# Probabilisitc Commonsense Knownledge Evaluation


Xiang Lorraine Li

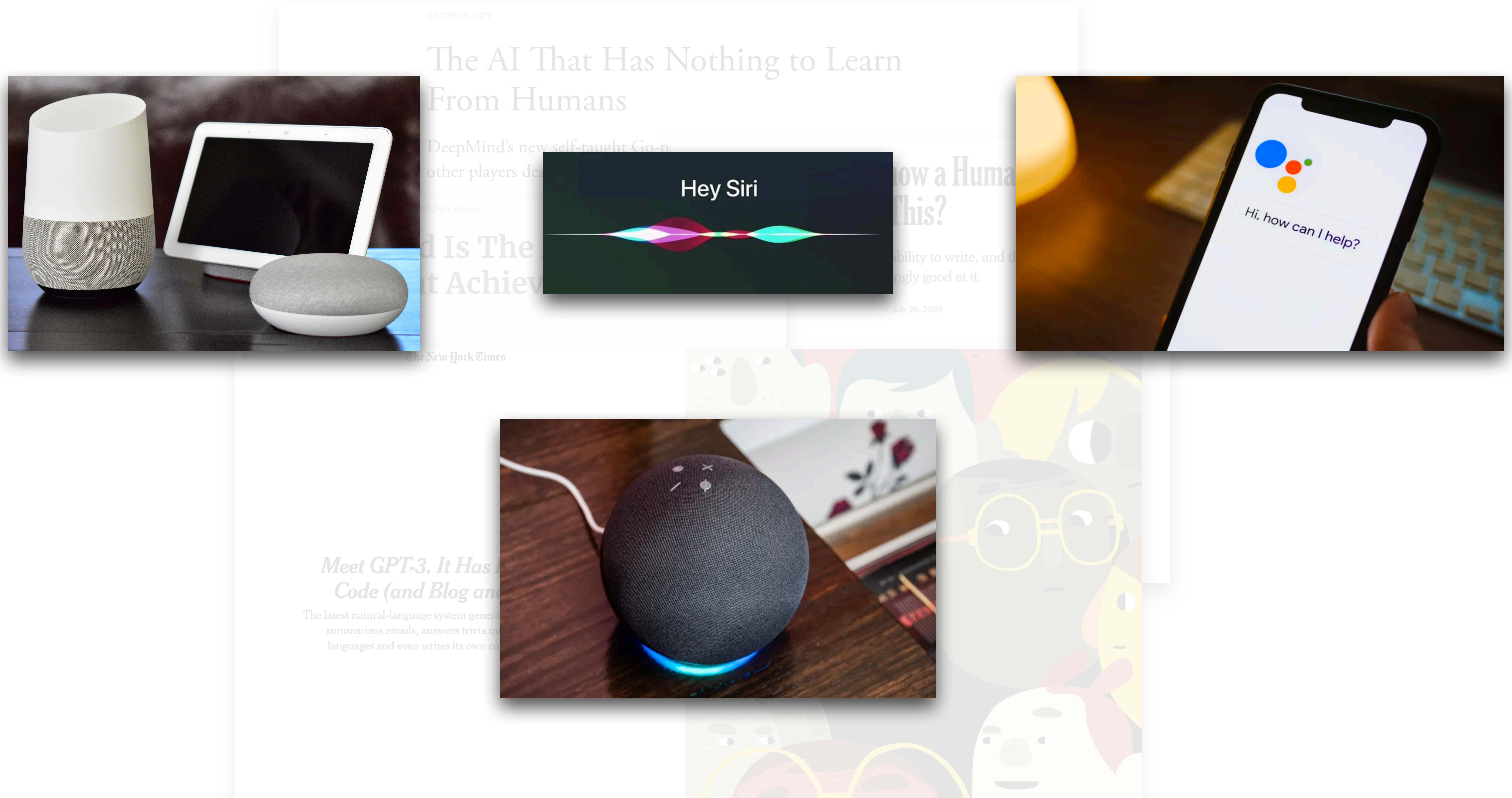
Assistant Professor at SCI Pitt

UMassAmherst
Robert and Donna Manning College of Information & Computer Sciences

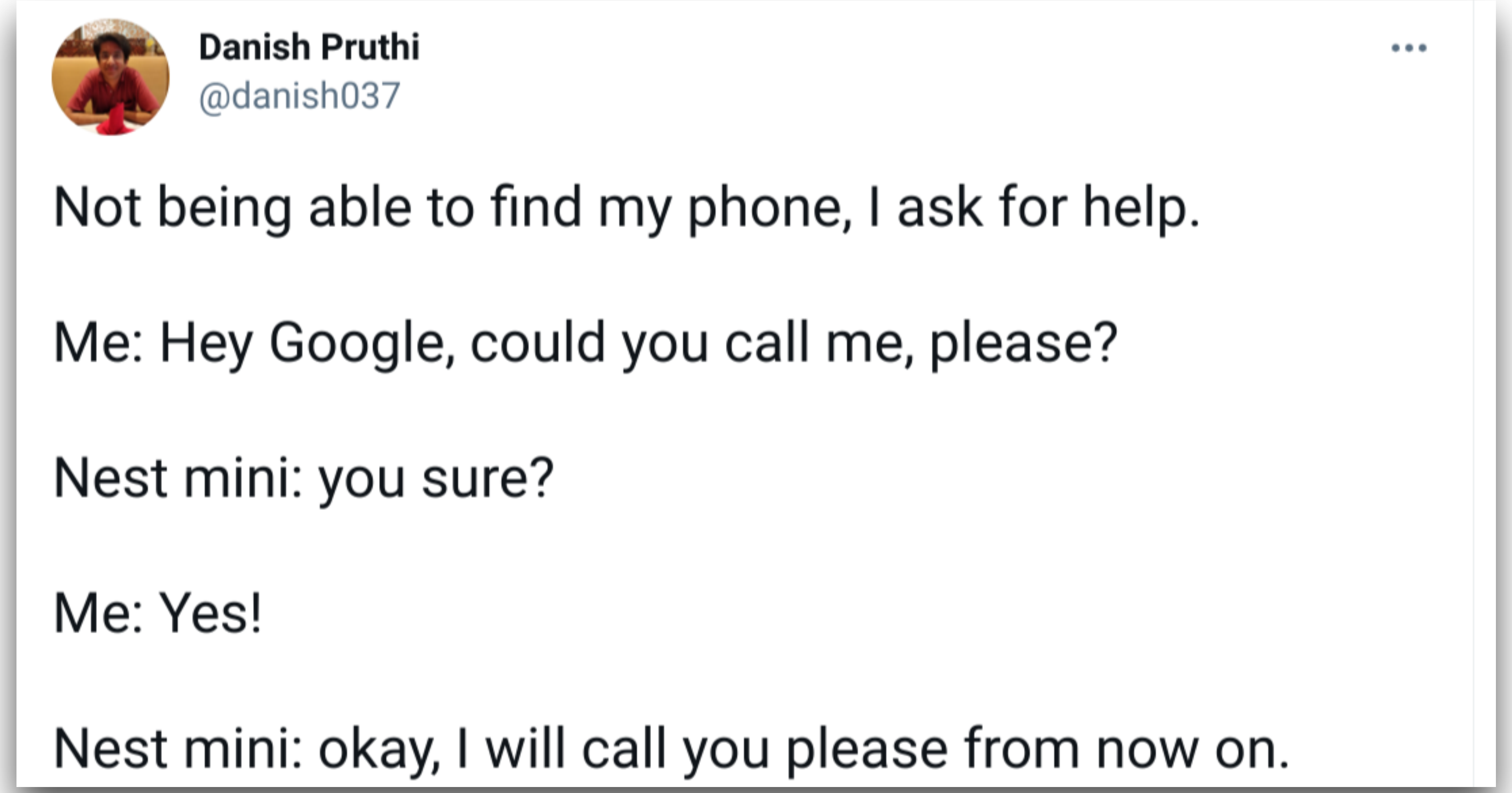
University of Pittsburgh

Ai2 Allen Institute for AI

# Impressive Progress in AI



# 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. Boiled water can cook food. Water needs to be held in a container.
Sugar can melt in water
There are usually waiter helping you order food.
Heat is needed to boil water.
Vater is liquid.
**Open the jelly jar.**
Water evaporates.
**They boiled the water, then added sugar.**
Heat is needed to boil water. When it's cloudy, sometimes there is no sunset.
Some people love sugar.
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.
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.
Human feel satisfied after having sweet stuff.
Heat is needed to boil water. Water can be used to wash clothes.
**They walked along the river at sunset time.**
Water needs to be h
Most bread is not sweet Sugar water is also liquid.
Boiled water can cook food. Boiled water is too hot to drink.
People are walking a
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.
**They boiled the water.**
Water is wet. Human feel satisfied after having sweet stuff.
Humans drink water. There are usually waiter helping you order food.
Water is wet.
People needs tools to put peanut butter on the bread. Opening a jar needs tool Sometimes the ordering is automatic too.
Person can open jar, but not dogs 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
**She walked into a restaurant and started ordering** Boiled water can cook food.
Heat is needed to boil water.
**Spread the peanut butter on the bread.**
Some people love sugar.
Human feel satisfied after having sweet stuff. Sometimes the ordering is automatic
Peanut butter on the bread is usually breakfast.
People walk into restaurant through door Person can open jar, but not cats.
Water can be used for cleaning.
People who wants to lose weight usually avoid peanut butter. Ordering food needs menu Restaurant serves food.
Sweet water tastes good
Boiled water can cook food.
Peanut butter is high calorie food. Water is wet. Walking into a restaurant usually at breakfast/lunch/dinner time.
Water needs to be held in a container.
Some people hate sugar.
Human feel satisfied after having sweet stuff. Most bread is not sweet
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
Ordering food needs menu
**Tom asked me how to get to the library.**
Water is wet. Boiled water can cook food.
Peanut butter is sweet
The kind of bread that can add peanut butter is flat.
Some people hate peanut butter.
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

# 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?

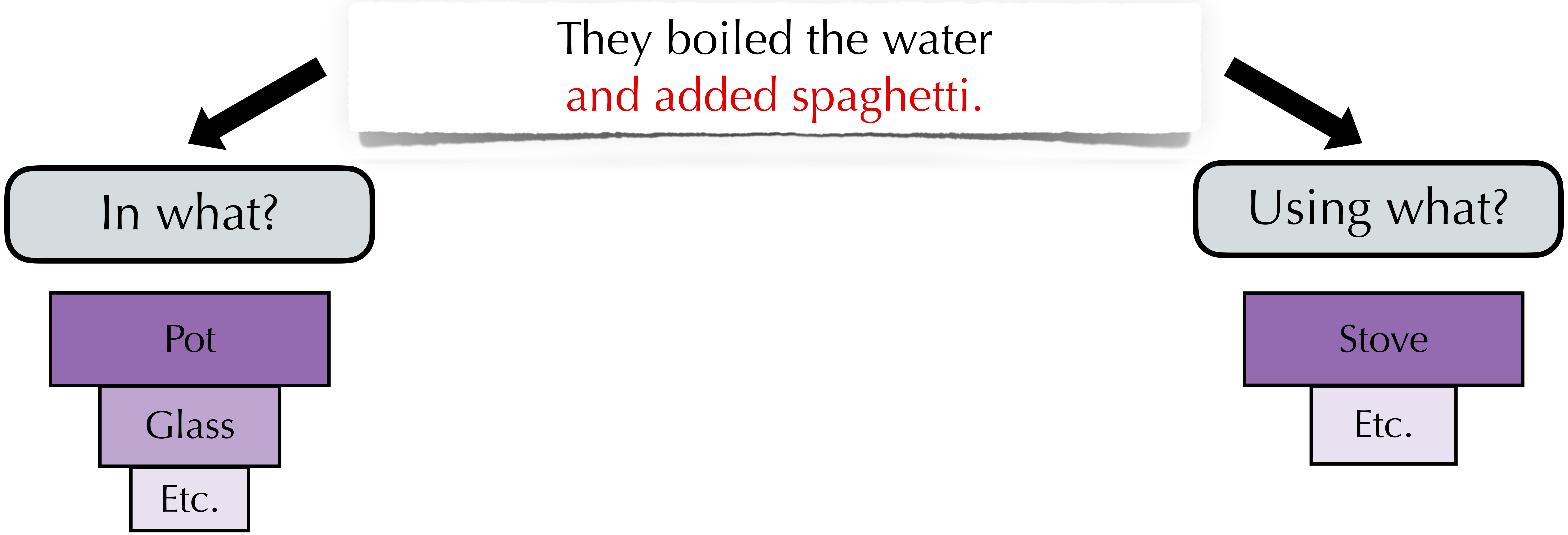
- Pot
- Glass
- Etc.

Using what?

- Stove
- Etc.

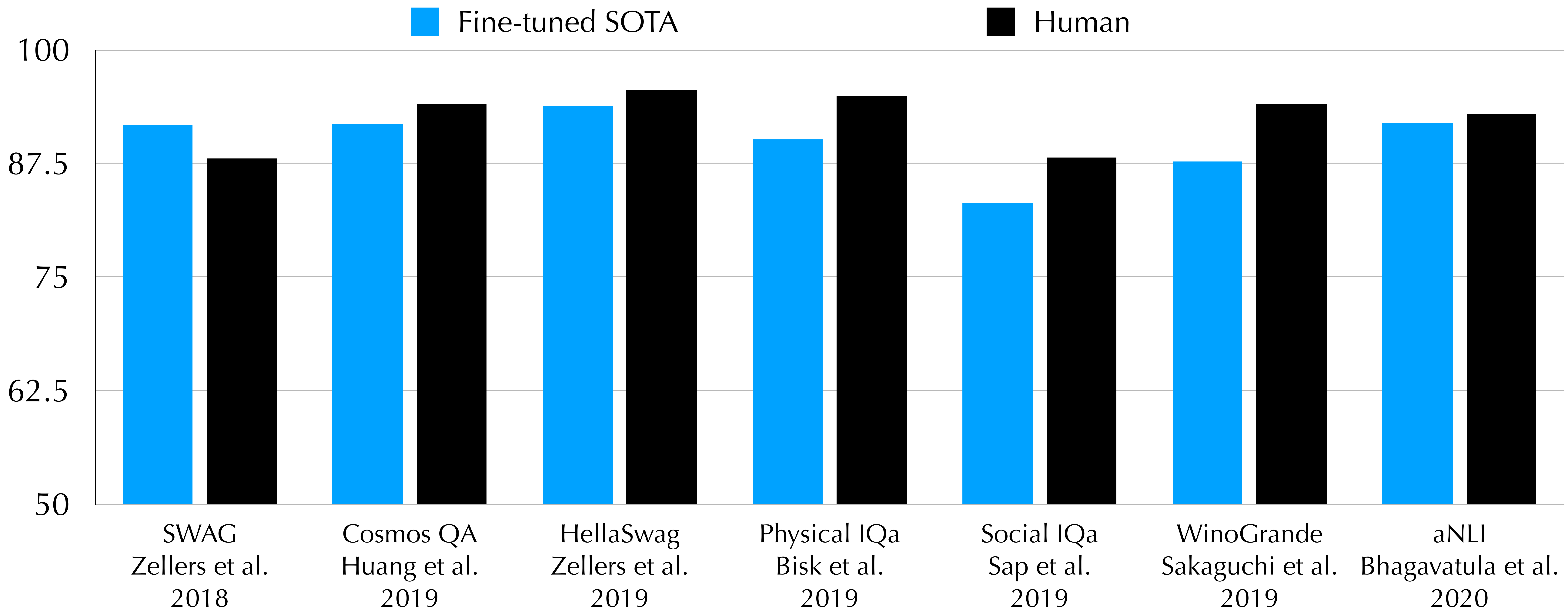
# Why is Common Sense Challenging?

