiling potatoes
boiling chicken
cook
cook food
making dinner
cooking
for an experiment
cleaning
to sanitize
steriliza instruments
cooking
to make hard boiled eggs
making food
cook dinner
kill parasites



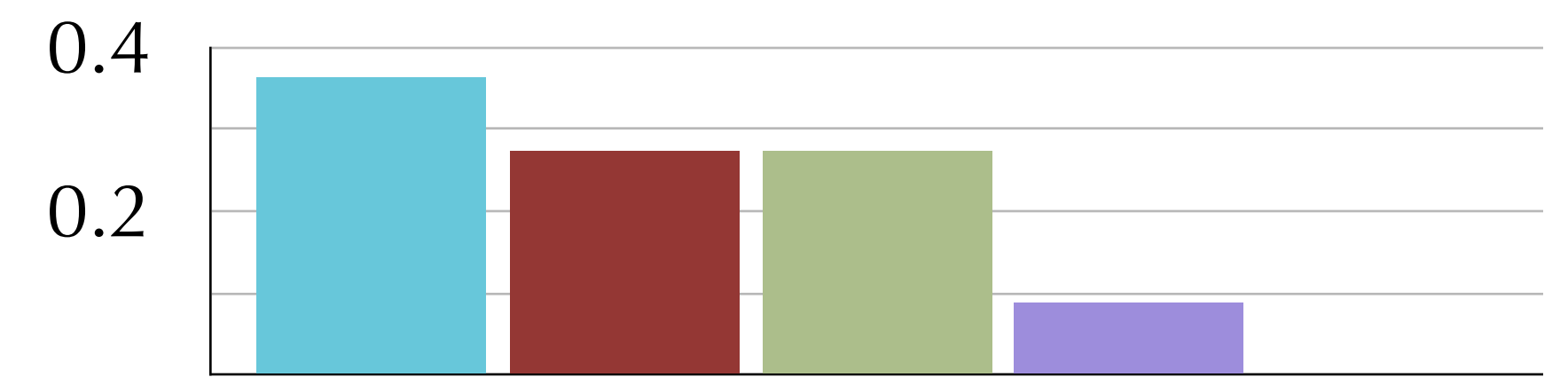
CFC Probabilistic Evaluation

“They boiled the water” Purpose?

for making tea
for a hot drink
make tea
clean
disinfect
disinfecting
cleaning
cleaning tools
make a cup of tea
for tea
making coffee
cooking spaghetti
making pasta
to cook
steaming vegetables
kill bacteria
purify
purification
boiling potatoes
make safe to drink
sterilization
cook
boiling chicken
steriliza instruments
cook food
making dinner
cooking
for an experiment
cooking
to make hard boiled eggs
making food
cook dinner
kill parasites



KL (P || Q)



CFC Automatic Evaluation

For each question:

$G \leftarrow$ *ground-truth answers (crowd-sourced)*

$H \leftarrow$ *evaluation answers (model)*

For each human scorer:

Cluster G

Match H to clusters of G

Calculate score

$\text{Score}(G, H) \leftarrow$ average of scores

CFC Automatic Evaluation

For each question:

$G \leftarrow$ *ground-truth answers (crowd-sourced)*

$H \leftarrow$ *evaluation answers (model)*

For each human scorer:

Cluster G

Match H to clusters of G

Calculate score

$\text{Score}(G, H) \leftarrow$ average of scores

CFC Automatic Evaluation

Embed G

Cluster G

Match H to cluster of G

Calculate Score

CFC Automatic Evaluation

Embed G

Cluster G

Match H to cluster of G

Calculate Score

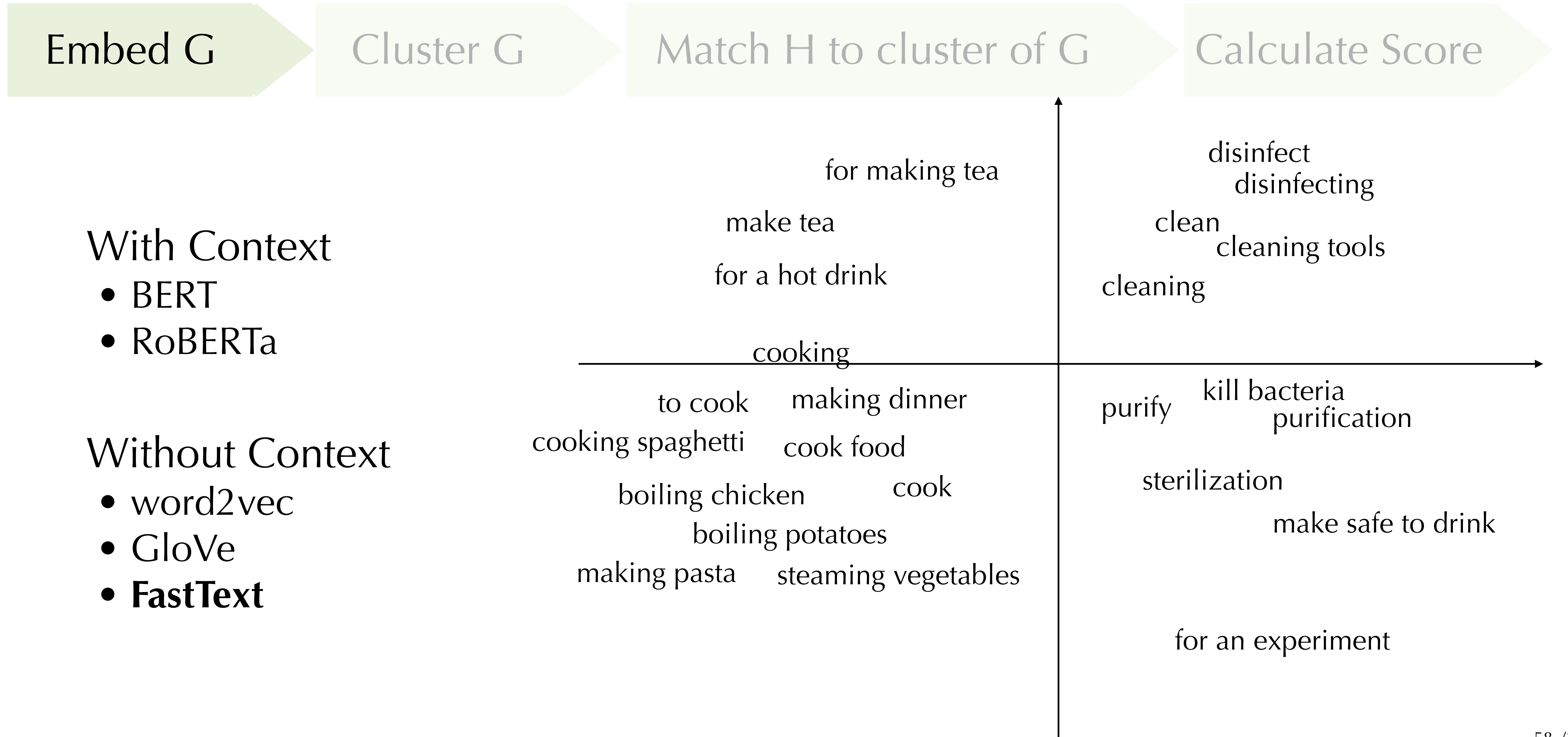
Ground truth: G

cooking
clean
disinfect
make tea
disinfecting
cook
making dinner
cleaning
cook food
cleaning tools
to cook
purification
cooking spaghetti
kill bacteria
steaming vegetables
for a hot drink
boiling potatoes
boiling chicken
purify
sterilization
make safe to drink
for an experiment
for making tea
making pasta

make a cup of tea
making coffee
for tea
cleaning
cooking
to sanitize
cook dinner
kill parasites
to make hard boiled eggs
making food
steriliza instruments

Model prediction: H

CFC Automatic Evaluation



CFC Automatic Evaluation

Embed G

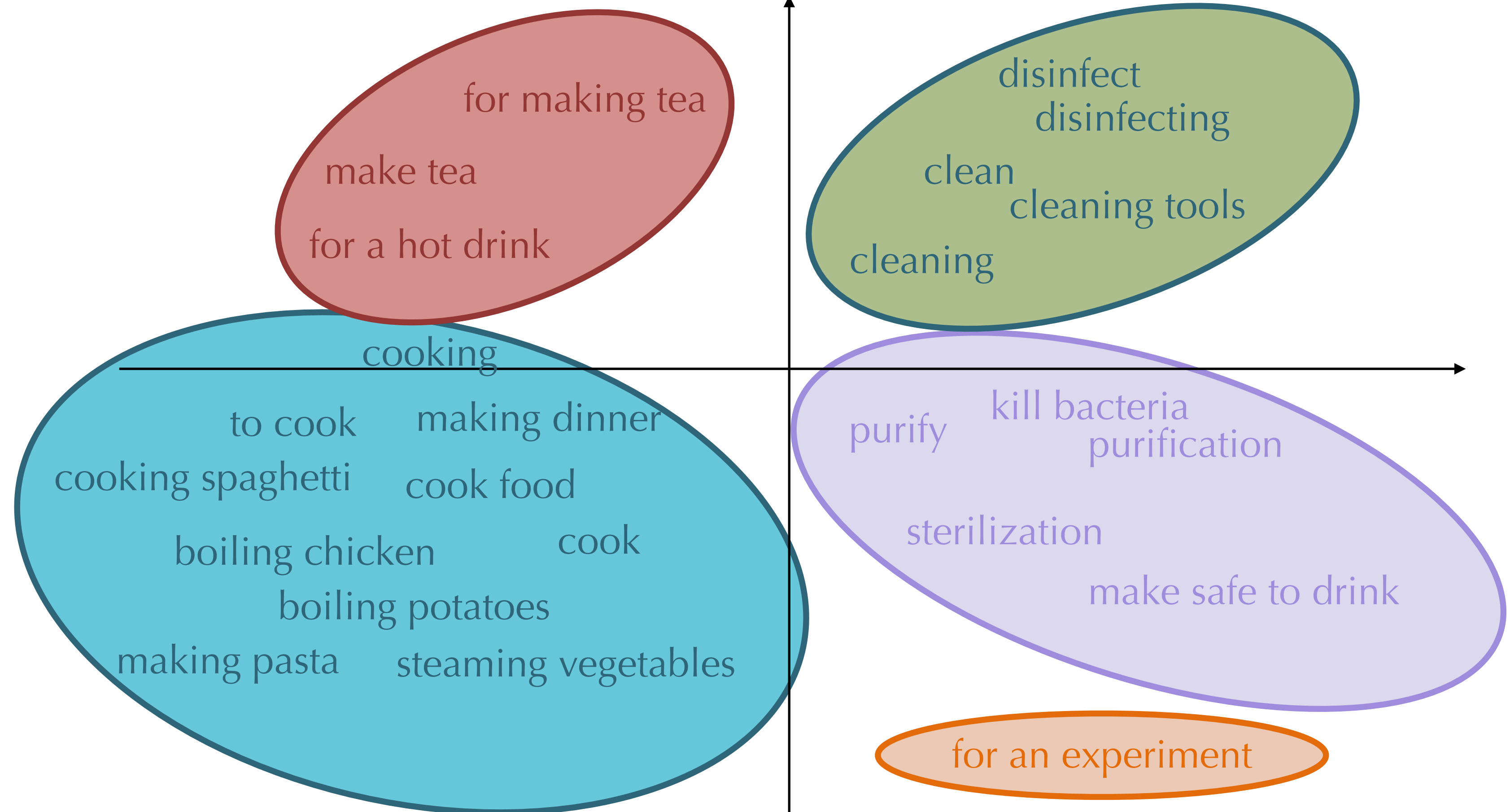
Cluster G

Match H to cluster of G

Calculate Score

Clustering Algorithm

- K-Means
- G-Means [1]
- **Hierarchical agglomerative clustering**



[1] Zhao, Zhonghua et al. "G-Means: A Clustering Algorithm for Intrusion Detection." *ICONIP* (2008).