**Massive** | **Probabilistic** | **Contextual**



# Common Sense in Language Model

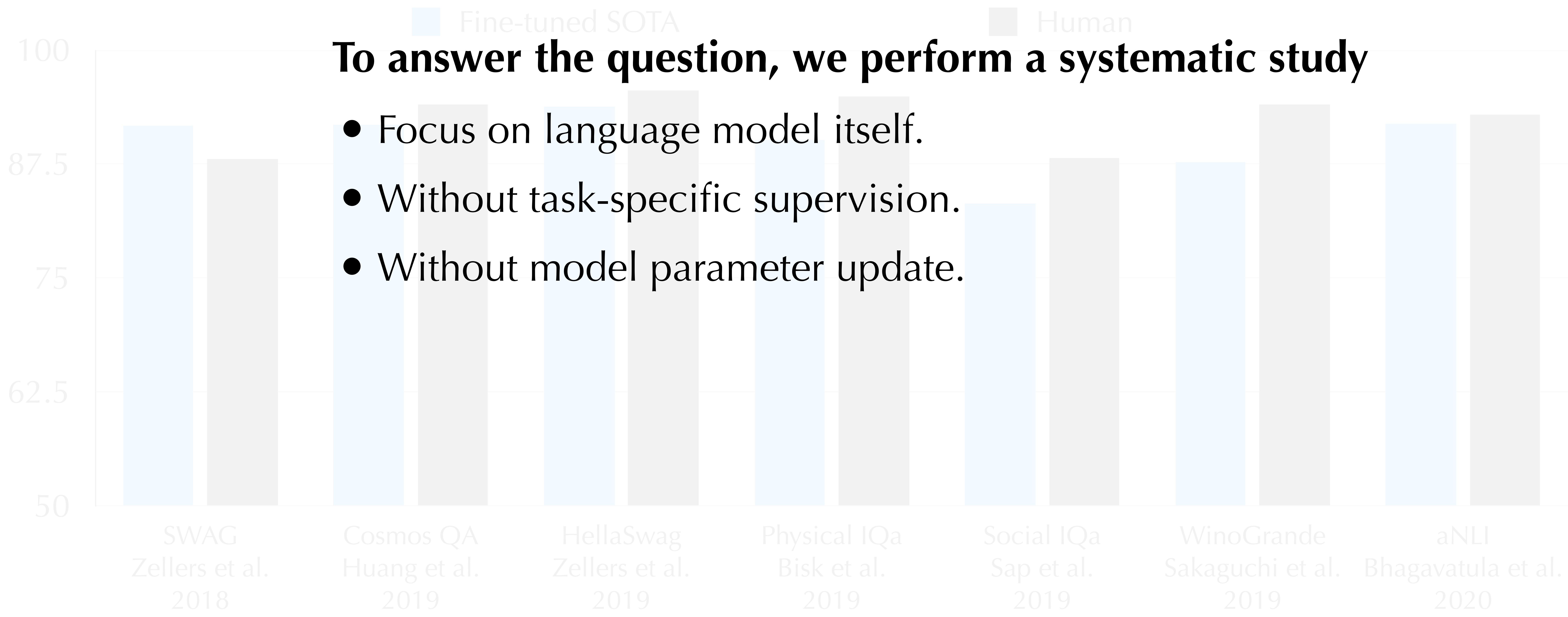
Models based on large language models show impressive performance on many **commonsense question answering** tasks.



# 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.

Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan et al. "Language models are few-shot learners." *NeurIPS 2020*

# 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 |

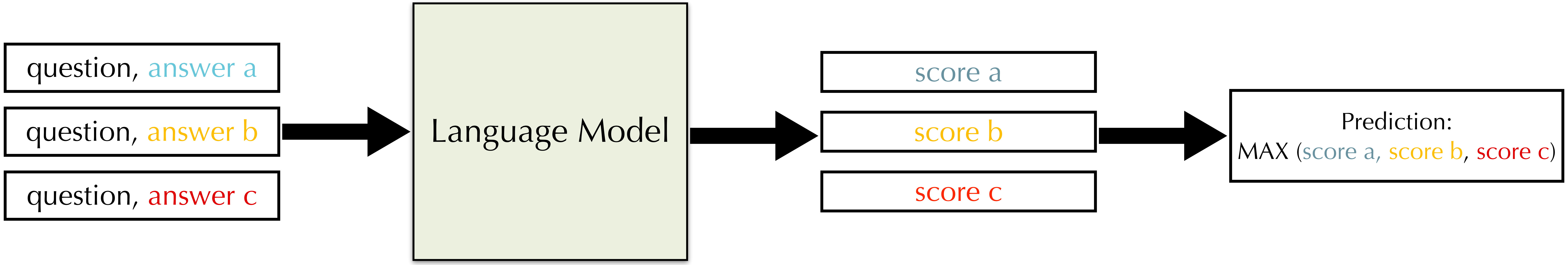
Four multiple choice selection QA datasets.

**Li**, Kuncoro, Hoffmann, d'Autume, Blunsom, Nematzadeh. ``A Systematic Investigation of Commonsense Knowledge in Large Language Models" EMNLP2022.

# 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

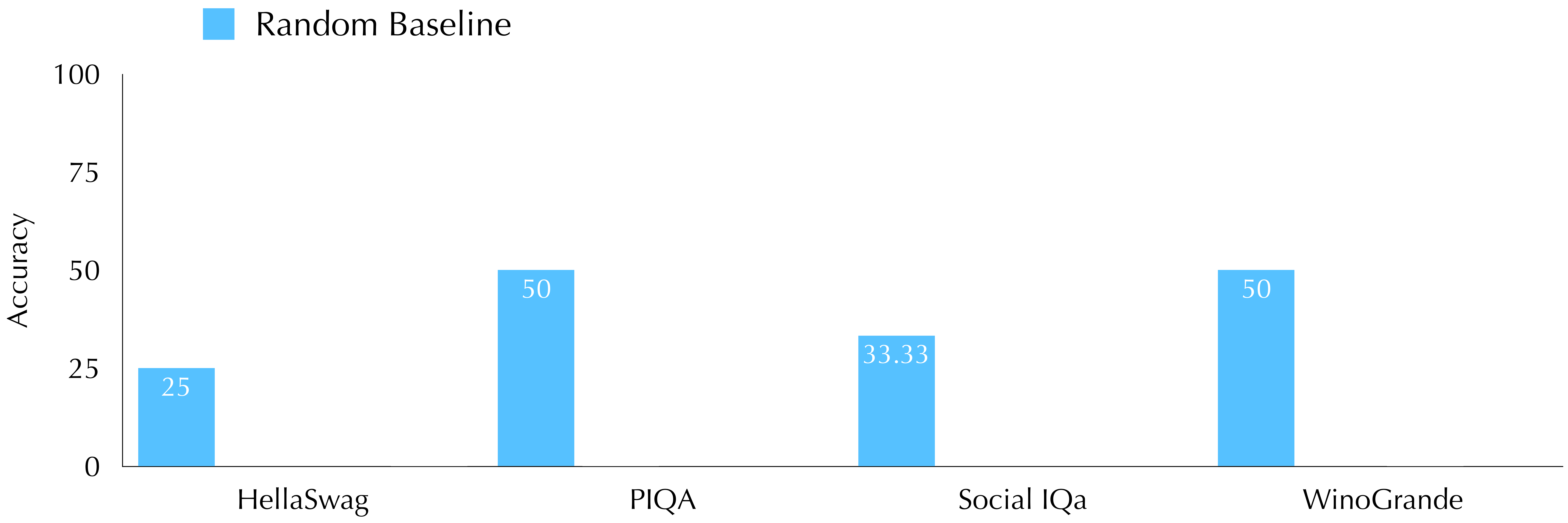



Figure: the dev accuracy for each dataset evaluated on Gopher.

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

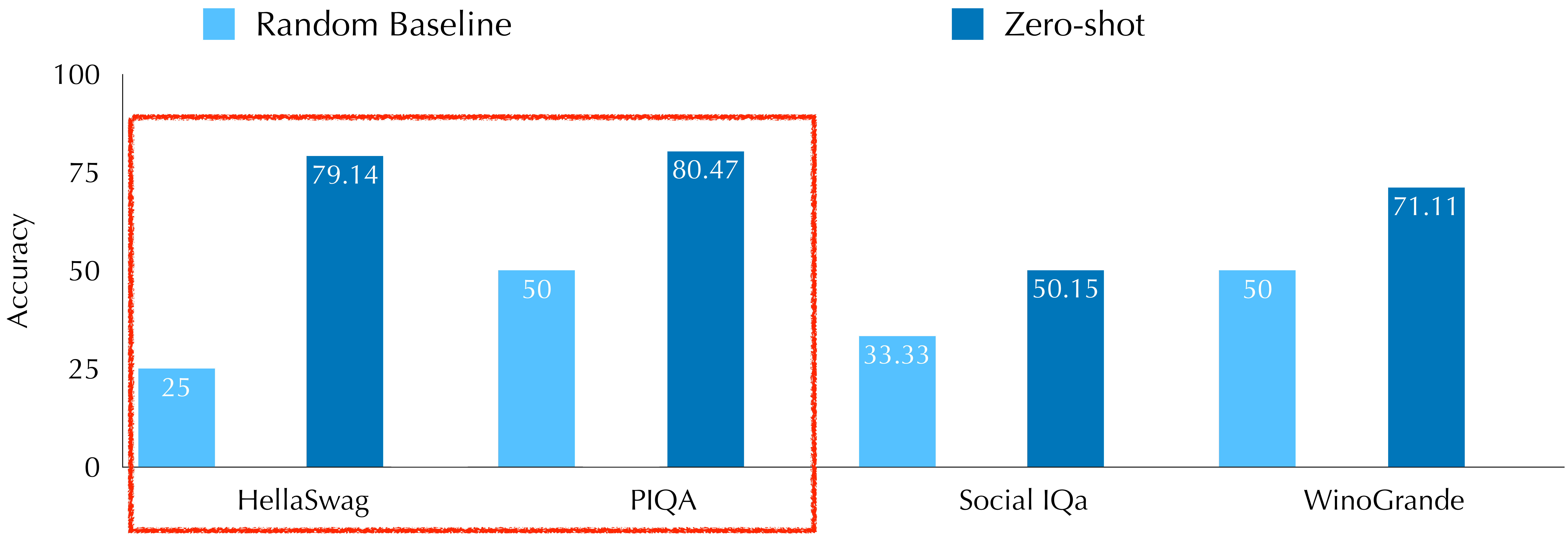



Figure: the dev accuracy for each dataset evaluated on Gopher.

**Li**, Kuncoro, Hoffmann, d'Autume, Blunsom, Nematzadeh. ``A Systematic Investigation of Commonsense Knowledge in Large Language Models" EMNLP2022.

# How much of the performance comes only from answers?

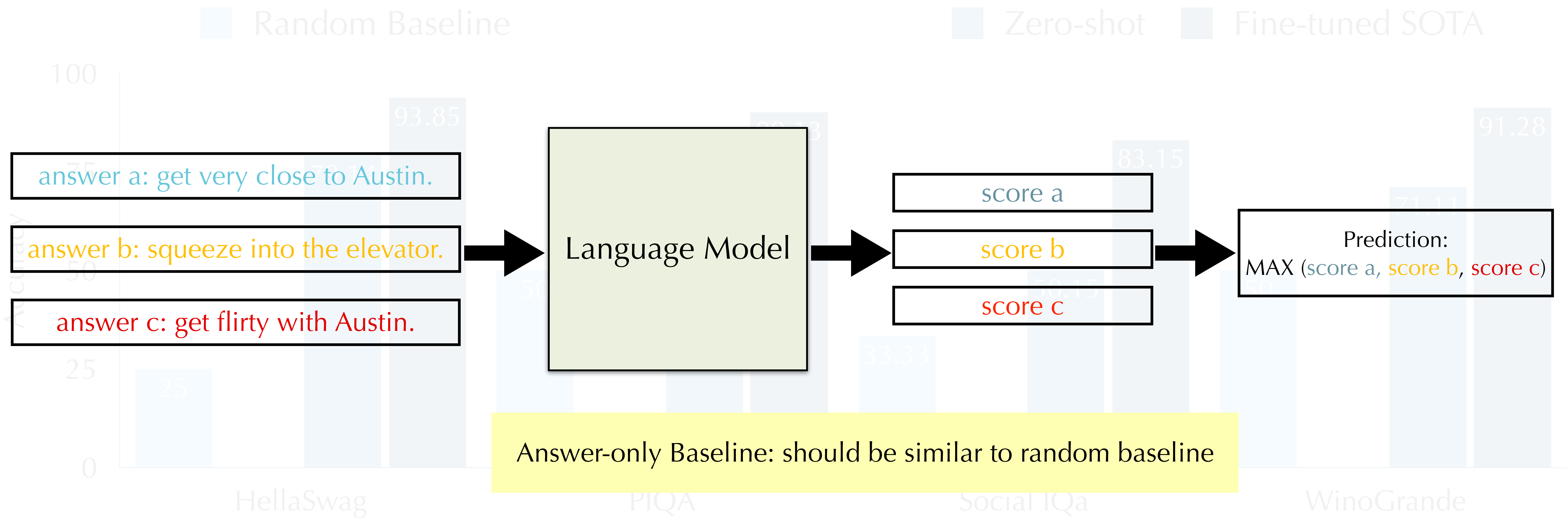



Figure: the dev accuracy for each dataset evaluated on Gopher.

# Models pick the correct answer without seeing the question

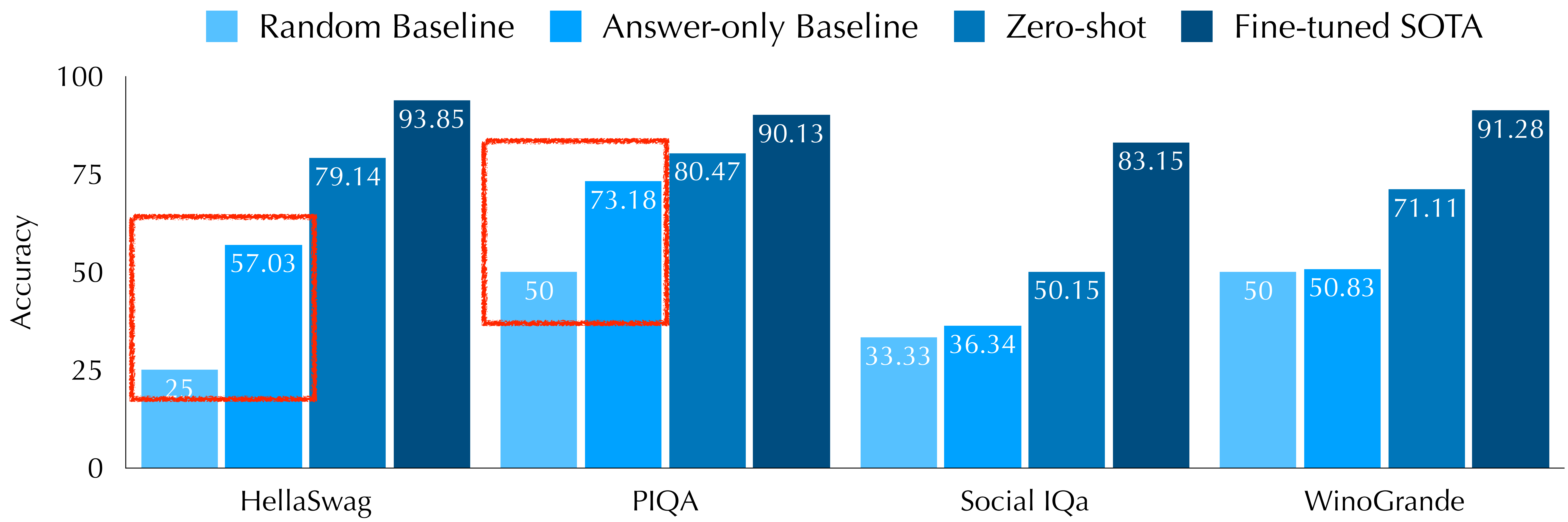



Figure: the dev accuracy for each dataset evaluated on Gopher.

# We need better commonsense evaluation!

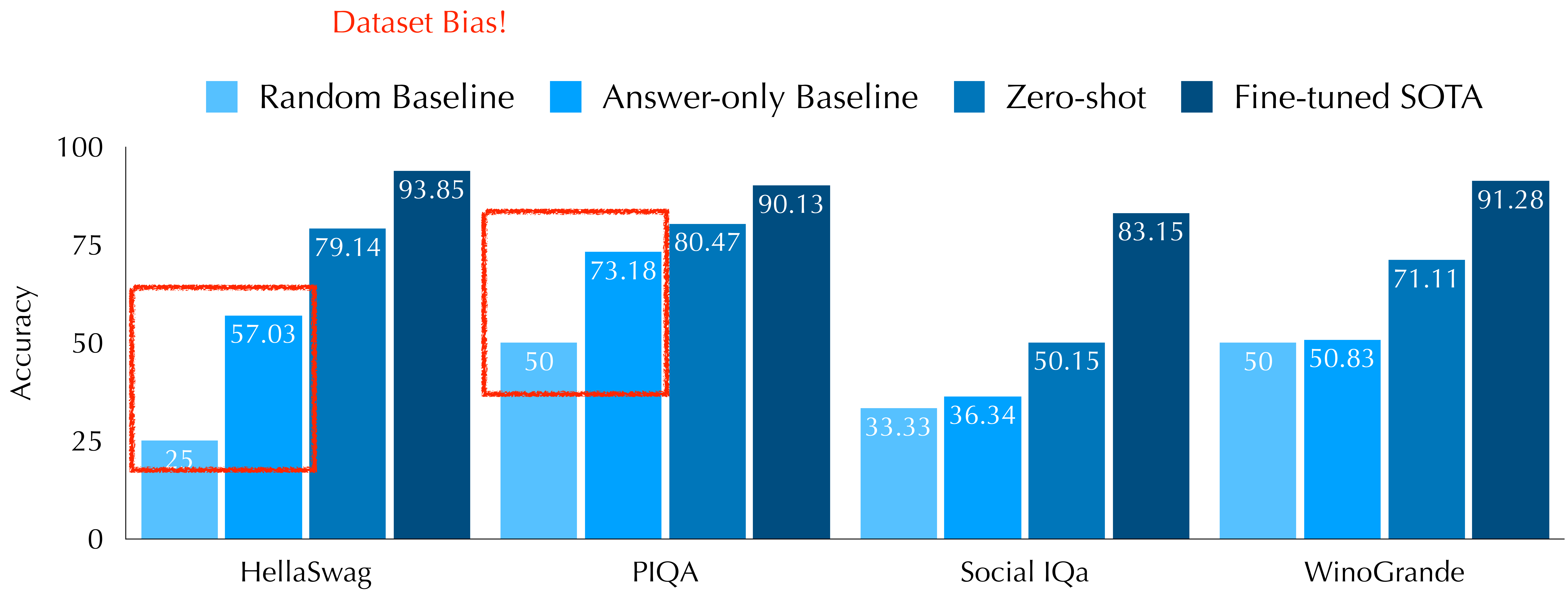



Figure: the dev accuracy for each dataset evaluated on Gopher.

**Li**, Kuncoro, Hoffmann, d'Autume, Blunsom, Nematzadeh. ``A Systematic Investigation of Commonsense Knowledge in Large Language Models" EMNLP2022.

# Outline

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

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

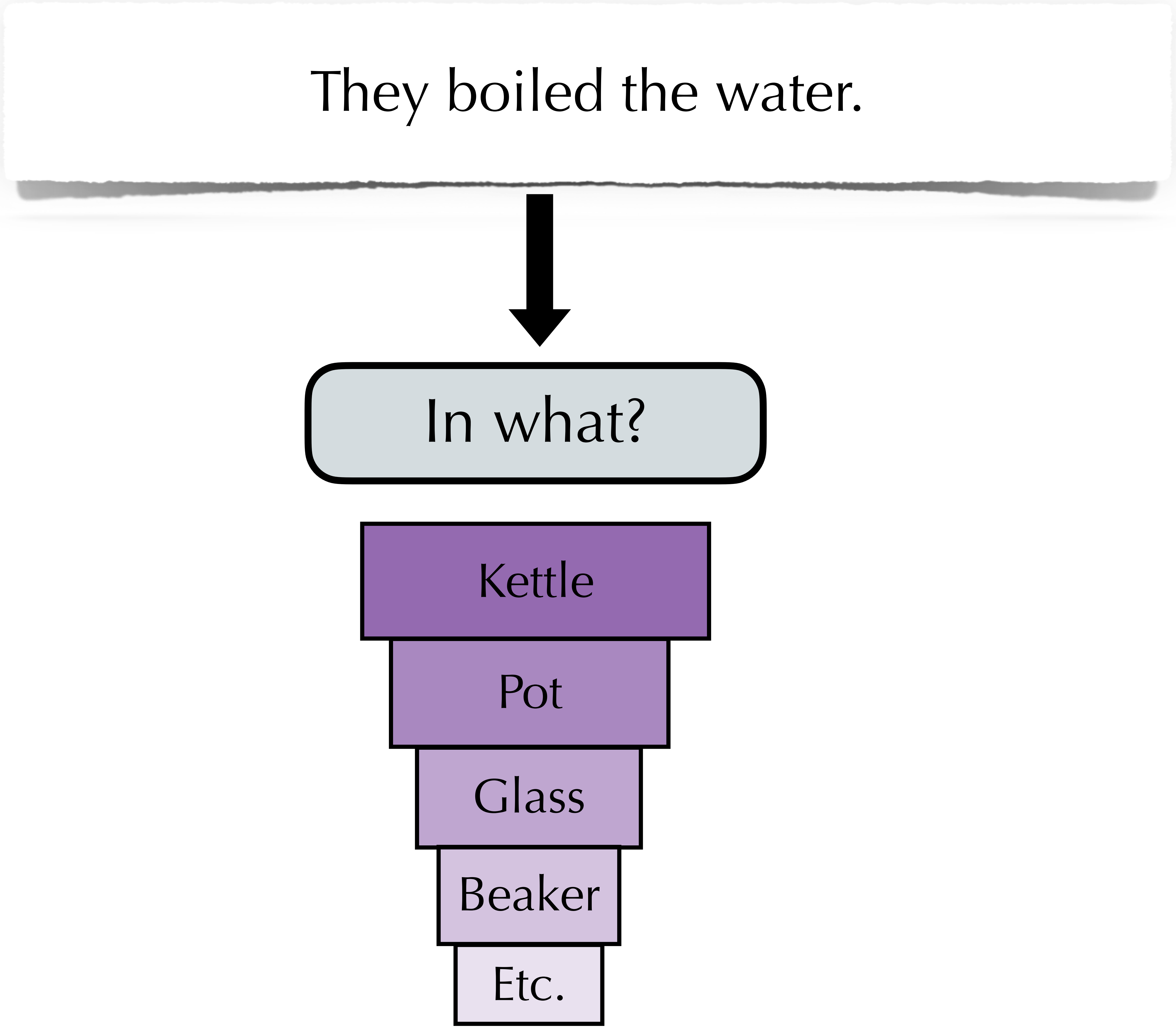
**Benchmark: Long-tail Question: Commonsense Reasoning Evaluation**

- 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]

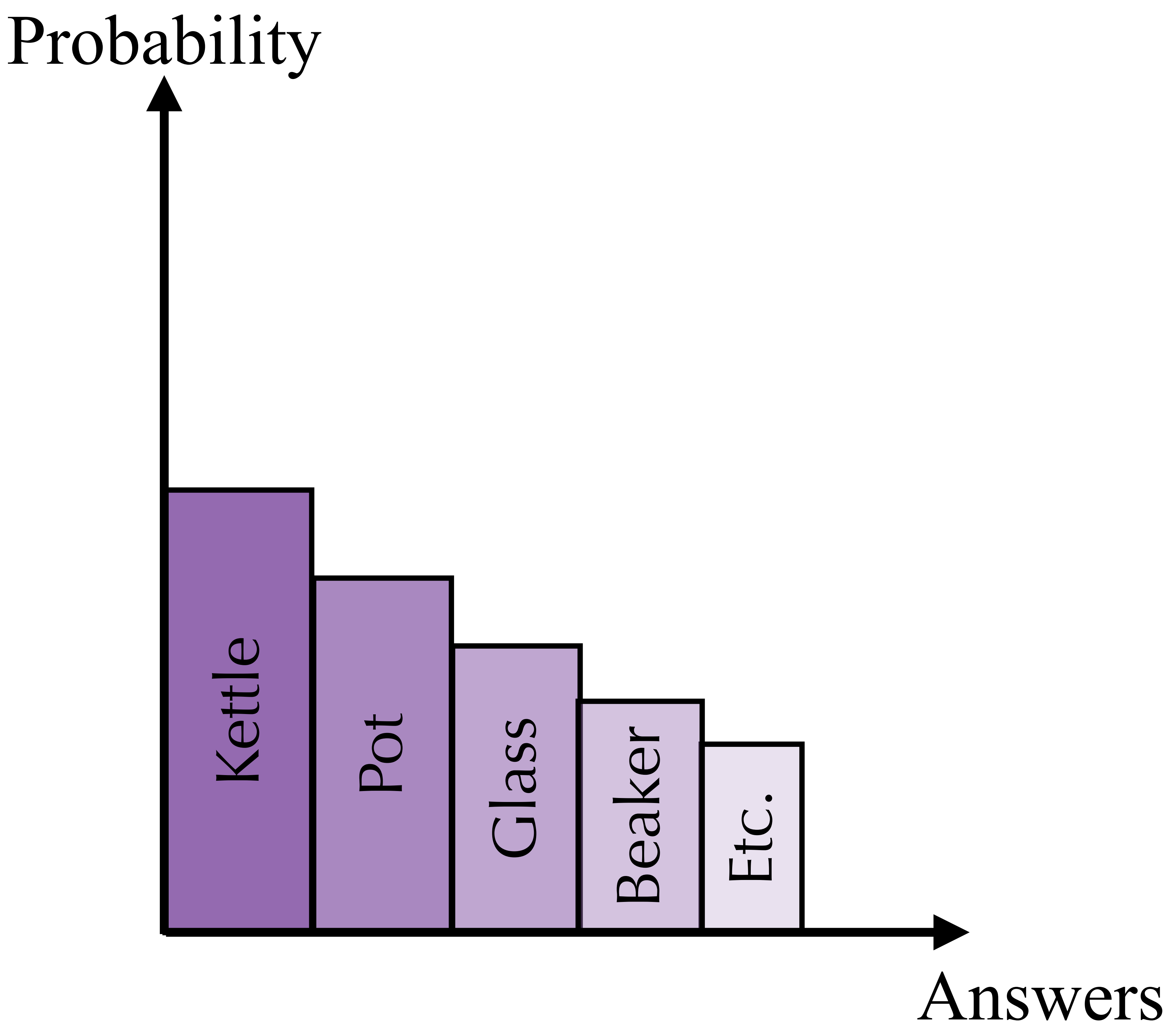
# Probabilistic Evaluation of Commonsense



# Probabilistic Evaluation of Commonsense

They boiled the water.

In what?




Question Answering

Dialogue

Any language tasks!



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

# 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

Context Sentence → Semantic Parsing

"Dog catches the thrown frisbee."


CommonGen (Image Captions)

AMR Parsing

# CFC Data Collection

We crowd-source high-quality evaluation data

Context Sentence → Semantic Parsing → Missing Slot Identification

"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)**

Context Sentence → Semantic Parsing → Missing Slot Identification

| Missing Slot | Definition | Examples |
|---|---|---|
| Arg0 | Who/what does the event? | Sentence: putting cheese on the pizza. Arg0?<br>Answers: person, cook |
| Purpose | What is the goal for doing the event? | Sentence: putting cheese on the pizza. Purpose?<br>Answers: get nutrition, stop being hungry |
| Instrument | What kind of tools are used to accomplish the event? | Sentence: putting cheese on the pizza. Instrument?<br>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?<br>Answers: lunch time, dinner time |
| Location | Where would the event usually happen? | Sentence: putting cheese on the pizza. Location?<br>Answers: kitchen, restaurant |