Qi, Boratko, Yelugam, O'Gorman, Singh, McCallum, Li. "Every Answer Matters: Evaluating Commonsense with Probabilistic Measures" *ACL* 2024

CFC Automatic Evaluation

Embed G

Cluster G

Match H to cluster of G

Calculate Score

make tea for a hot drink
for making tea

clean cleaning disinfect
disinfecting cleaning tools

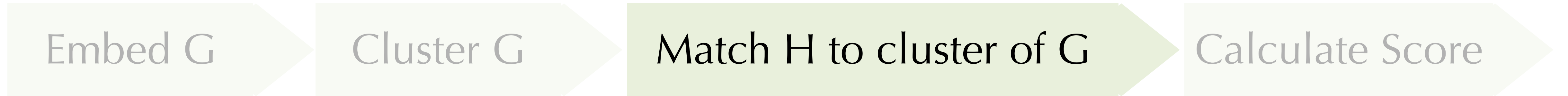
purify kill bacteria
make safe to drink
purification sterilization

for an experiment

cooking making dinner
to cook cook food
cook boiling chicken
boiling potatoes
steaming vegetables
making pasta
cooking spaghetti

**make a cup of tea
making coffee
for tea
cleaning
cooking
to sanitize
cook dinner
kill parasites
to make hard boiled eggs
making food
steriliza instruments**

CFC Automatic Evaluation



Embeddings Based

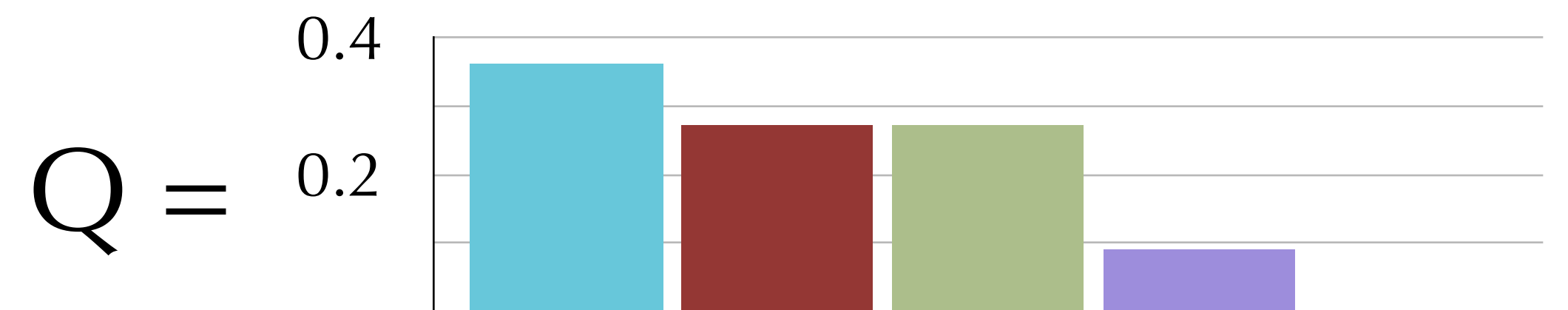
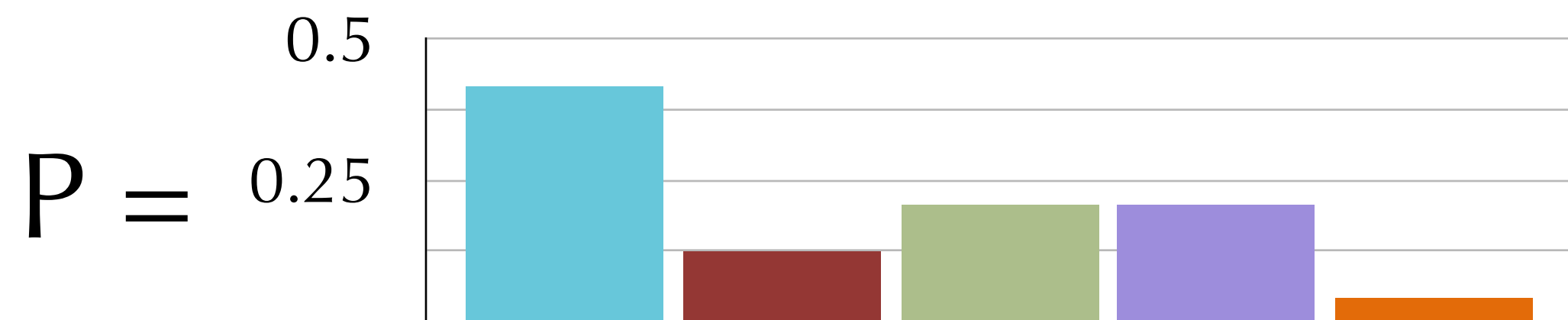
- FastText

Lexical Token Based

- WordNet



CFC Automatic Evaluation



$$\text{Score}(G, H) = \text{KL}(P \parallel Q)$$

Evaluating Automatic Metric

Given a question, and a large prediction set

- Sample **n** predicted answer sets.

s1, s2, s3, s4, s5...

- Using **human** annotations, score answer sets:

H: [s2, s5, s4, s3, s1...]

- Using **automatic** evaluation, score answer sets:

A: [s2, s4, s3, s1, s5]

- Calculate Spearman correlation between **H** and **A**

Evaluating Automatic Metric

Clustering	Gmeans		Xmeans		Hierarchical agglomerative clustering (HAC)	
Matching	FastText	WordNet	FastText	WordNet	FastText	WordNet
ProtoQA Correlation	0.528	0.681	0.525	0.668	0.593	0.698
CFC Correlation	0.561	0.721	0.503	0.728	0.564	0.728

Table: Spearman correlation between human KL score and automatic KL score

Evaluating Automatic Metric

Clustering	Gmeans		Xmeans		Hierarchical agglomerative clustering (HAC)	
Matching	FastText	WordNet	FastText	WordNet	FastText	WordNet
ProtoQA Correlation	0.528	0.681	0.525	0.668	0.593	0.698
CFC Correlation	0.561	0.721	0.503	0.728	0.564	0.728

Table: Spearman correlation between human KL score and automatic KL score

Evaluating Automatic Metric

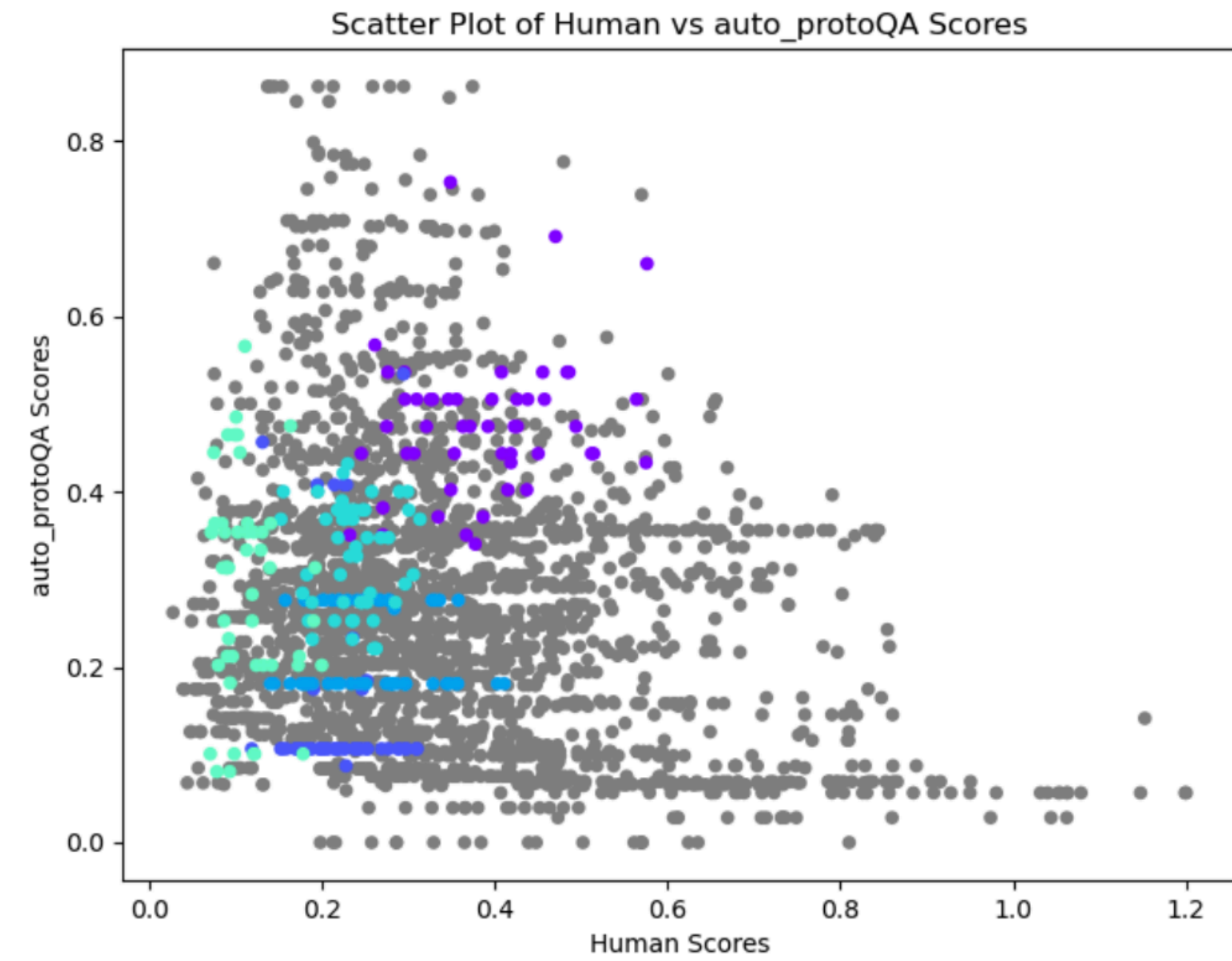
Clustering	Gmeans		Xmeans		Hierarchical agglomerative clustering (HAC)	
Matching	FastText	WordNet	FastText	WordNet	FastText	WordNet
ProtoQA Correlation	0.528	0.681	0.525	0.668	0.593	0.698
CFC Correlation	0.561	0.721	0.503	0.728	0.564	0.728

Evaluating Automatic Metric - PROBEVAL

X-axis: KL with human cluster and matching
Y-axis: automatic evaluator score (kl or 1-protoqa score)
Five random questions are annotated with different colors