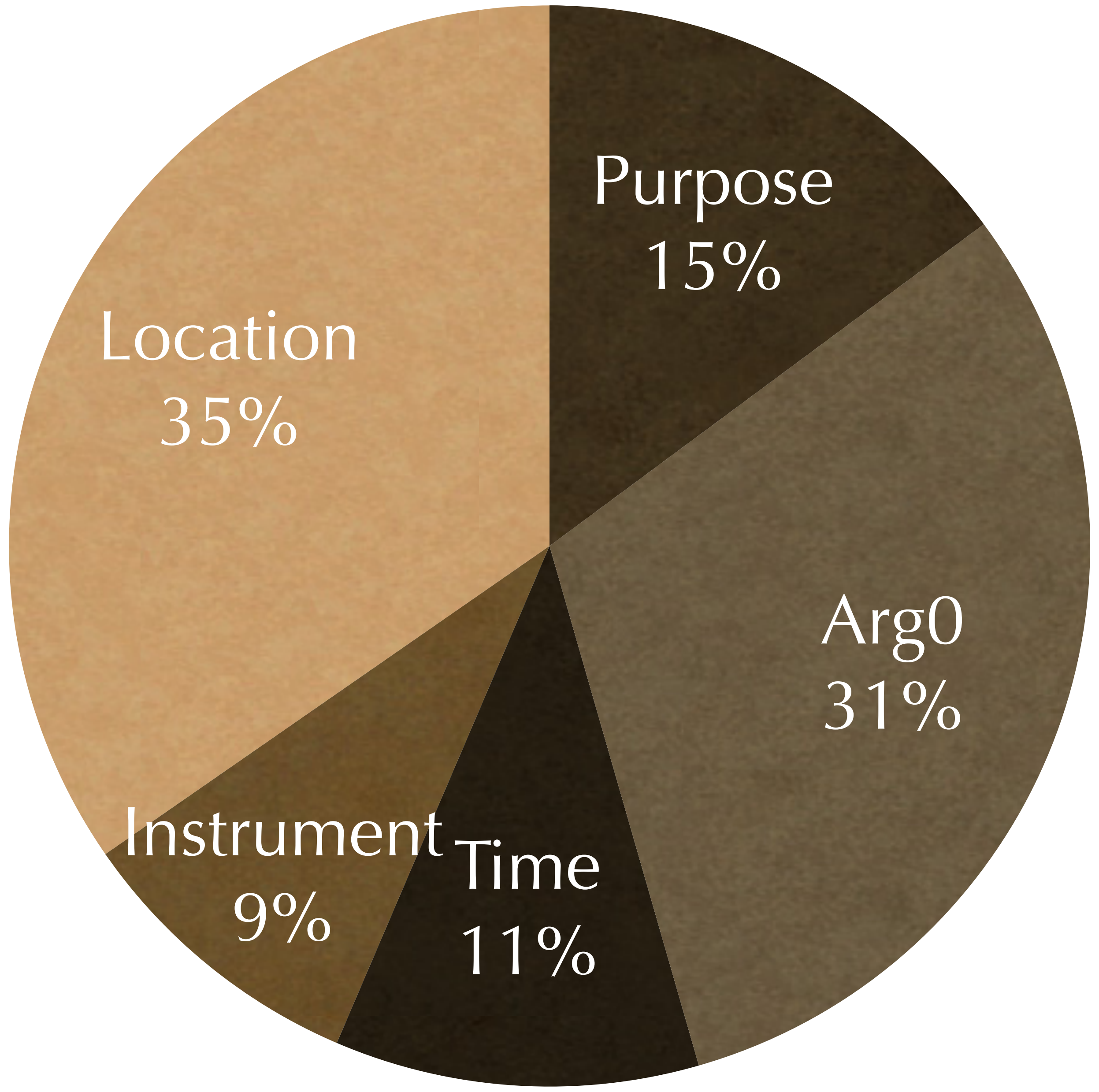
Qi, Boratko, Yelugam, O'Gorman, Singh, McCallum, Li. "Every Answer Matters: Evaluating Commonsense with Probabilistic Measures" ACL 2024

# CFC Data Collection

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

Context Sentence → Semantic Parsing → Missing Slot Identification

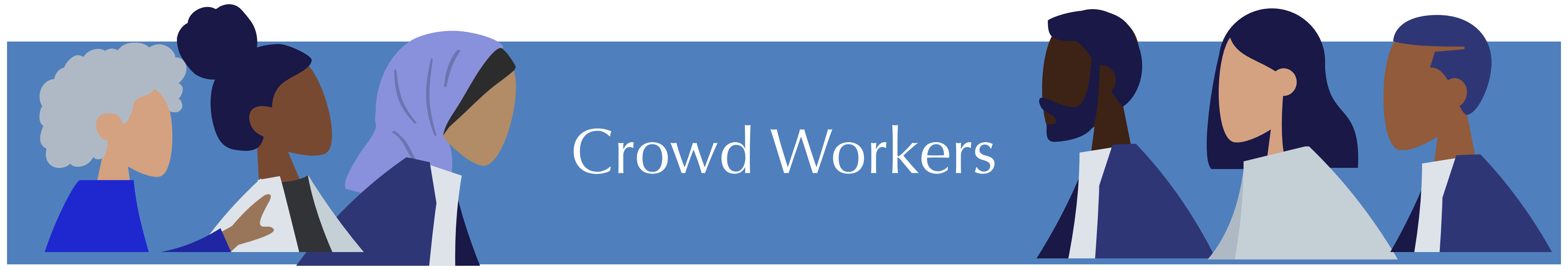
| Missing Slot | Definition | Examples |
|---|---|---|
| Arg0 | Who/what does the event? | Sentence: putting cheese on the pizza. Arg0?<br>Answers: person, cook |
| Purpose | What is the goal for doing the event? | Sentence: putting cheese on the pizza. Purpose?<br>Answers: get nutrition, stop being hungry |
| Instrument | What kind of tools are used to accomplish the event? | Sentence: putting cheese on the pizza. Instrument?<br>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?<br>Answers: lunch time, dinner time |
| Location | Where would the event usually happen? | Sentence: putting cheese on the pizza. Location?<br>Answers: kitchen, restaurant |



# CFC Data Collection

"They boiled the water" Purpose?

cooking
clean
disinfect
cook
make tea
disinfecting
making dinner
for making tea
cleaning
for a hot drink
to cook
cook food
cleaning tools
making pasta
cooking spaghetti
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?

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

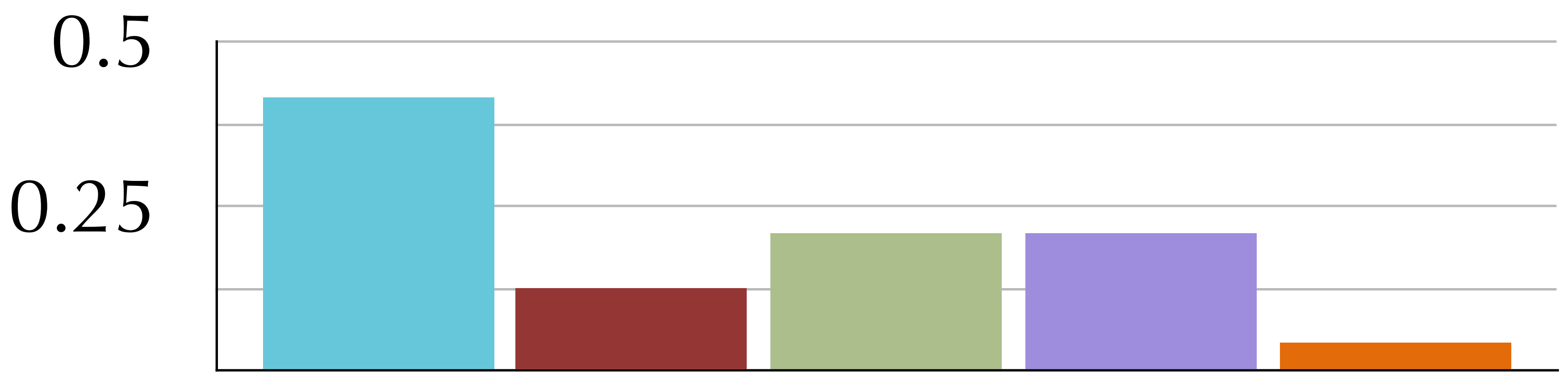
for making tea
for a hot drink
make tea

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

clean
disinfect
disinfecting
cleaning
cleaning tools

kill bacteria
purify
purification
make safe to drink
sterilization

for an experiment



# 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]

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

[1] Mardia, Jay, Jiantao Jiao, Ervin Tánczos, 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.

# 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]

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

- $n$: number of samples
- $k$: number of category in the categorical distribution
- $\epsilon$: KL error rate

[1] Mardia, Jay, Jiantao Jiao, Ervin Tánczos, 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

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

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

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

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

[1] Mardia, Jay, Jiantao Jiao, Ervin Tánczos, 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

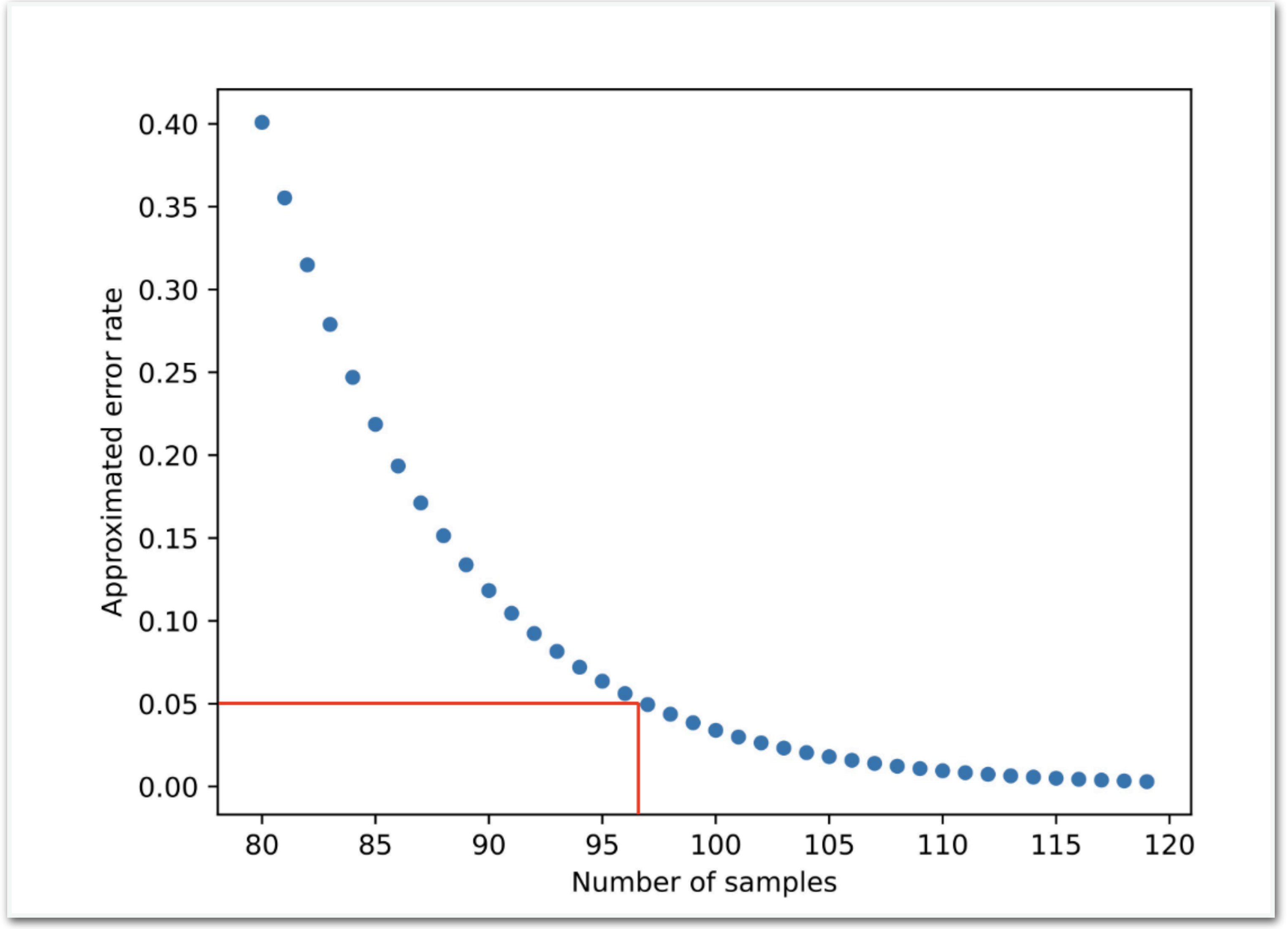
# CFC Data Collection

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

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

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

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



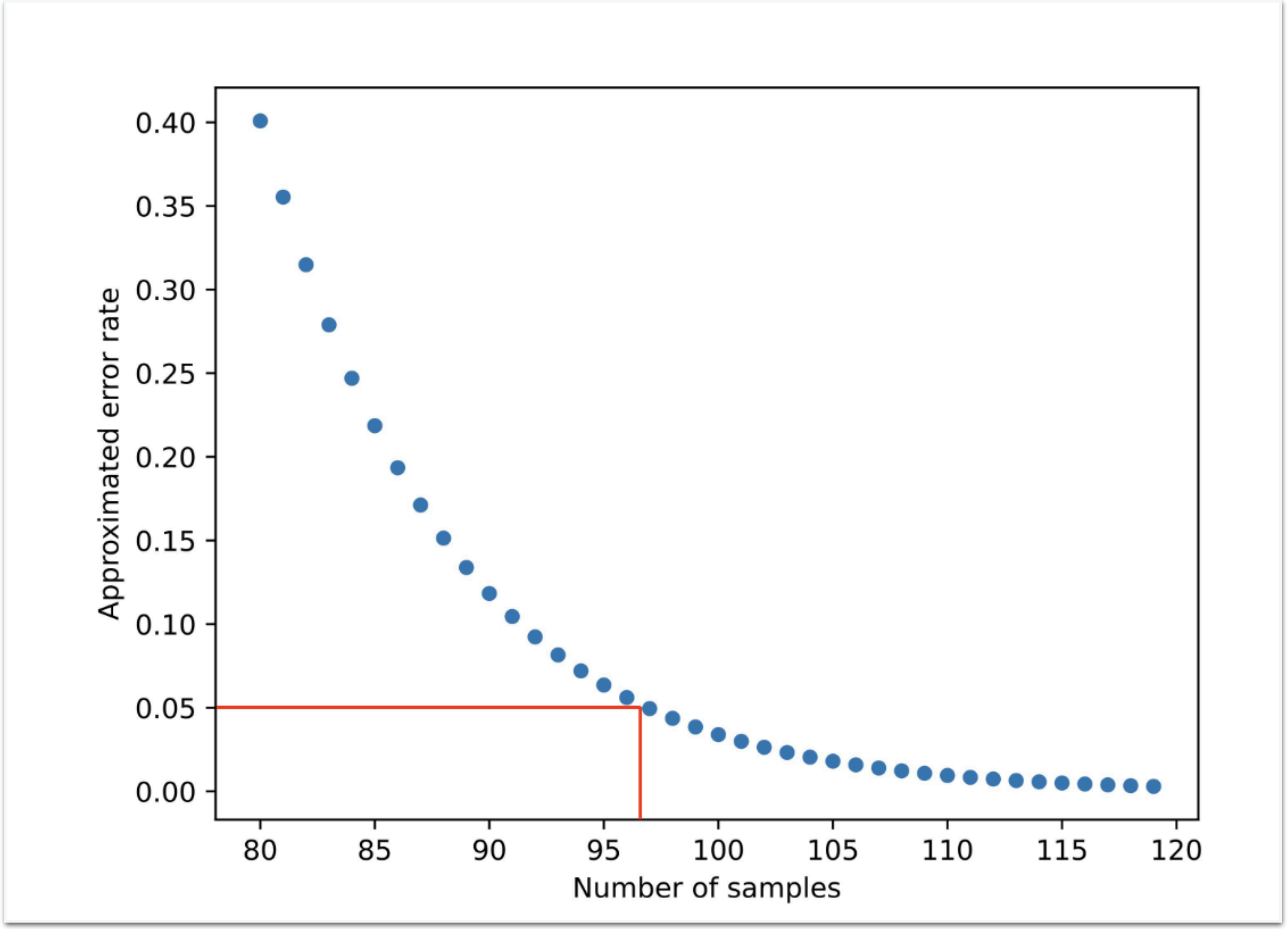

[1] Mardia, Jay, Jiantao Jiao, Ervin Tánczos, Robert D. Nowak, and Tsachy Weissman. "Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types." *Information and Inference: A Journal of the IMA* 9, no. 4 (2020): 813-850.