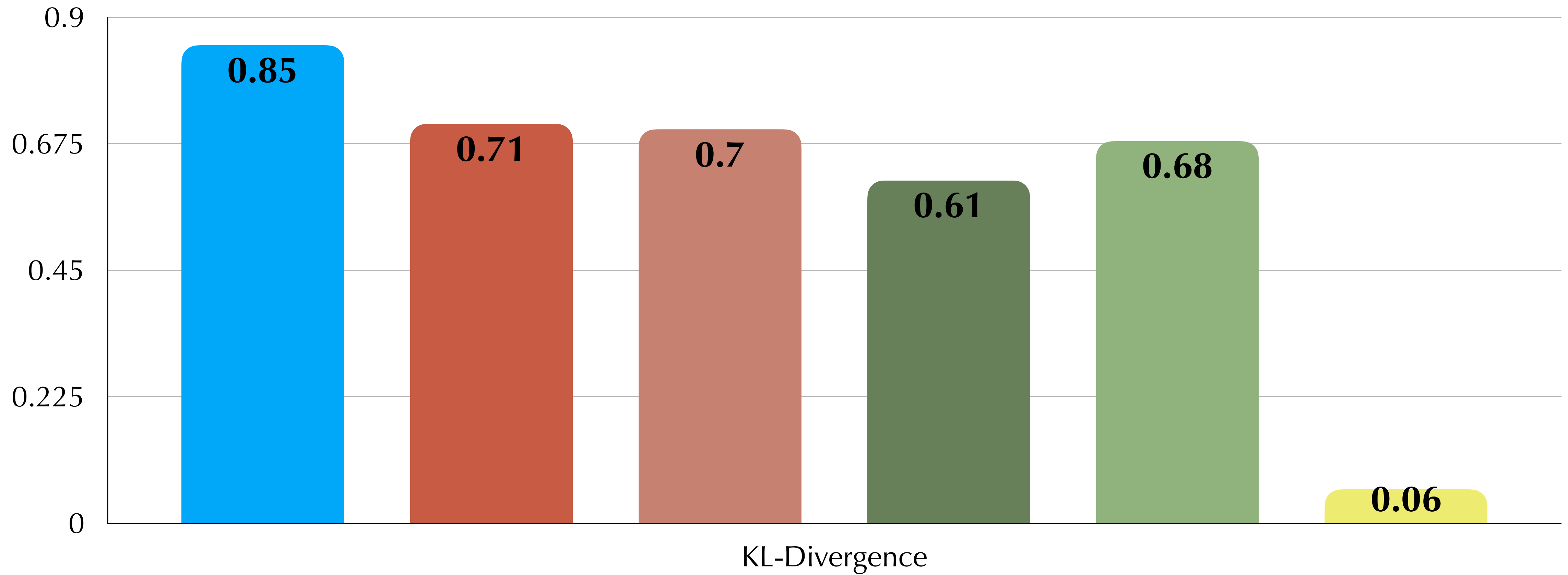


Ours



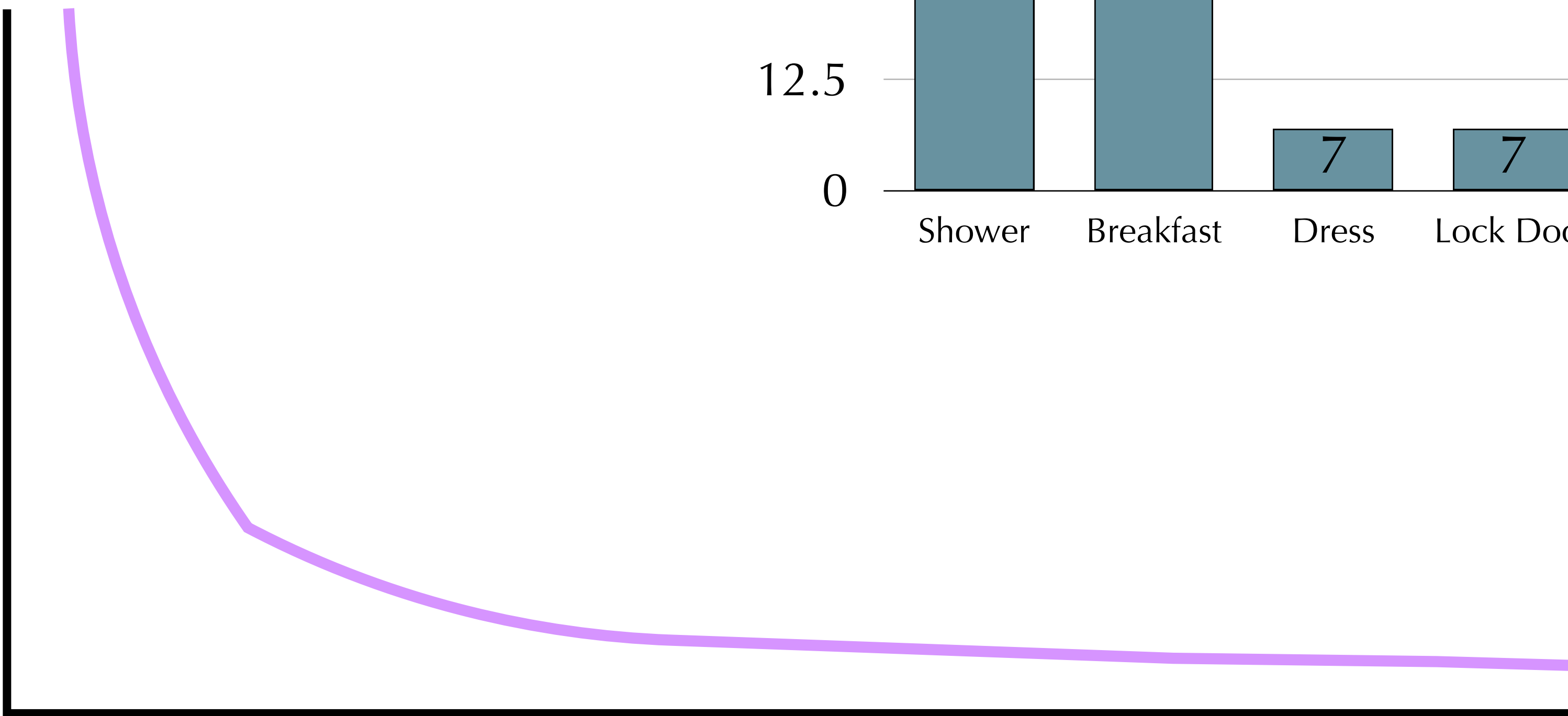
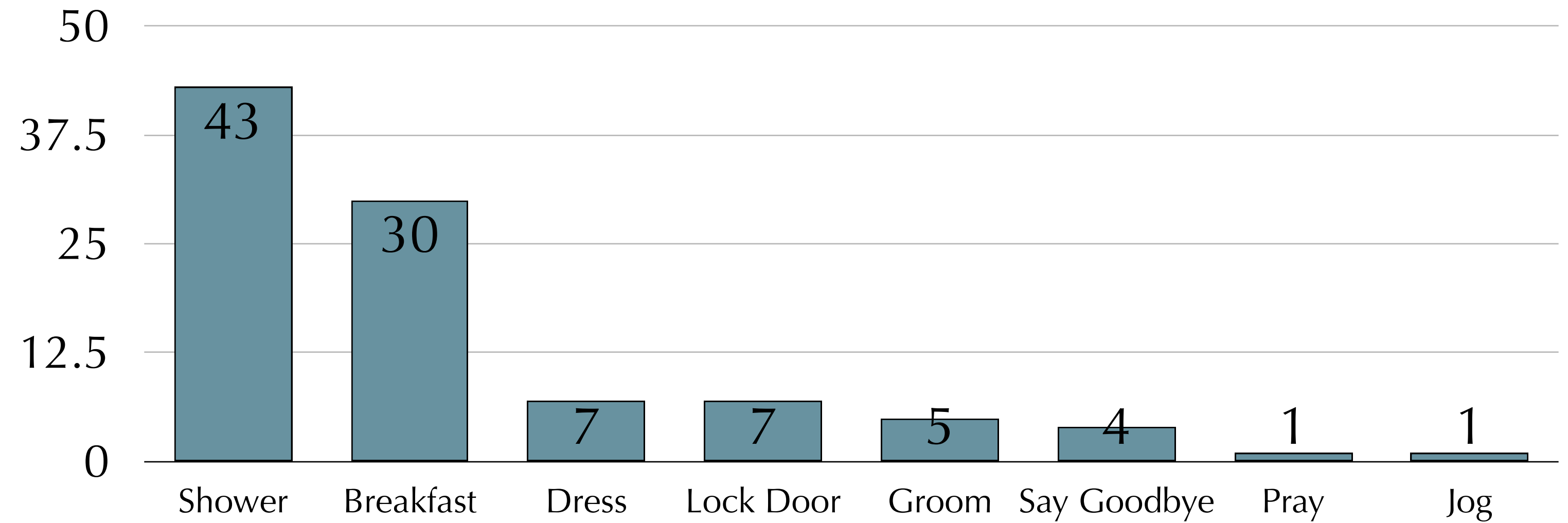
ProtoQA Evaluator

Model Performance

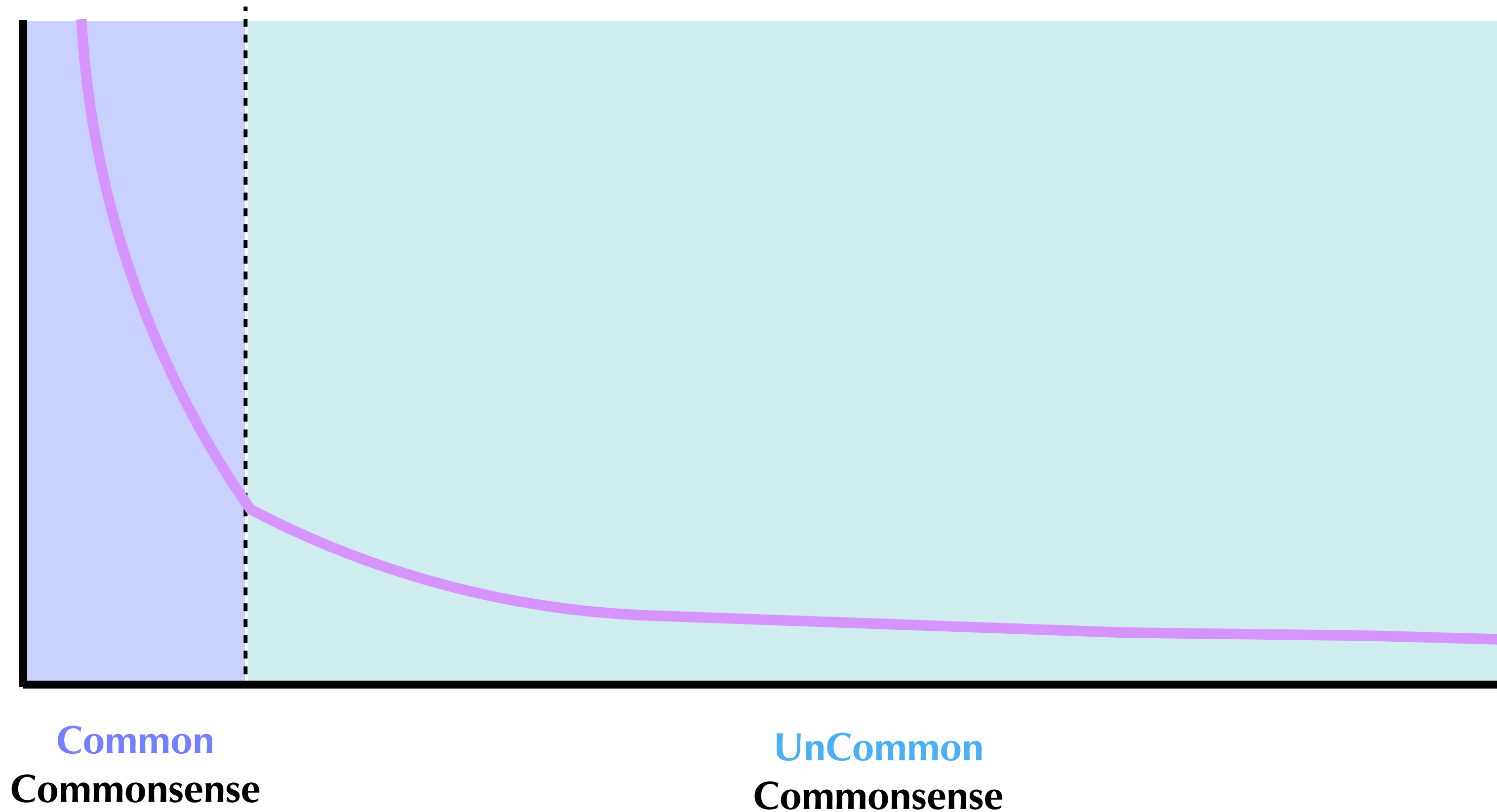


Why is performance bad?