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.




[1] Mardia, Jay, Jiantao Jiao, Ervin Tánczos, 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

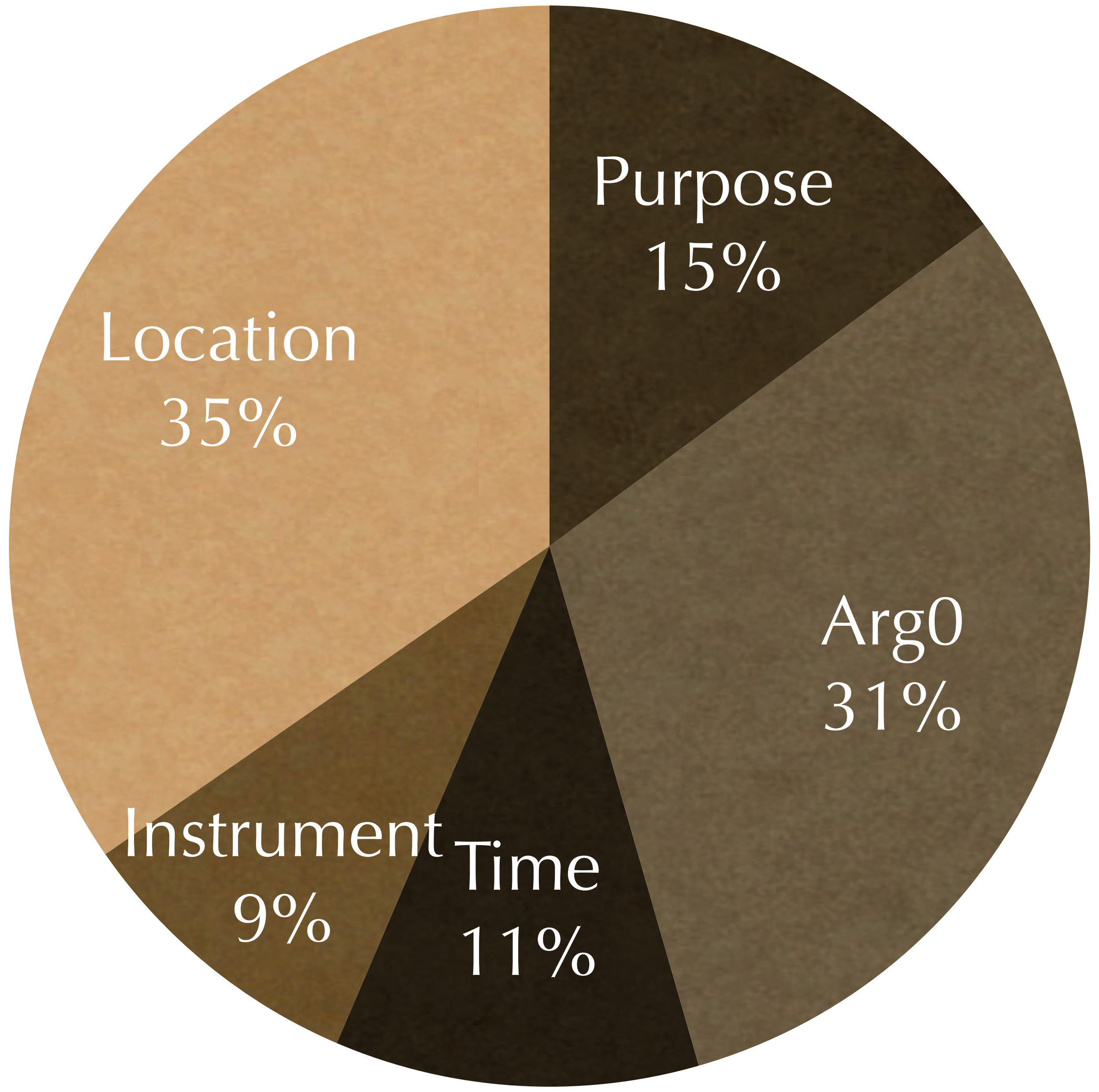
# CFC Data Statistics

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**:



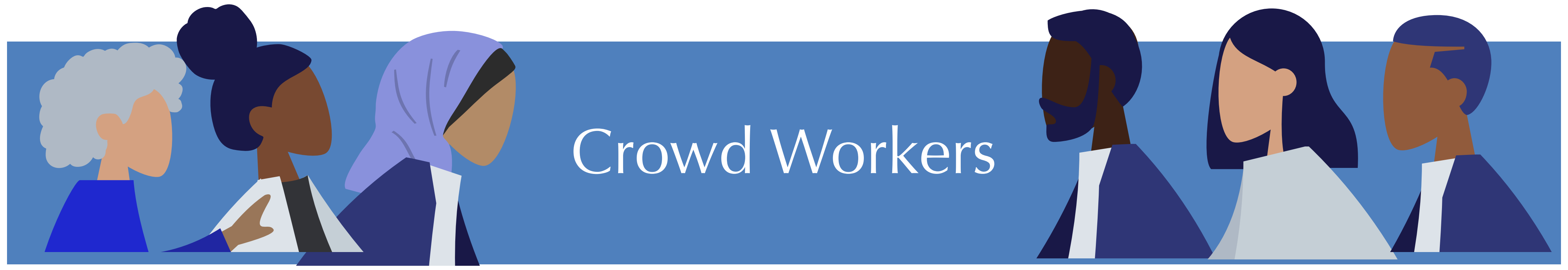

Question Slot Type

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 …

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

# CFC Probabilistic Evaluation

“They boiled the water” Purpose?



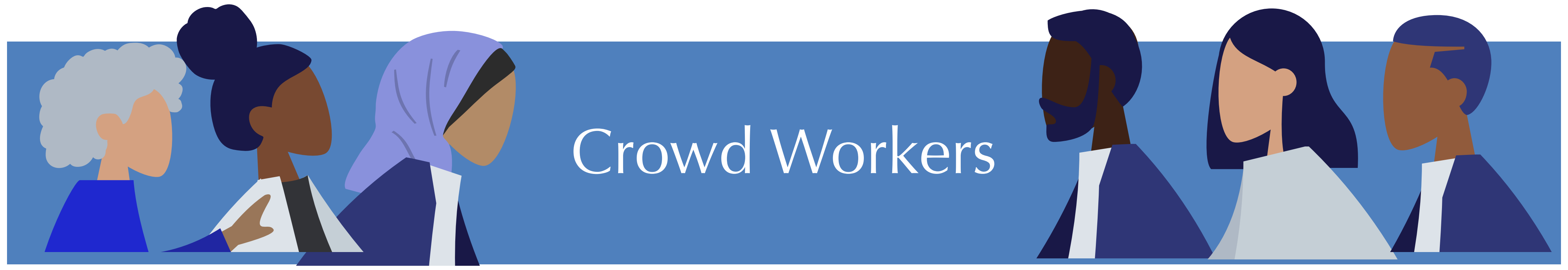


Qi, Boratko, Yelugam, O’Gorman, Singh, McCallum, Li. “Every Answer Matters: Evaluating Commonsense with Probabilistic Measures” ACL 2024

# CFC Probabilistic Evaluation

"They boiled the water" Purpose?

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



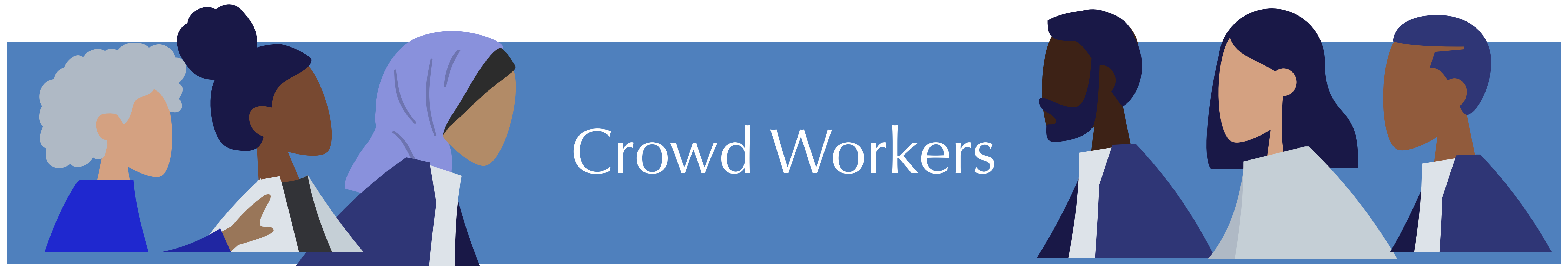


Qi, Boratko, Yelugam, O'Gorman, Singh, McCallum, Li. "Every Answer Matters: Evaluating Commonsense with Probabilistic Measures" ACL 2024

# CFC Probabilistic Evaluation

“They boiled the water” Purpose?

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




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


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

# CFC Probabilistic Evaluation

“They boiled the water” Cause?

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

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

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

# CFC Probabilistic Evaluation

“They boiled the water” Purpose?

for making tea
for a hot drink
make tea
cooking spaghetti
making pasta
to cook
steaming vegetables
boiling potatoes
cook
boiling chicken
cook food
making dinner
cooking
for an experiment

clean
disinfect
disinfecting
cleaning
cleaning tools
kill bacteria
purify
purification
make safe to drink
sterilization

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

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

# CFC Probabilistic Evaluation

"They boiled the water" Purpose?

for making tea
for a hot drink
make tea
cooking spaghetti
making pasta
to cook
steaming vegetables
boiling potatoes
cook
boiling chicken
cook food
making dinner
cooking
for an experiment

clean
disinfect
disinfecting
cleaning
cleaning tools
kill bacteria
purify
purification
make safe to drink
sterilization

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

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

# CFC Probabilistic Evaluation

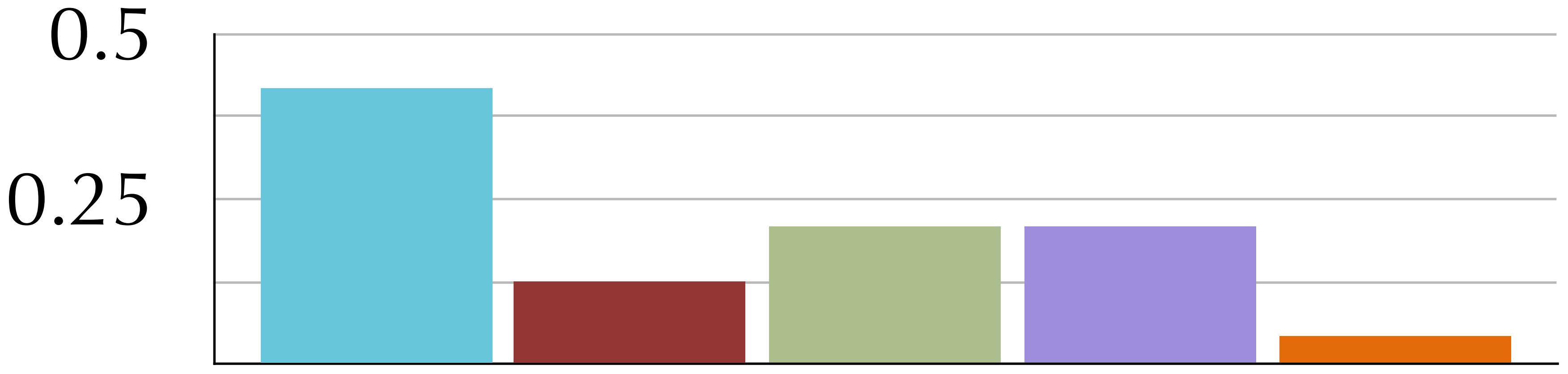
“They boiled the water” Purpose?

for making tea
for a hot drink
make tea
cooking spaghetti
making pasta
to cook
steaming vegetables
boiling potatoes
cook
boiling chicken
cook food
making dinner
cooking
for an experiment

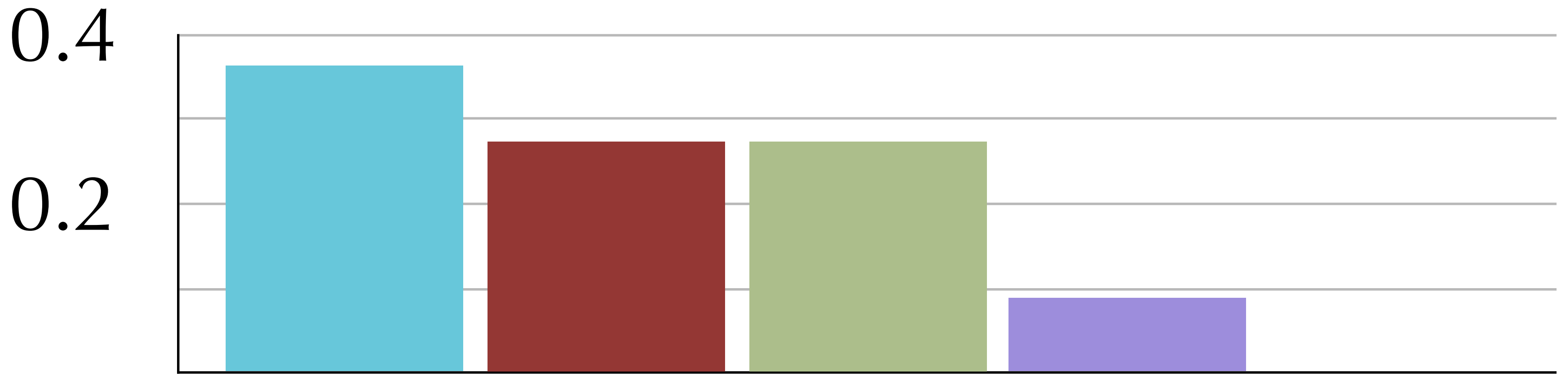
clean
disinfect
disinfecting
cleaning
cleaning tools
kill bacteria
purify
purification
make safe to drink
sterilization

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

P =




Q =



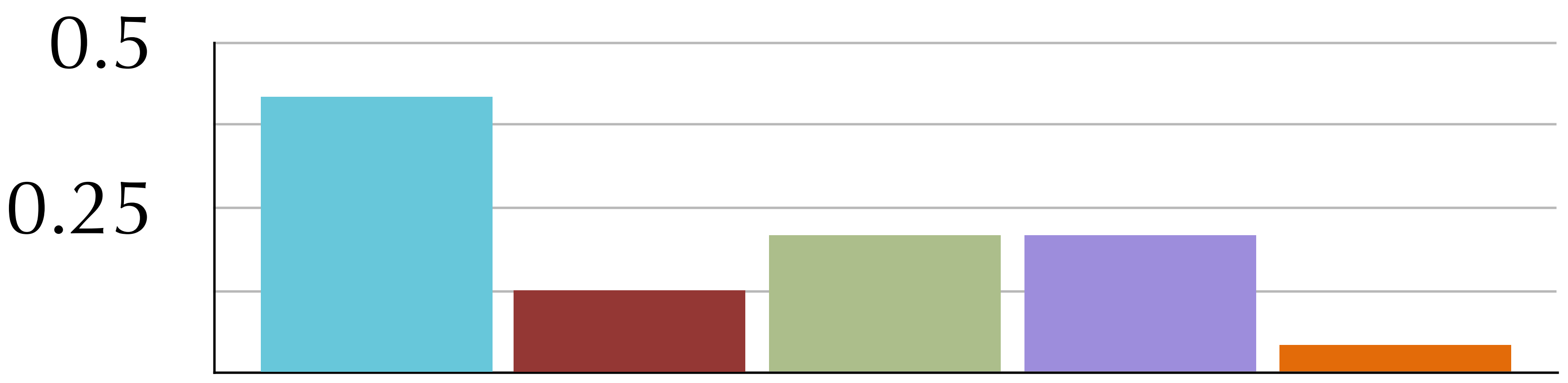
# CFC Probabilistic Evaluation

“They boiled the water” Purpose?