- The long-tail problem.



Why is performance so bad?



Probabilistic View of Commonsense Questions

They boiled the water and added spaghetti.



Why?

Probabilistic View of Commonsense Questions

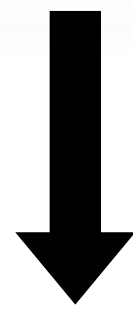
They boiled the water and added spaghetti. They invited their friend Kate to try the spaghetti. Kate didn't like the spaghetti but kept eating.



Why?

Probabilistic View of Commonsense Questions

They boiled the water and added spaghetti. They invited their friend Kate to try the spaghetti. Kate didn't like the spaghetti but kept eating.



Why?

UnCommon
Commonsense

Reasoning

Context: Cameron tried sushi for the first time, and really disliked it.



Despite disliking the taste of sushi, Cameron decided to stay and eat more sushi plates to avoid disappointing his partner, who was excited about sharing...

UnCommon

Commonsense

Uncommon Outcome: Cameron will want to stay and eat more sushi.

UNcommonsense Abductive Reasoning

Context: Cameron tried sushi for the first time, and really disliked it.

Explanations:

- ✓ Makes outcome more likely.
- ✓ Naturally follows the context.
- ✓ Leaves little information gap in-between.

Despite disliking the taste of sushi, Cameron decided to stay and eat more sushi plates to avoid disappointing his partner, who was excited about sharing...

Uncommon Outcome: Cameron will want to stay and eat more sushi.

UNcommonsense Abductive Reasoning

- Uncommon Outcomes
 - “Incorrect” answers from **SocialQA & RocStories**
 - human written



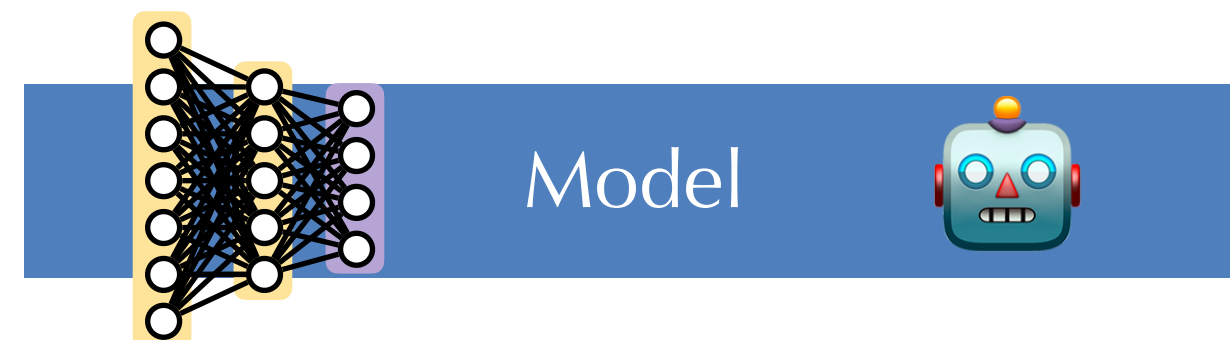
Uncommon Outcome: Cameron will want to stay and eat more sushi.

UNcommonsense Abductive Reasoning

Despite disliking the taste of sushi, Cameron decided to stay and eat more sushi plates to avoid disappointing his partner, who was excited about sharing...

- Explanations for uncommon outcomes

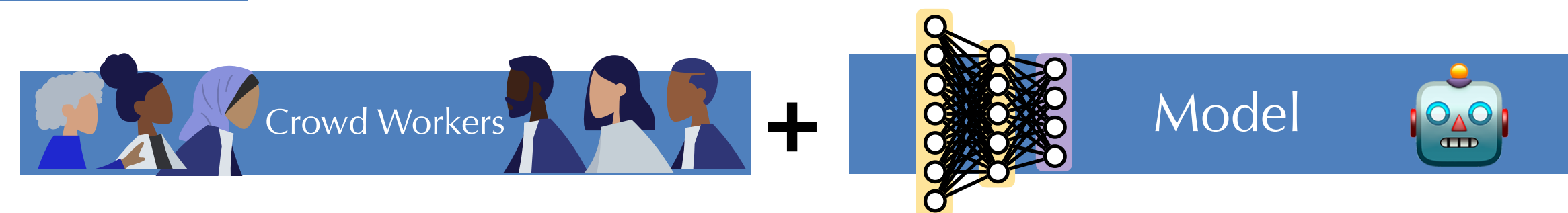
- LLM generated



- human written



- human written + LLM modification



UNcommonsense Abductive Reasoning

- Uncommon Outcomes
 - “Incorrect” answers from SocialQA & RocStories
 - human written
- Explanations for Uncommon outcomes
 - LLM generated
 - human written
 - human written + LLM modification

UNcommonsense Abductive Reasoning

Explanation Analysis: Quality

	un-SocialQA			un-RocStories		
	<i>Crowd</i>	<i>C+LLM</i>	<i>LLM²</i>	<i>Crowd</i>	<i>C+LLM</i>	<i>LLM²</i>
Win	30.8	43.2	33.8	19.2	28.4	26.4
Eql. good	33.4	34.8	41.2	37.0	45.6	42.4
Eql. bad	3.4	2.0	3.8	12.0	3.0	3.0
Lose	32.4	20.0	21.2	42.6	23.0	28.2
Non-Lose:	67.6	80	78.8	57.4	77	71.8

Figure 1: Win rates judged by Crowdworkers of Human+LLM versus LLM.

- LLM explanations are preferred over Crowd explanations

UNcommonsense Abductive Reasoning

Explanation Analysis: Length & Entropy

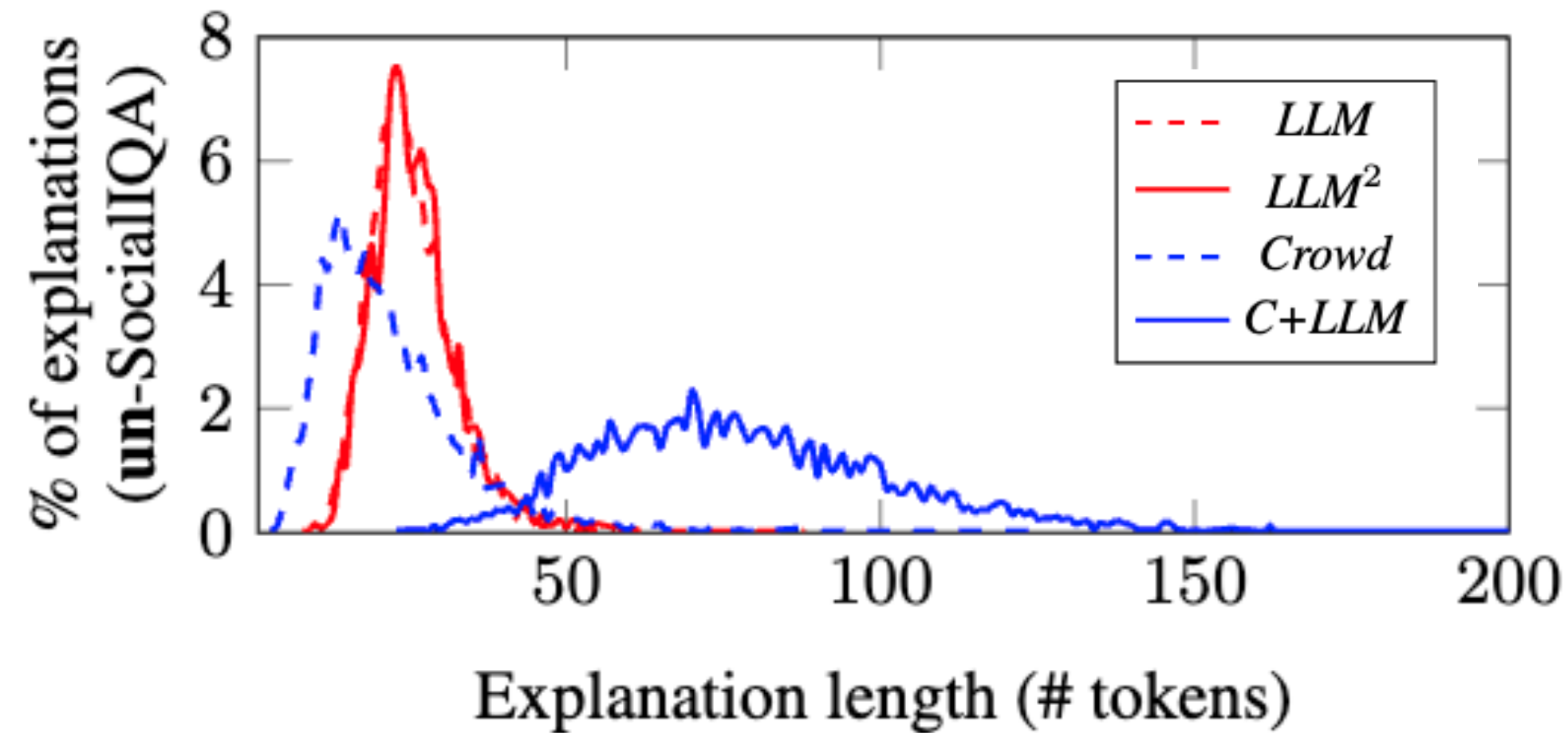


Figure 2: Distribution of explanation lengths in un-SocialQA. Computed on the development sets.

- Crowd explanations are significantly shorter than LLM.
- Enhancing crowd-written explanations with an LLM significantly increases their lengths over LLM.

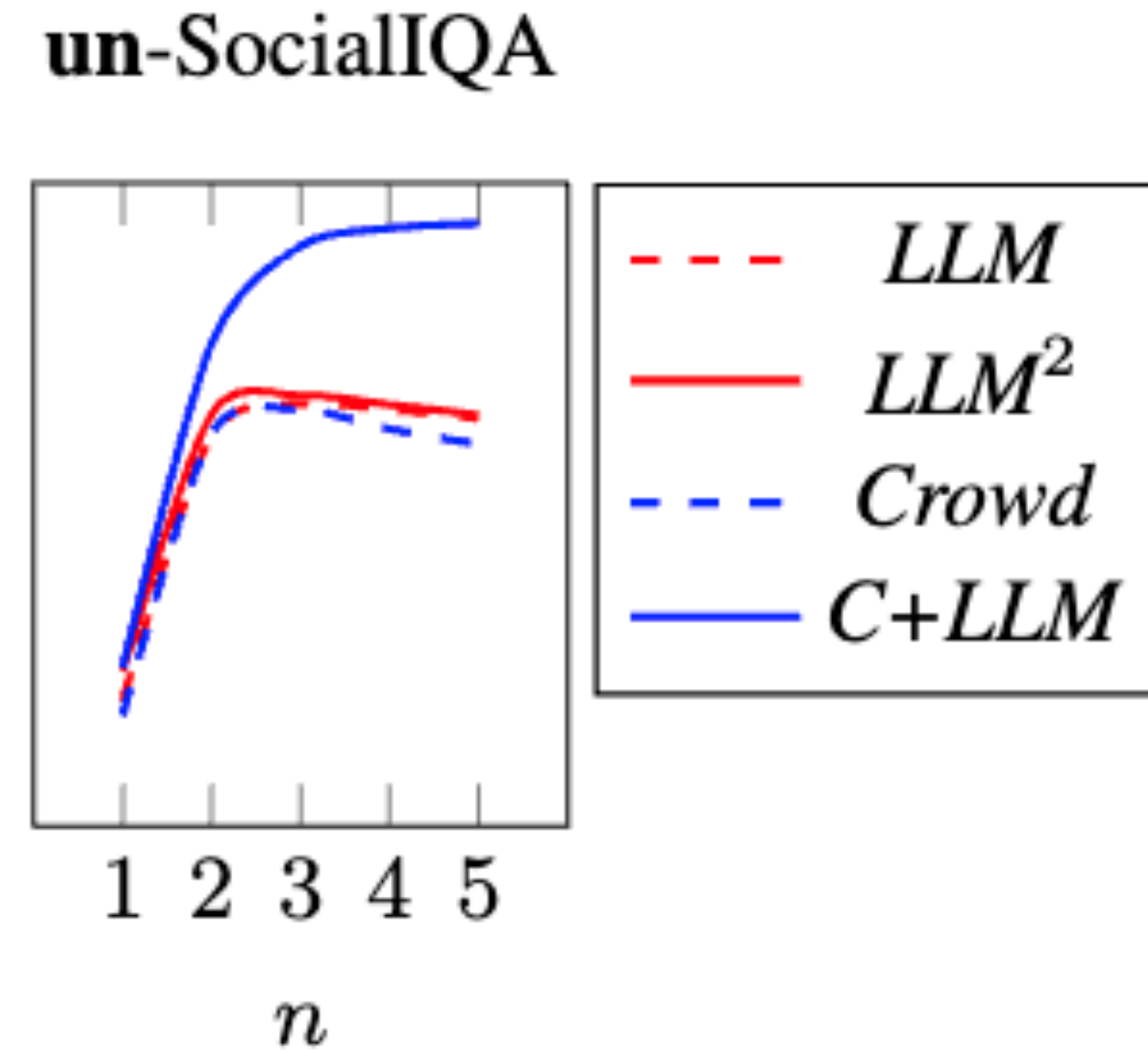


Figure 3: Entropies of n-gram distributions in un-SocialQA. Computed on the development sets.

- Entropy as a measure for lexical diversity.
- Crowd has generally lower entropy than LLM.
- LLM enhancement of crowd-written explanations results in significantly higher entropy.

UNcommonsense Abductive Reasoning

Explanation Analysis: Outcome Likelihood

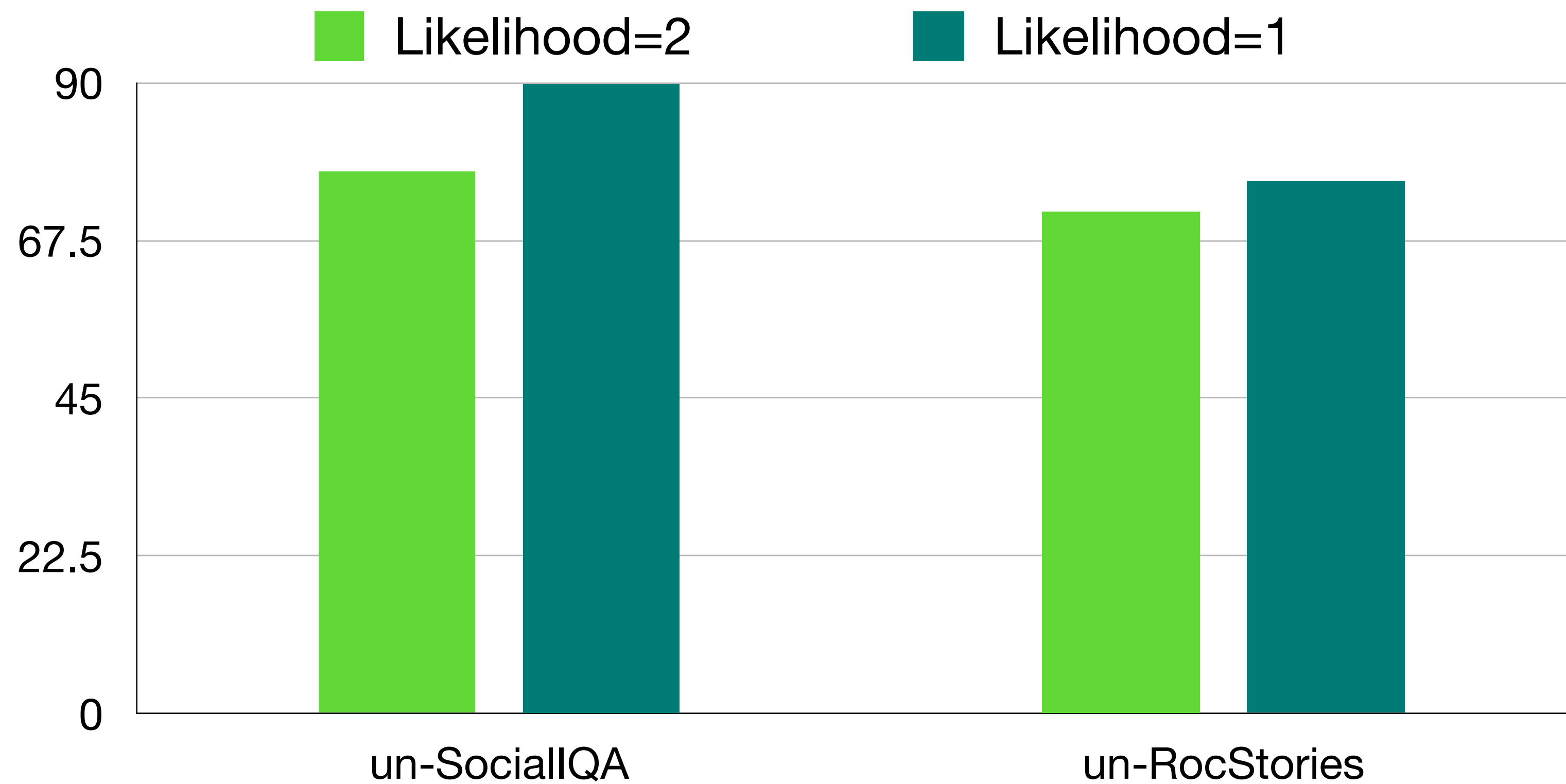


Figure 1: Non-lose rates of Human+LLM versus LLM, broken down by the likelihoods of outcomes. Likelihood=1 is less likely. (annotated by human)

Human+LLM explanations become more preferable as the likelihood of outcomes decreases.

UNcommonsense Abductive Reasoning

Takeaways

- GPT4 is not bad for explaining uncommon situations. So, are we done?
 - We argue that the uncommon situation in the uncommon sense dataset can still be explained with common arguments, i.e., not that “uncommon” such that it requires complicated reasoning.
 - Can we evaluate data directly using complicated reasoning?
 - One type of complicated reasoning can be compositional reasoning

Compositional Reasoning Evaluation

- a case study in puzzle game

- We aim to better understand what is possible and not possible with Transformers with these highly compositional tasks that require **multi-step reasoning**.

Reasoning Task: Einstein's Puzzle

General Unique Rules

There are 3 houses (numbered 1 on the left, 3 on the right). They have different characteristics:

- Each person has a unique **name**: Peter, Eric, Arnold
- People have different favorite **sports**: Soccer, Tennis, Basketball
- People own different **car models**: Tesla, Ford, Camry



House	1	2	3
Name			
Sports			
Car			

Reasoning Task: Einstein's Puzzle

General Unique Rules

There are 3 houses (numbered 1 on the left, 3 on the right). They have different characteristics:

- Each person has a unique name: Peter, Eric, Arnold
- People have different favorite sports: Soccer, Tennis, Basketball
- People own different car models: Tesla, Ford, Camry

Clues

1. The person who owns a Ford is the person who loves tennis.
2. Arnold is in the third house.
3. The person who owns a Camry is directly left of the person who owns a Ford.
4. Eric is the person who owns a Camry.
5. The person who loves basketball is Eric.
6. The person who loves tennis and the person who loves soccer are next to each other.



House	1	2	3
Name			
Sports			
Car			

Reasoning Task: Einstein's Puzzle

General Unique Rules

There are 3 houses (numbered 1 on the left, 3 on the right). They have different characteristics:

- Each person has a unique name: Peter, Eric, Arnold
- People have different favorite sports: Soccer, Tennis, Basketball
- People own different car models: Tesla, Ford, Camry



Clues

1. The person who owns a Ford is the person who loves tennis.
2. Arnold is in the third house.
3. The person who owns a Camry is directly left of the person who owns a Ford.
4. Eric is the person who owns a Camry.
5. The person who loves basketball is Eric.
6. The person who loves tennis and the person who loves soccer are next to each other.

House	1	2	3
Name			Arnold
Sports			
Car			

Reasoning Task: Einstein's Puzzle

General Unique Rules

There are 3 houses (numbered 1 on the left, 3 on the right). They have different characteristics:

- Each person has a unique name: Peter, Eric, Arnold
- People have different favorite sports: Soccer, Tennis, Basketball
- People own different car models: Tesla, Ford, Camry



Clues

1. The person who owns a Ford is the person who loves tennis.
2. Arnold is in the third house.
3. The person who owns a Camry is directly left of the person who owns a Ford.
4. Eric is the person who owns a Camry.
5. The person who loves basketball is Eric.
6. The person who loves tennis and the person who loves soccer are next to each other.

House	1	2	3
Name	Eric	Peter	Arnold
Sports	Basketball		
Car			

Reasoning Task: Einstein's Puzzle

General Unique Rules

There are 3 houses (numbered 1 on the left, 3 on the right). They have different characteristics:

- Each person has a unique **name**: Peter, Eric, Arnold
- People have different favorite **sports**: Soccer, Tennis, Basketball
- People own different **car models**: Tesla, Ford, Camry



Clues

1. The person who owns a Ford is the person who loves tennis.
2. Arnold is in the third house.
3. The person who owns a Camry is directly left of the person who owns a Ford.
4. Eric is the person who owns a Camry.
5. The person who loves basketball is Eric.
6. The person who loves tennis and the person who loves soccer are next to each other.

House	1	2	3
Name	Eric	Peter	Arnold
Sports	Basketball	Tennis	Soccer
Car	Camry	Ford	Tesla

Zero-shot Performance

GPT4 zero-shot (Puzzle)

2	1	0.9	0.72	0.6	0.6
3	0.7	0.6	0.2	0.1	0.3
4	0.8	0.2	0.1	0	0
5	0.1	0.1	0	0	0
6	0	0	0	0	0
	2	3	4	5	6

Figure 1: Zero-shot accuracy. Axes refer to problem sizes, number of houses and attributes in puzzle.

Transformers' accuracy decreases to near zero as task complexity increases, measuring task complexity by the problem size.

Does it mean models can't solve the tasks?

We fine-tuned the model

- Finetuned **GPT3 (large model)** with a **large amount of data** within a reasonable budget.