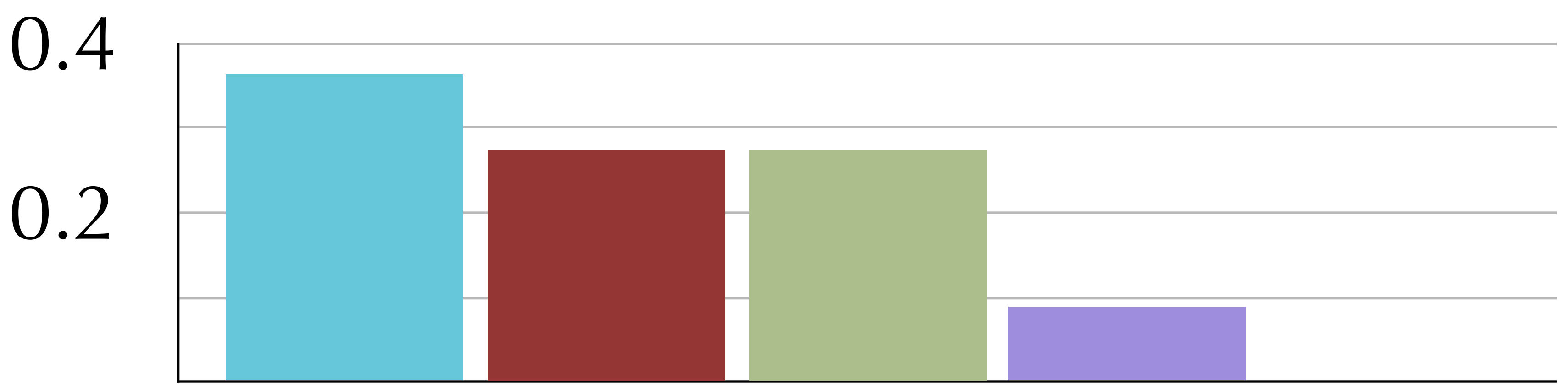
for making tea
for a hot drink
make tea
cooking spaghetti
making pasta
to cook
steaming vegetables
boiling potatoes
cook
boiling chicken
cook food
making dinner
cooking
for an experiment

clean
disinfect
disinfecting
cleaning
cleaning tools
kill bacteria
purify
purification
make safe to drink
sterilization

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




KL (P || Q)



# CFC Automatic Evaluation

*For each question:*

*G ← ground-truth answers (crowd-sourced)*

*H ← evaluation answers (model)*

For each human scorer:

Cluster G

Match H to clusters of G

Calculate score

Score(G, H ) ← average of scores

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

# CFC Automatic Evaluation

*For each question:*

*G ← ground-truth answers (crowd-sourced)*

*H ← evaluation answers (model)*

For each human scorer:

Cluster G

Match H to clusters of G

Calculate score

Score(G, H ) ← 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**

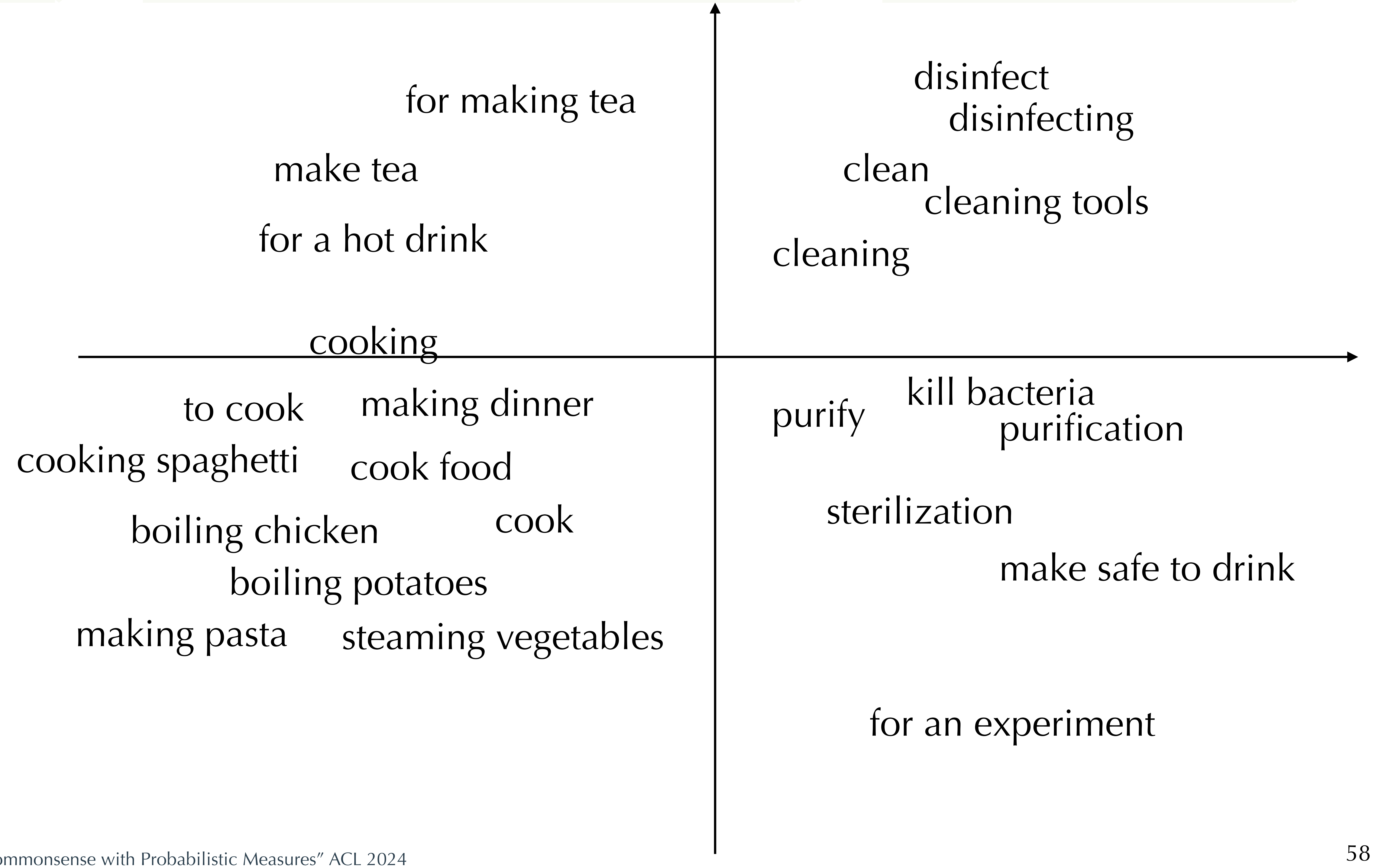
Qi, Boratko, Yelugam, O'Gorman, Singh, McCallum, Li. "Every Answer Matters: Evaluating Commonsense with Probabilistic Measures" ACL 2024

# CFC Automatic Evaluation

**Embed G** → Cluster G → Match H to cluster of G → Calculate Score

With Context
- BERT
- RoBERTa

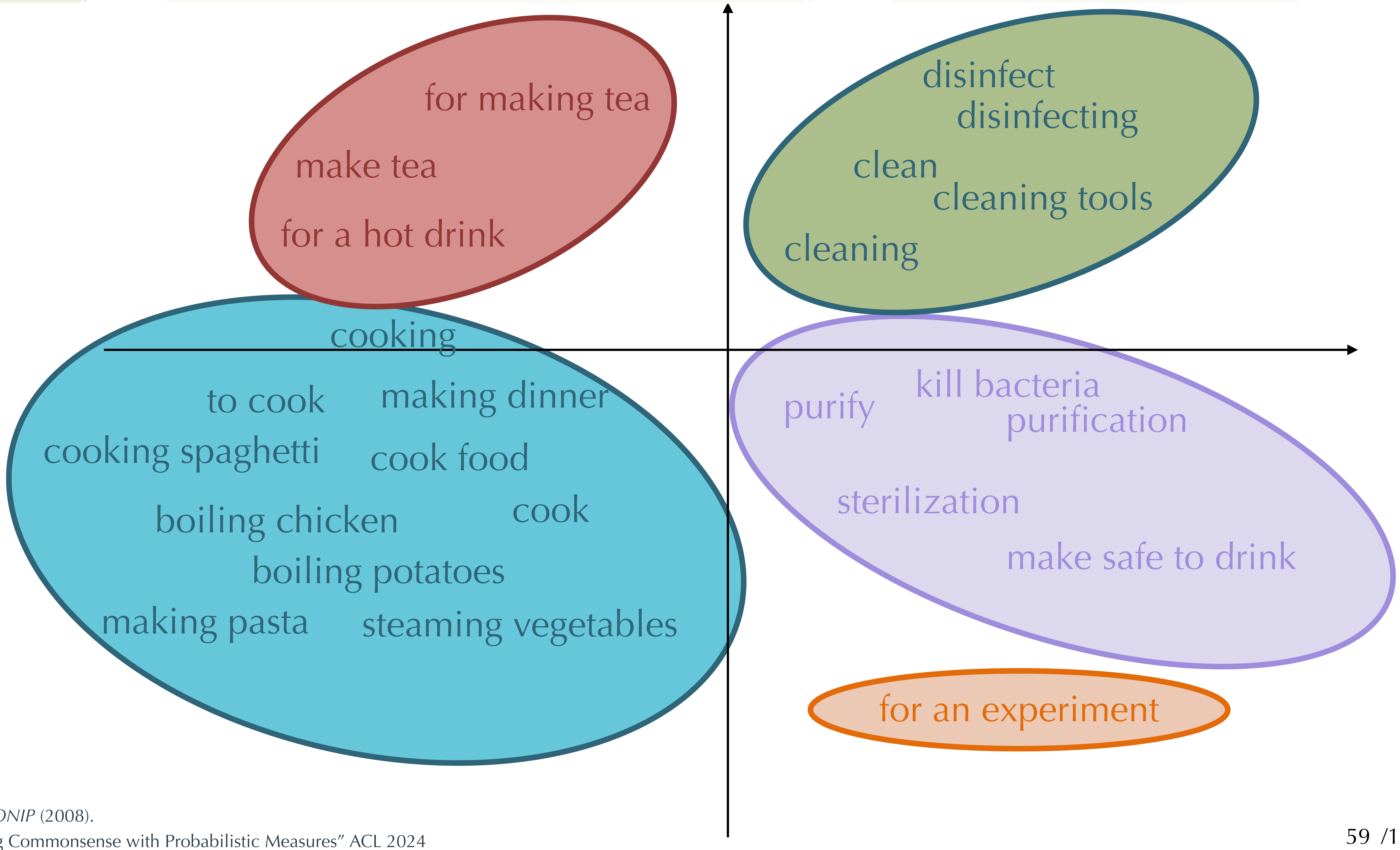
Without Context
- word2vec
- GloVe
- **FastText**