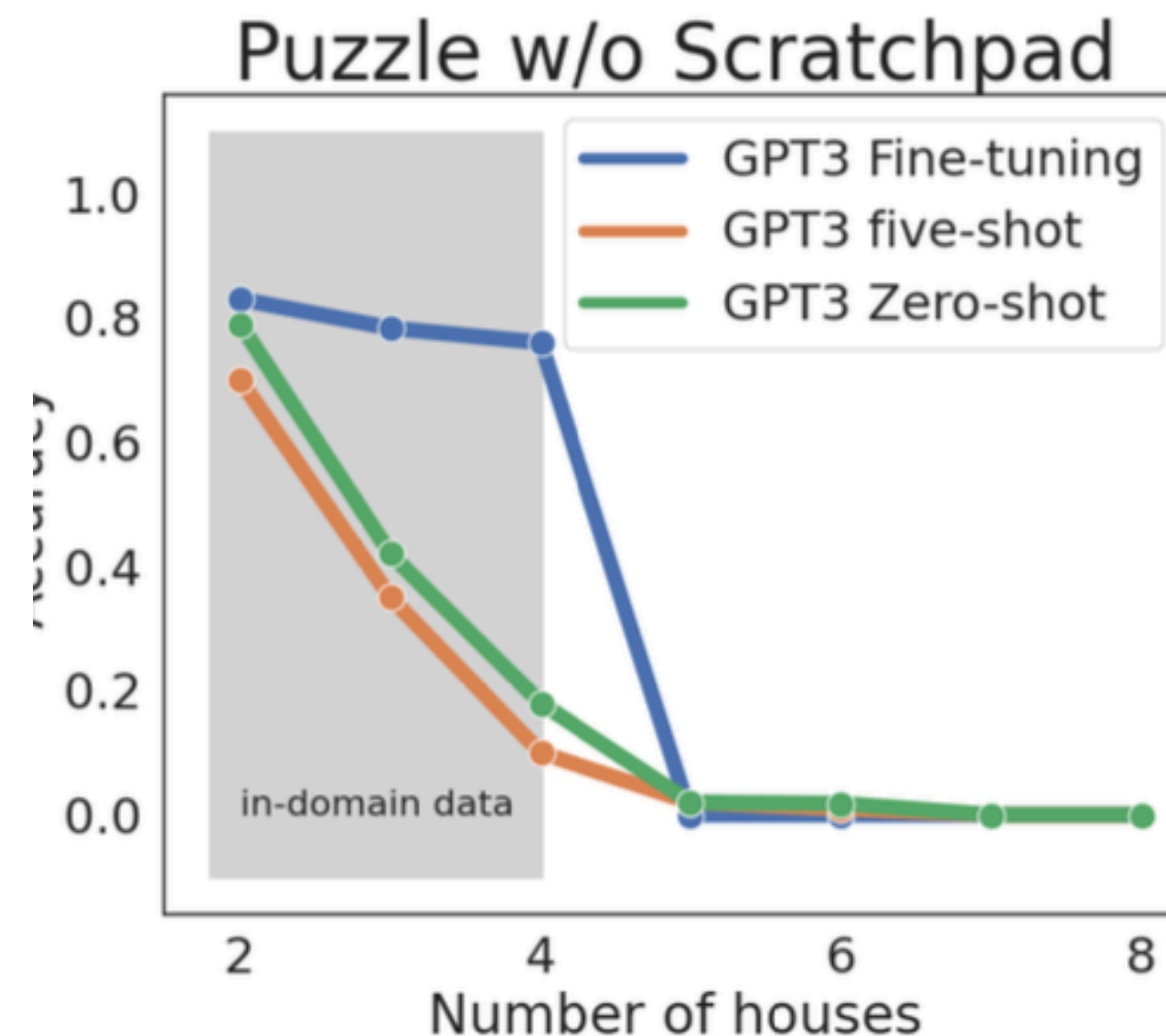


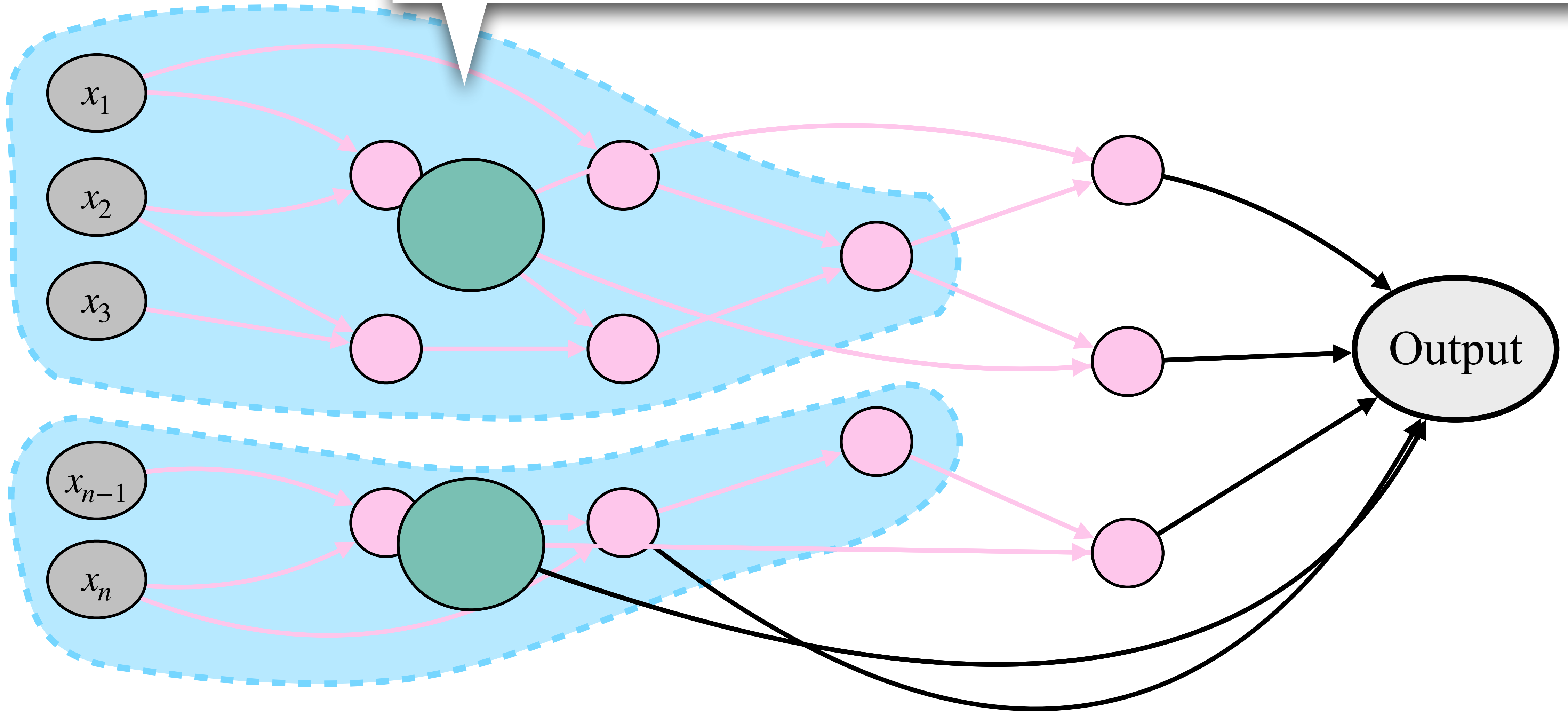
Figure 3: fine-tuning performance with in-domain data and out-of-domain data.

Systematic problem-solving capabilities do not emerge via exhaustive training on task-specific data.

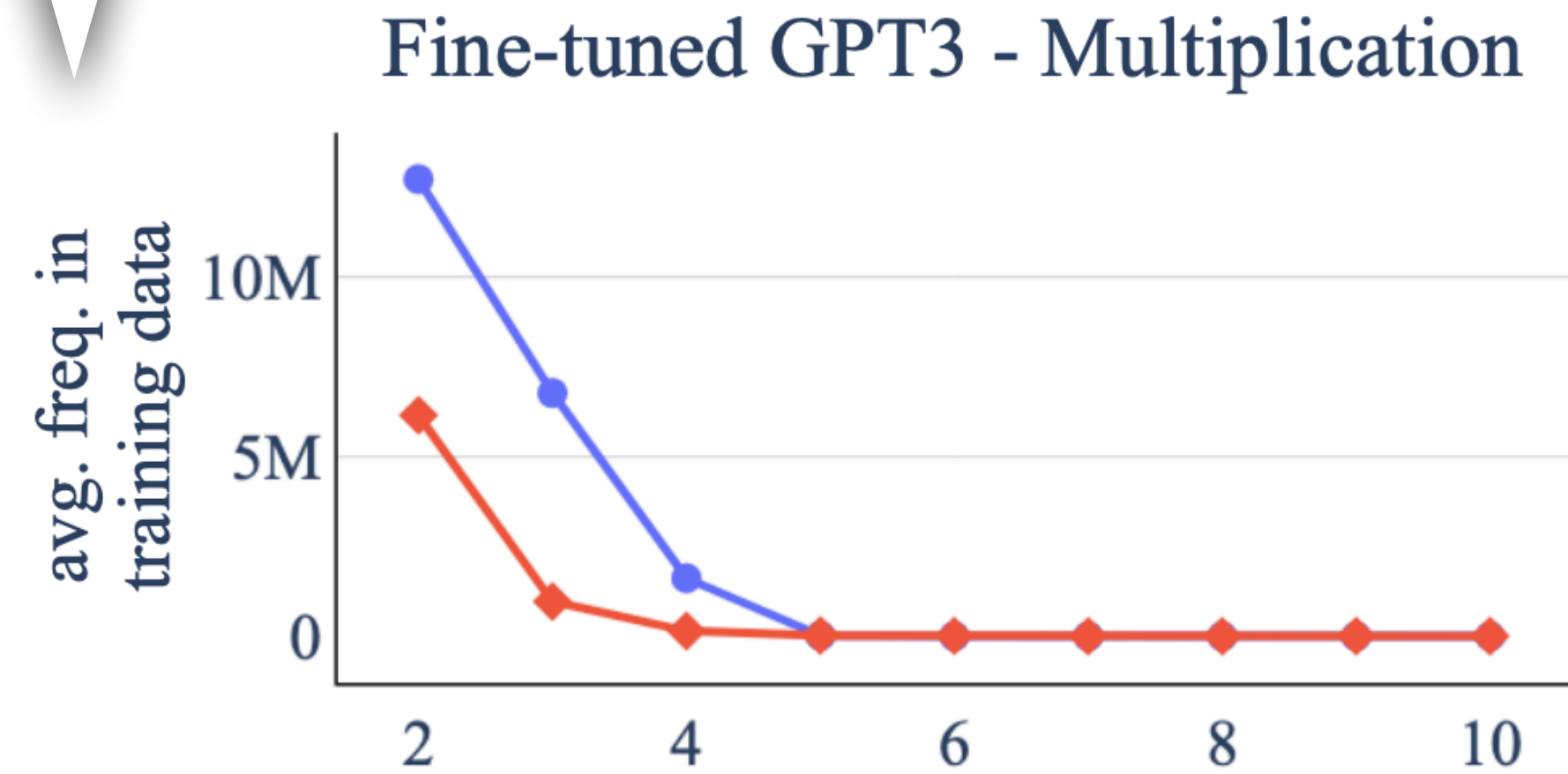
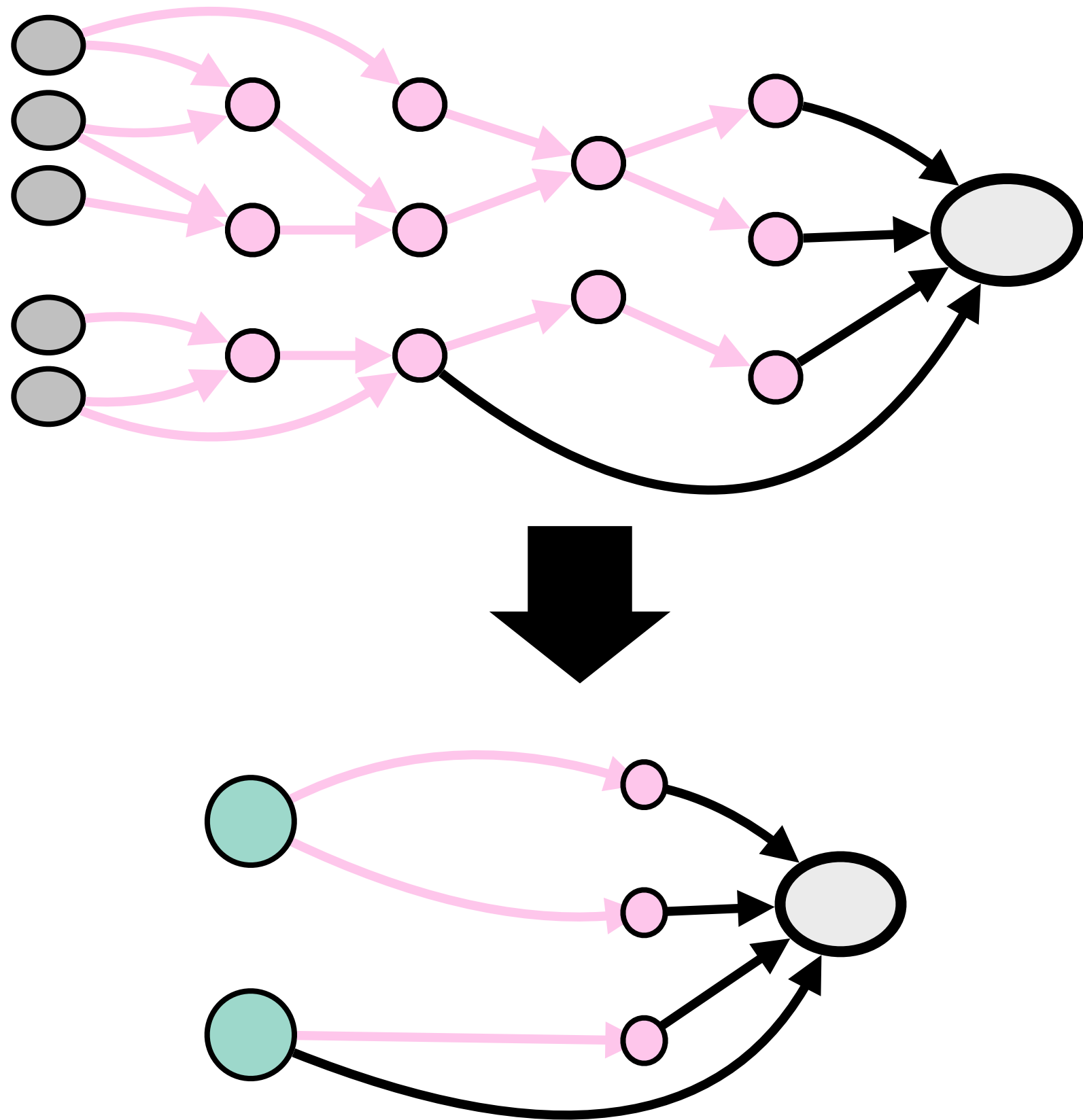
One of the key findings.