# 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**

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

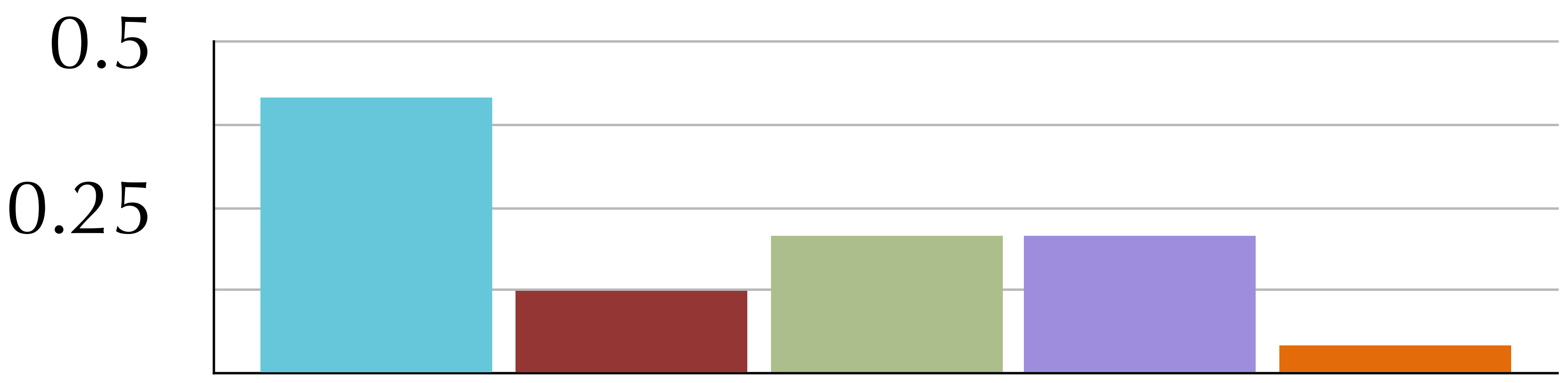
Embeddings Based
- FastText

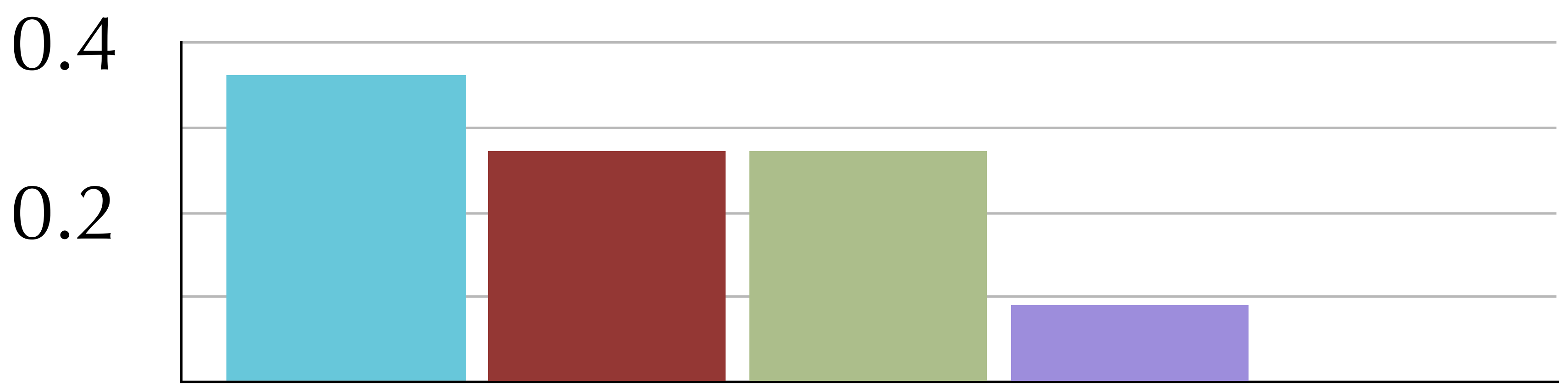
Lexical Token Based
- WordNet

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

# CFC Automatic Evaluation

Embed G → Cluster G → Match H to cluster of G → **Calculate Score**

P = 


Q = 


Score (G, H) = KL (P || 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**

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

# 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** |

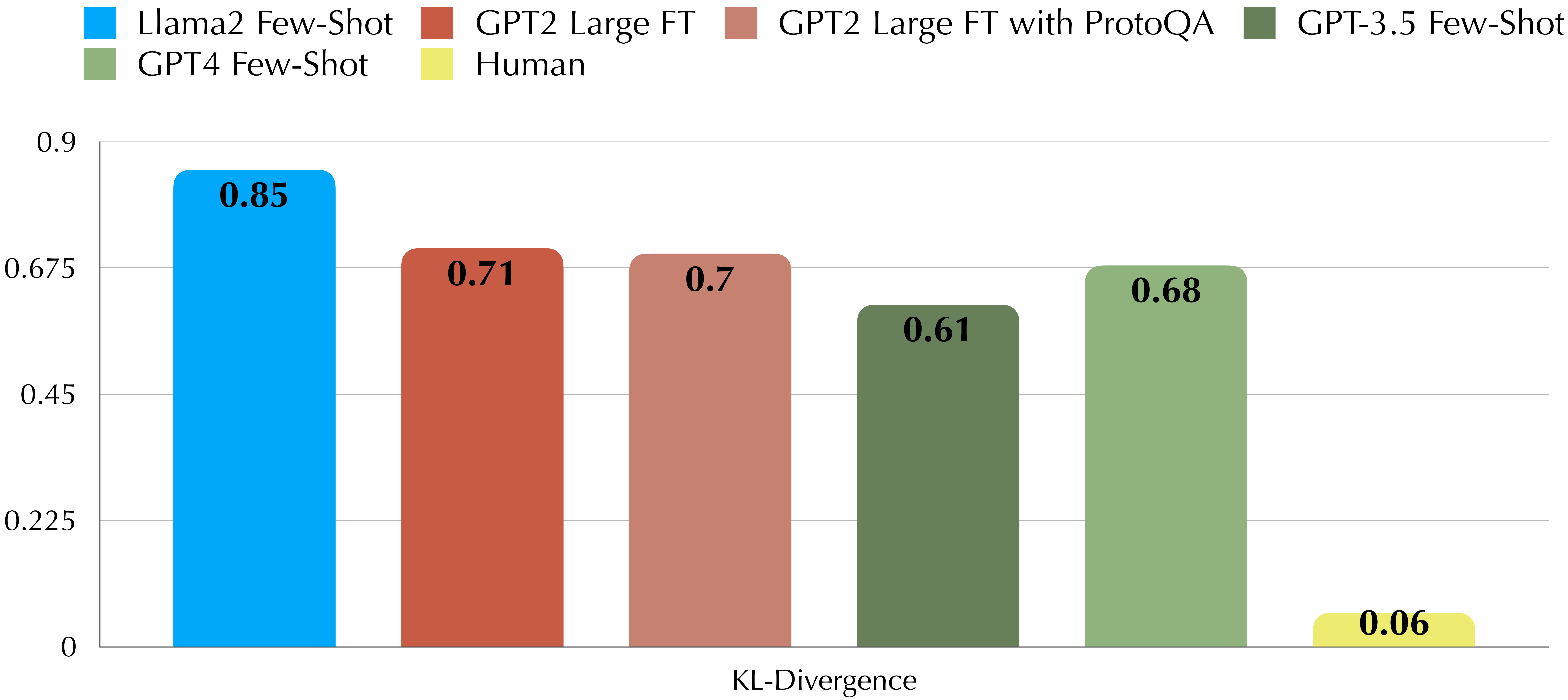
# Evaluating Automatic Metric - PROBEVAL

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

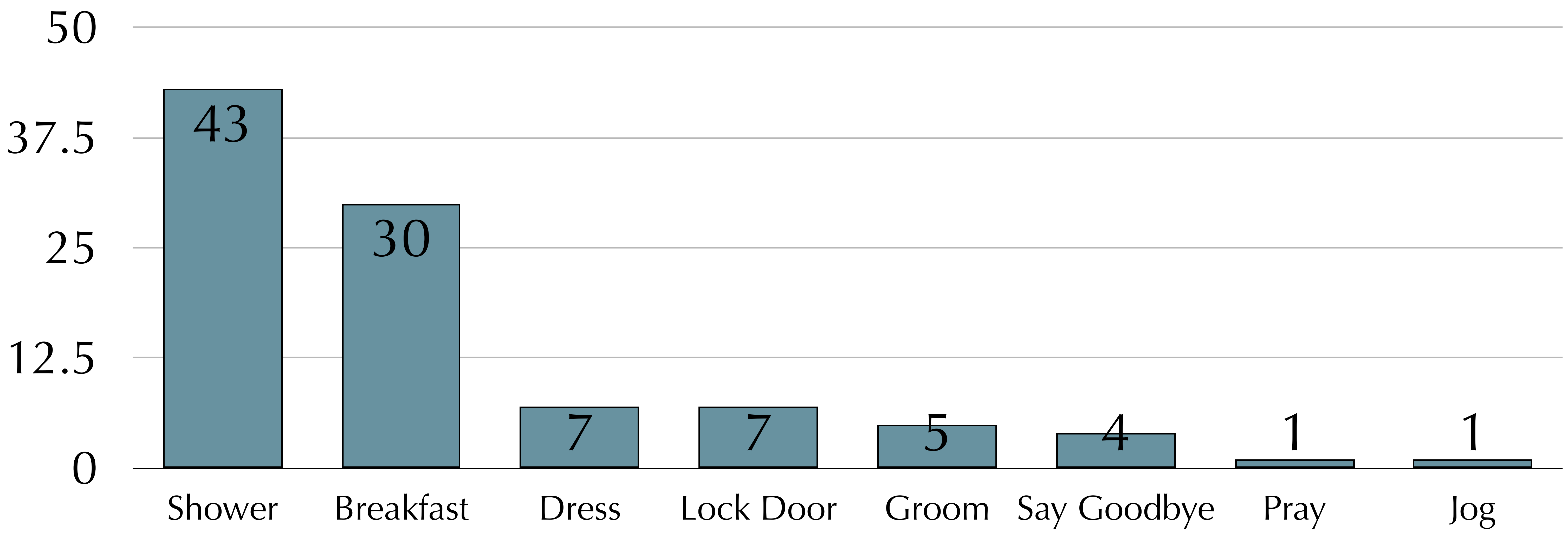

ProtoQA Evaluator

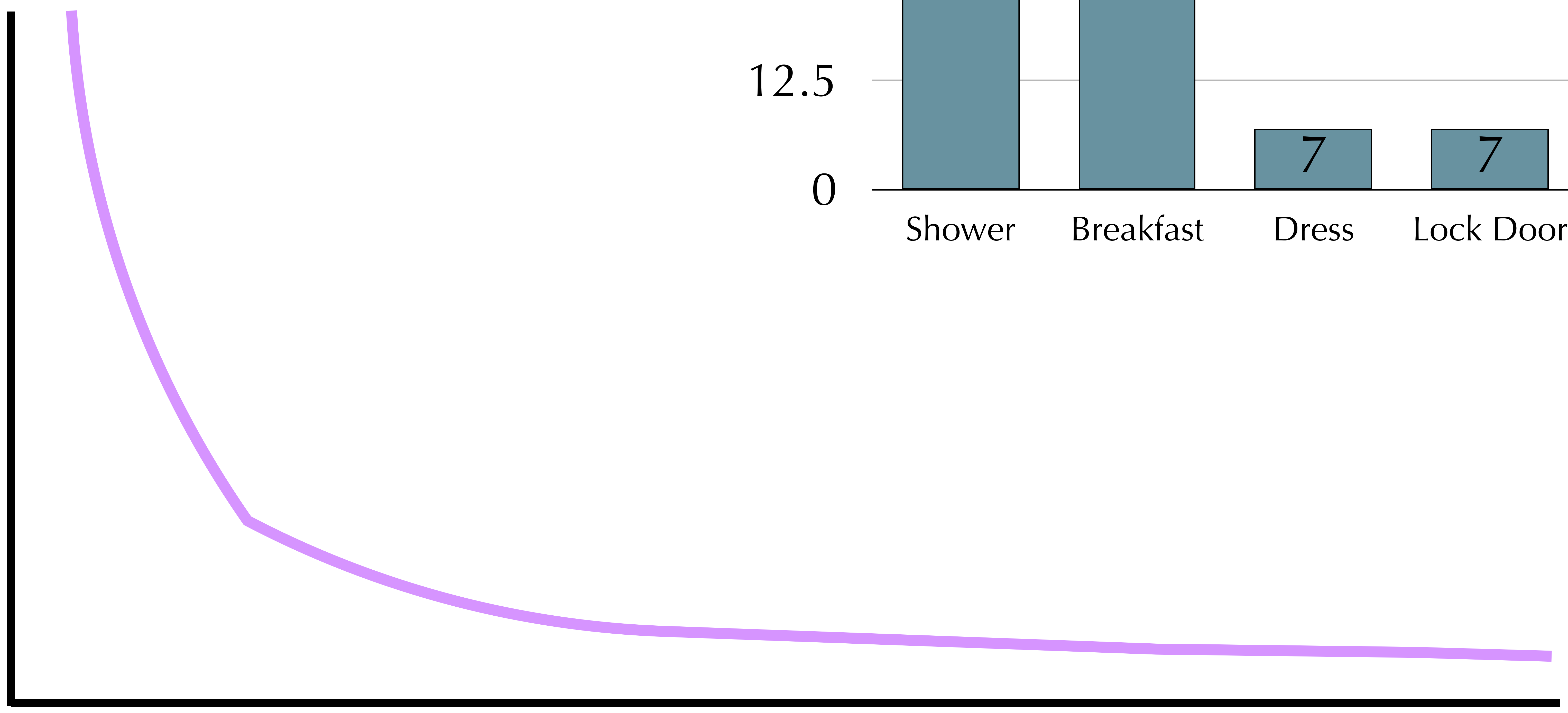
# Model Performance



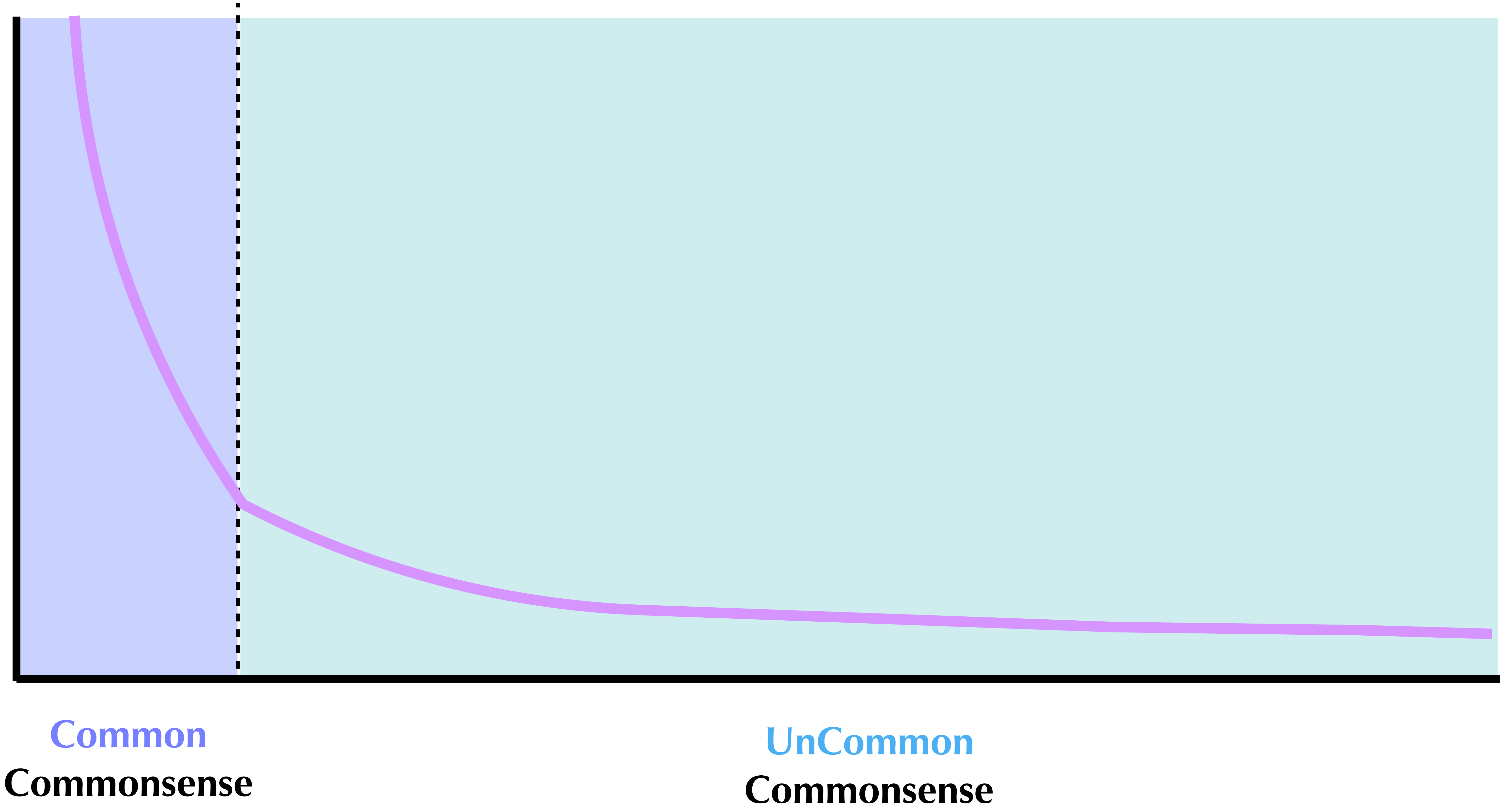
# Why is performance bad?

- 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 <u>avoid disappointing his partner</u>, 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 <u>avoid disappointing his partner</u>, who was excited about sharing…

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

Zhao, Chiu, Huang, Brahman, Hessel, Choudhury, Choi, **Li***, Suhr*. "UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations ." NAACL 2024.