What is the correlation between a model generating a correct output and having seen relevant subgraphs during training?

Detect subgraphs already seen during training: *Identical* subgraphs during training, the inference is only *seemingly* highly compositional

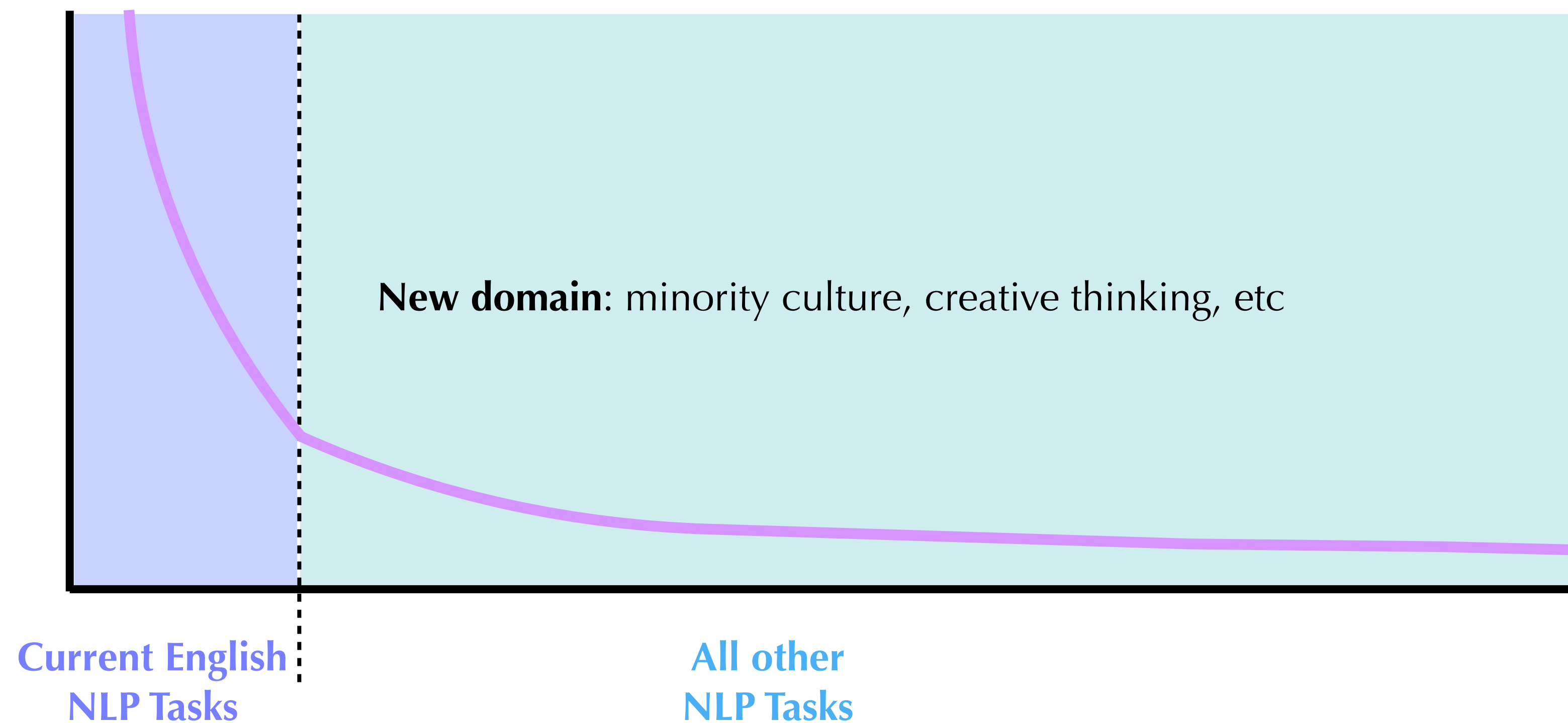


Transformers' *successes are heavily linked to having seen significant portions of the required computation graph during training*

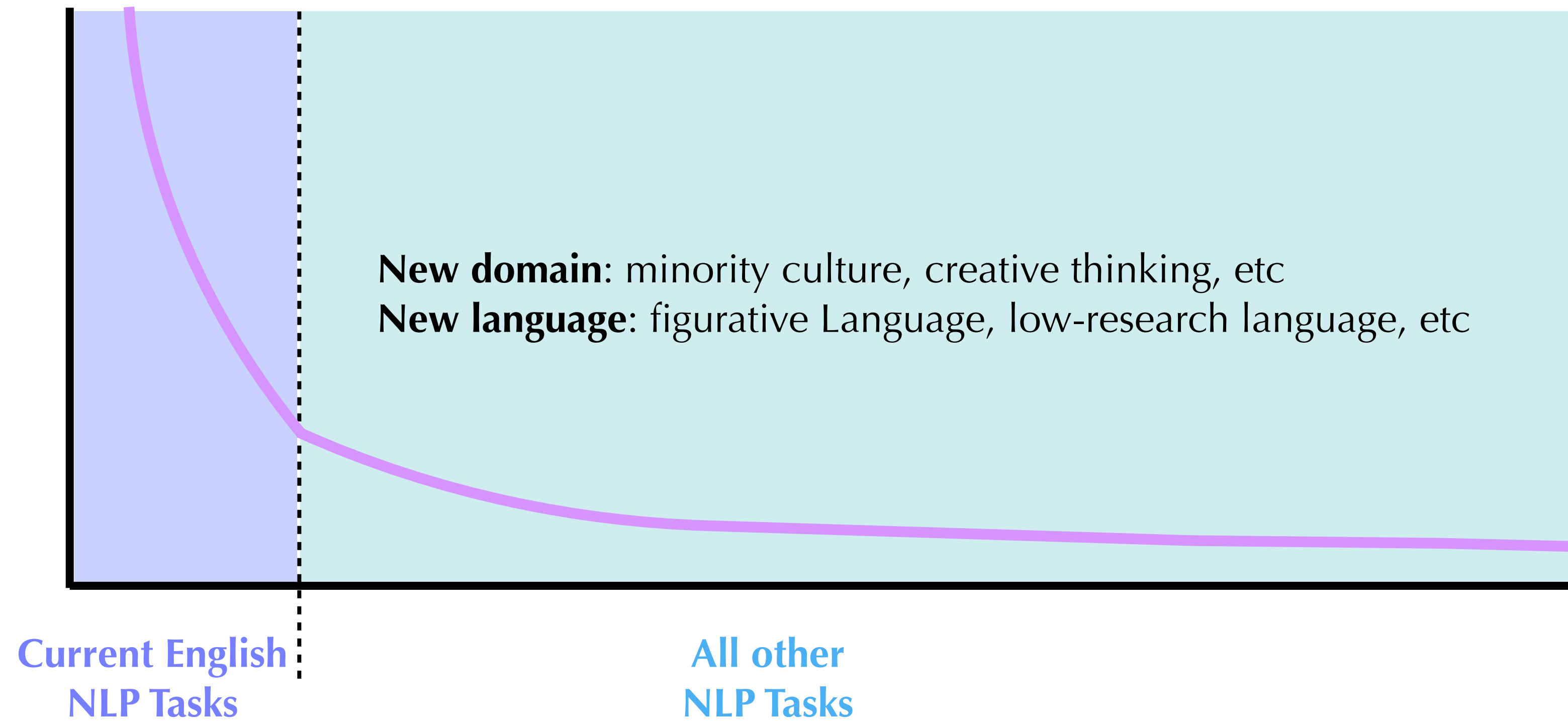


● Correct Final Answer ◆ Incorrect Final Answer

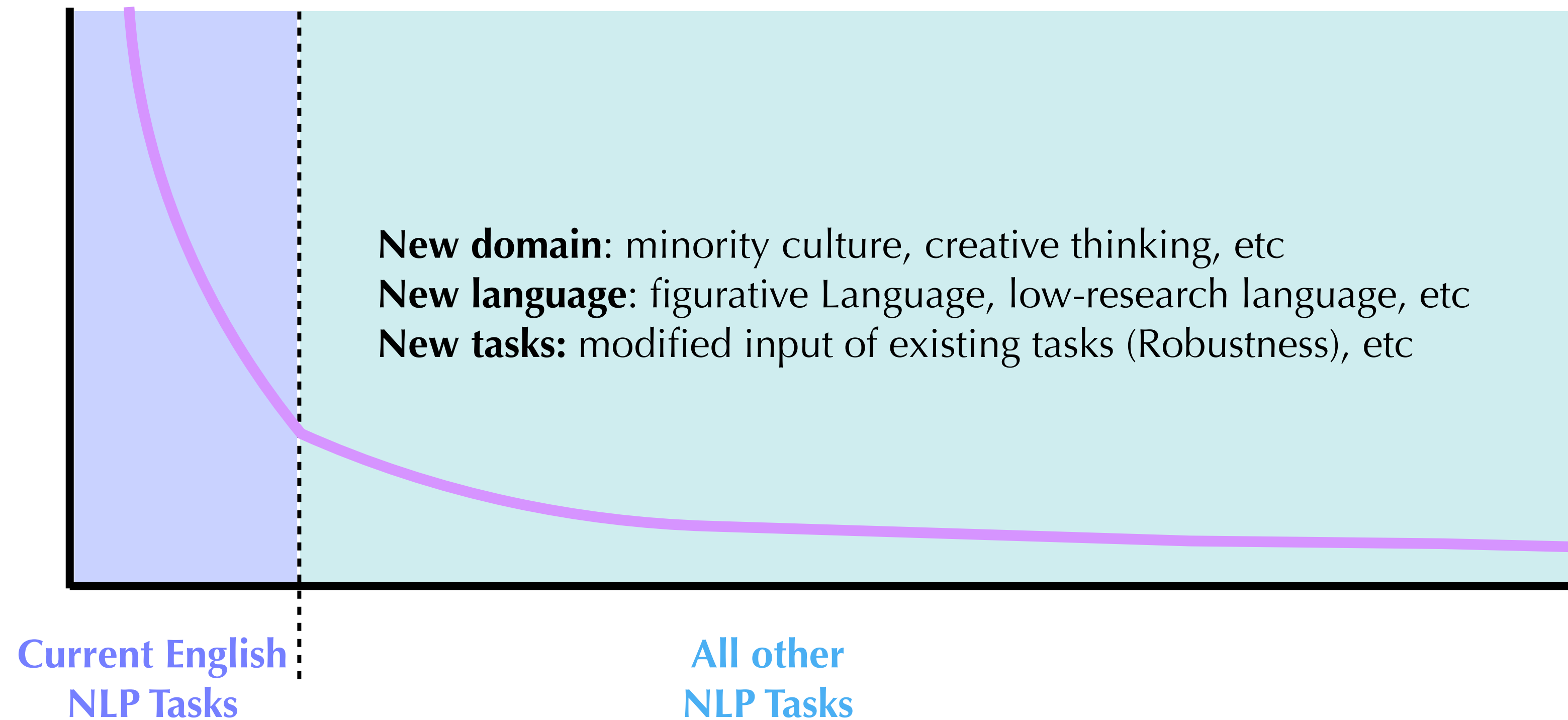
Takeaways: evaluation



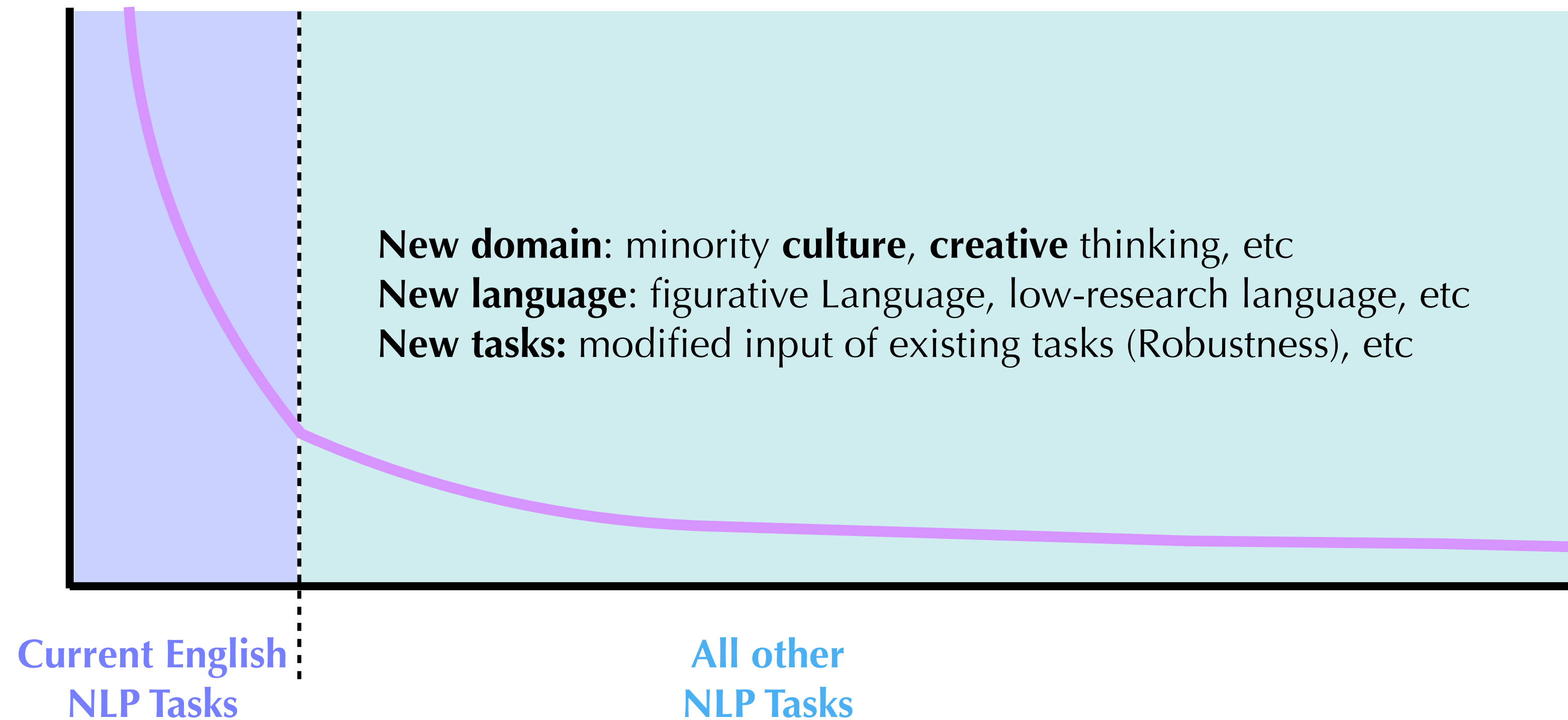
Takeaways: evaluation



Takeaways: evaluation



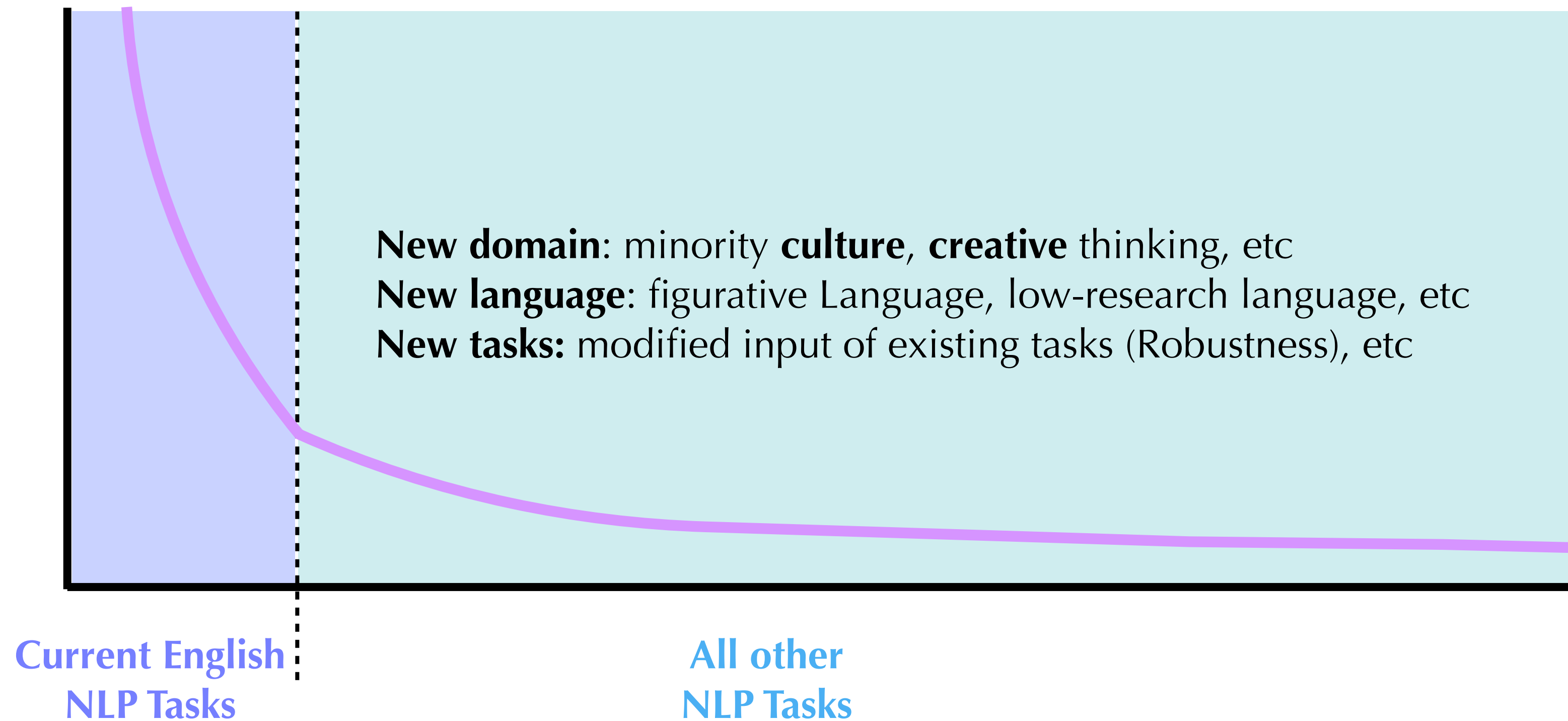
Takeaways: evaluation



Reasoning is the ability

1. to perform multiple rounds of computation before arriving at an answer (Karthik Narasimhan)

Takeaways: evaluation



Reasoning is the ability

- 1. to perform multiple rounds of computation before arriving at an answer (Karthik Narasimhan)**
- 2. to accurately adapt to new situations/new domains and new tasks.**

Takeaway: Model

- LLMs (rephrased by ChatGPT)
 - Long-Tail Challenges: Since LLMs are trained on prevalent data patterns, they might not effectively handle rare events or specialized knowledge that resides in the long tail of data distributions.
 - Reasoning Abilities: While LLMs can mimic reasoning to an extent, genuine logical reasoning, especially in multi-step or abstract contexts, remains a challenge.
- **Hybrid Models:**
 - **LLMs provide candidate sets, and statistical models provide exact solutions/probabilities.**
 - **In cases (long compositional reasoning problems) where LLMs can not give us ample or correct candidate sets, trace back to the model predictions (structural reasoning, knowledge graph) and correct them at their location (model editing, etc.)**