# UNcommonsense Abductive Reasoning

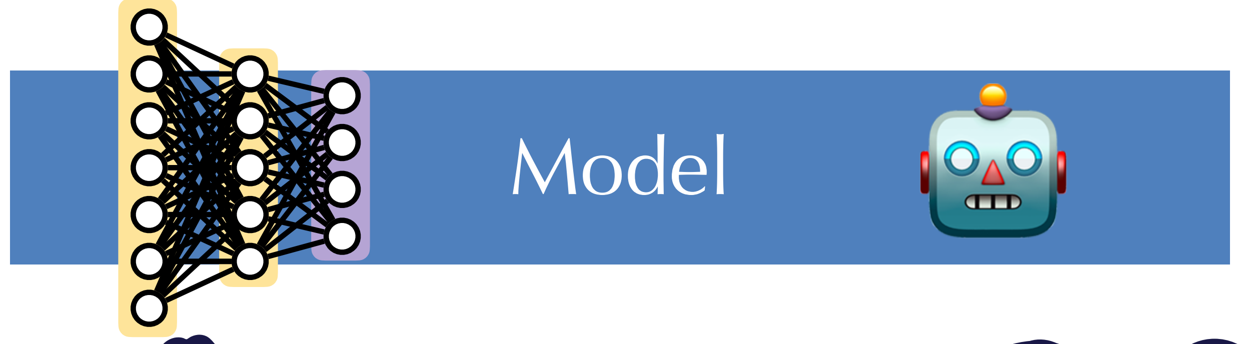
- Uncommon Outcomes
  - "Incorrect" answers from **SocialIQA** & **RocStories**
  - human written 


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

Zhao, Chiu, Huang, Brahman, Hessel, Choudhury, Choi, **Li***, Suhr*. "UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations ." NAACL 2024.

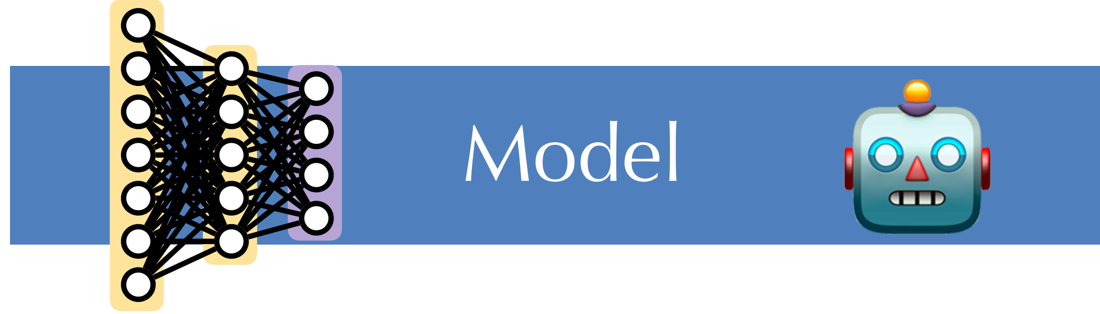
# UNcommonsense Abductive Reasoning

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

- Explanations for uncommon outcomes
  - LLM generated 

  - human written 

  - human written + LLM modification 

+ 


Zhao, Chiu, Huang, Brahman, Hessel, Choudhury, Choi, **Li***, Suhr*. "UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations ." NAACL 2024.

# UNcommonsense Abductive Reasoning

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

Zhao, Chiu, Huang, Brahman, Hessel, Choudhury, Choi, **Li***, Suhr*. “UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations .” NAACL 2024.

# UNcommonsense Abductive Reasoning

## Explanation Analysis: Quality

| | **un**-SocialIQA | | | **un**-RocStories | | |
|---|---|---|---|---|---|---|
| | *Crowd* | *C+LLM* | *LLM²* | *Crowd* | *C+LLM* | *LLM²* |
| 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

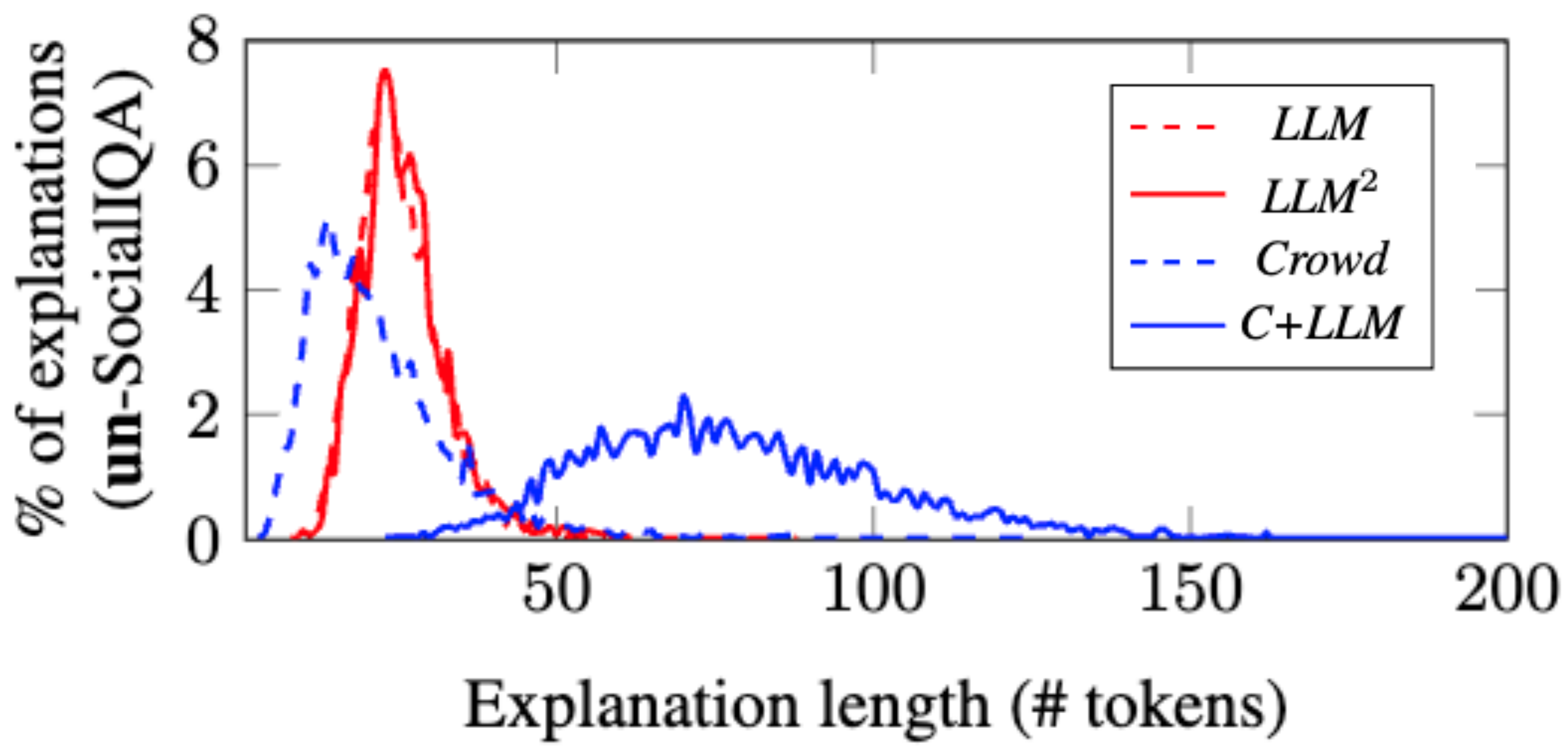



**Figure 2: Distribution of explanation lengths in un-SocialIQA.** Computed on the development sets.

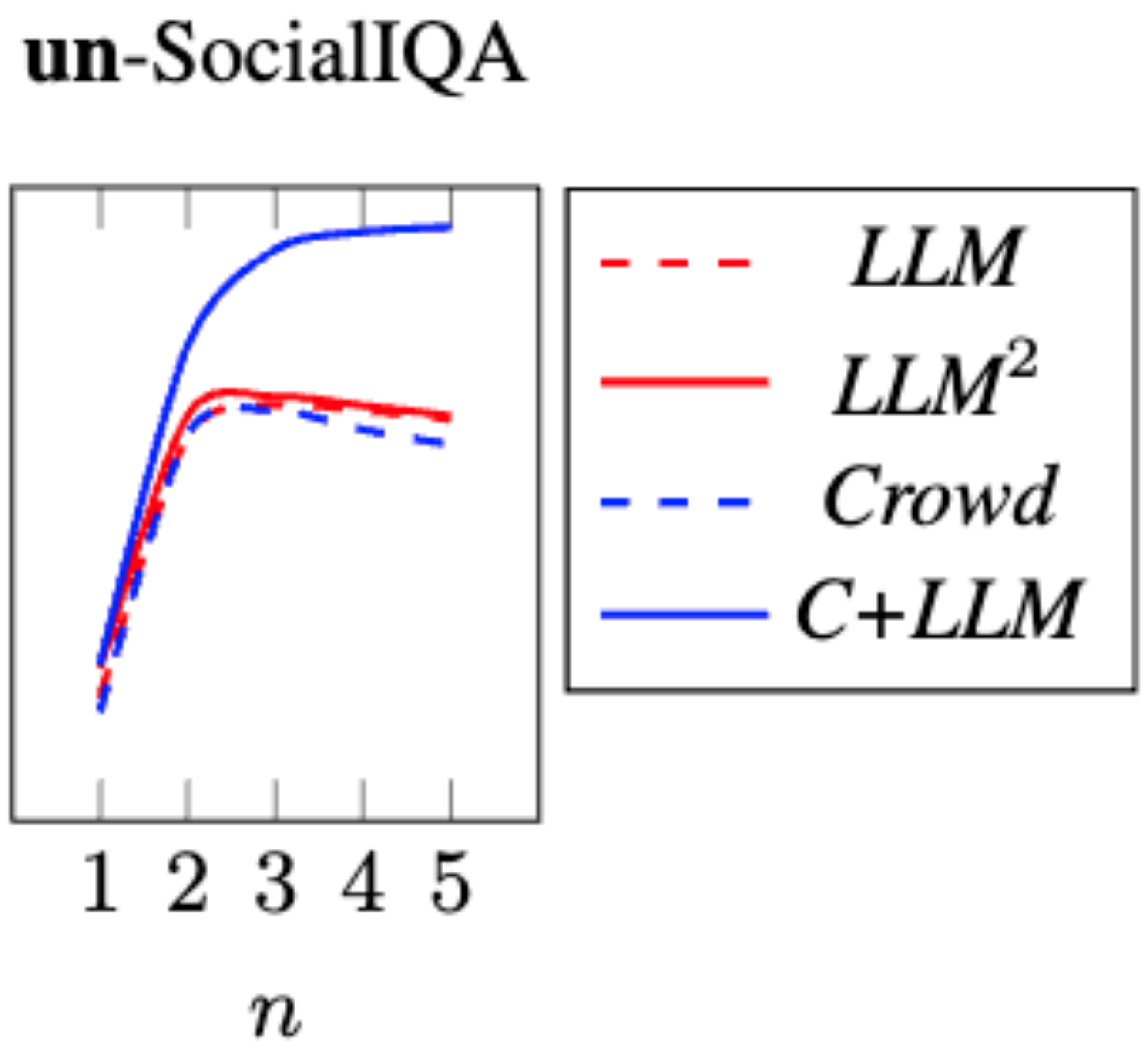



**Figure 3: Entropies of n-gram distributions in un-SocialIQA.** Computed on the development sets.

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

- 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

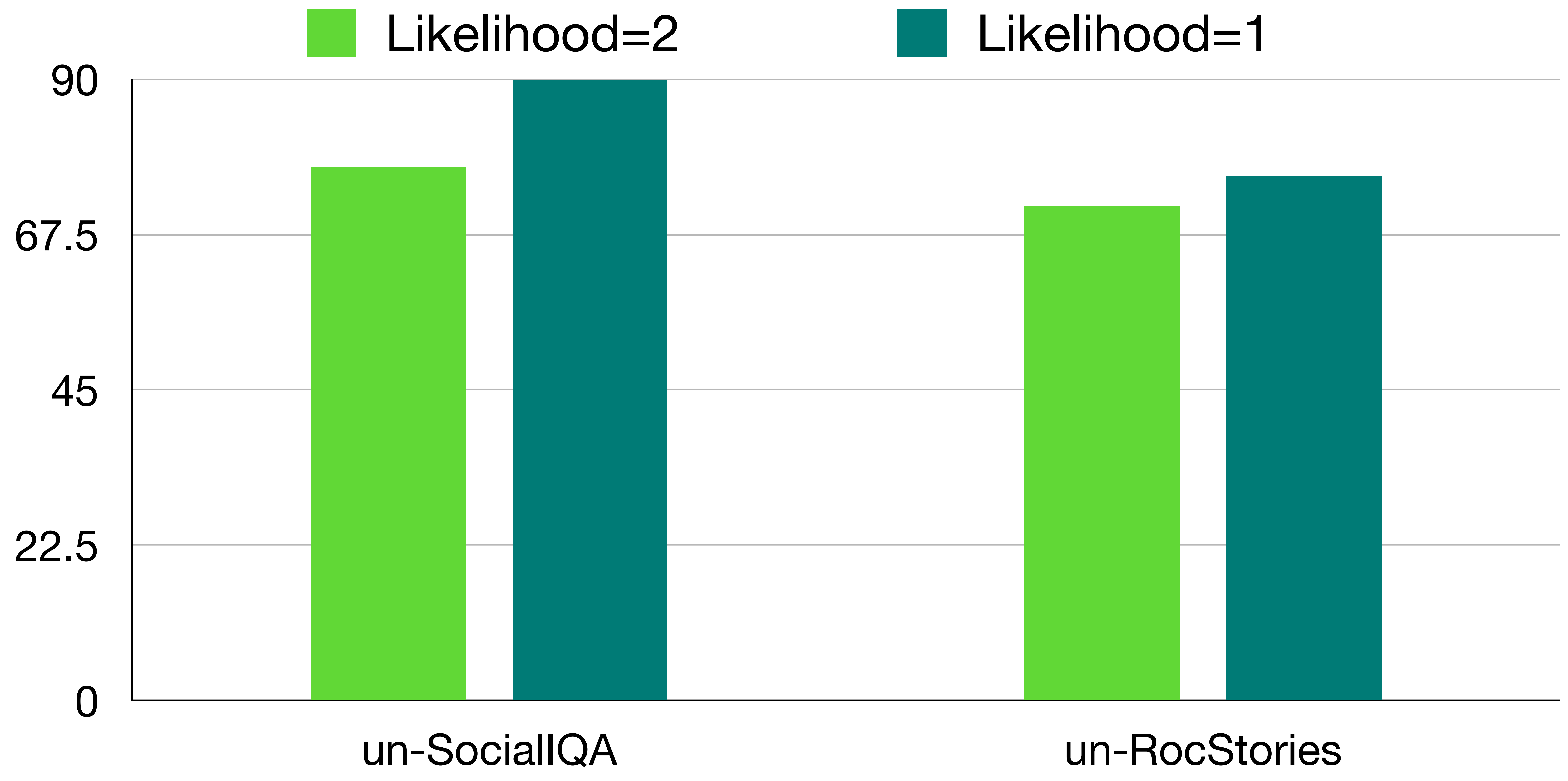



**Figure 1: Non-lose rates of Human+*LLM* versus LLM,** broken down by the likehoods of outcomes. Likehood=1 is lease likely. (annoated by human)

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

Zhao, Chiu, Huang, Brahman, Hessel, Choudhury, Choi, **Li***, Suhr*. "UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations ." NAACL 2024.

# 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**.

Dziri*, Lu*, Sclar*, **Li****, Jiang**, Lin**, West, Bhagavatula, Bras, Hwang, Sanyal, Welleck, Ren, Ettinger, Harchaoui, Choi. ``Fate and Faith: Limits of Transformers on Compositionality" NeurIPS 2023.

# 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

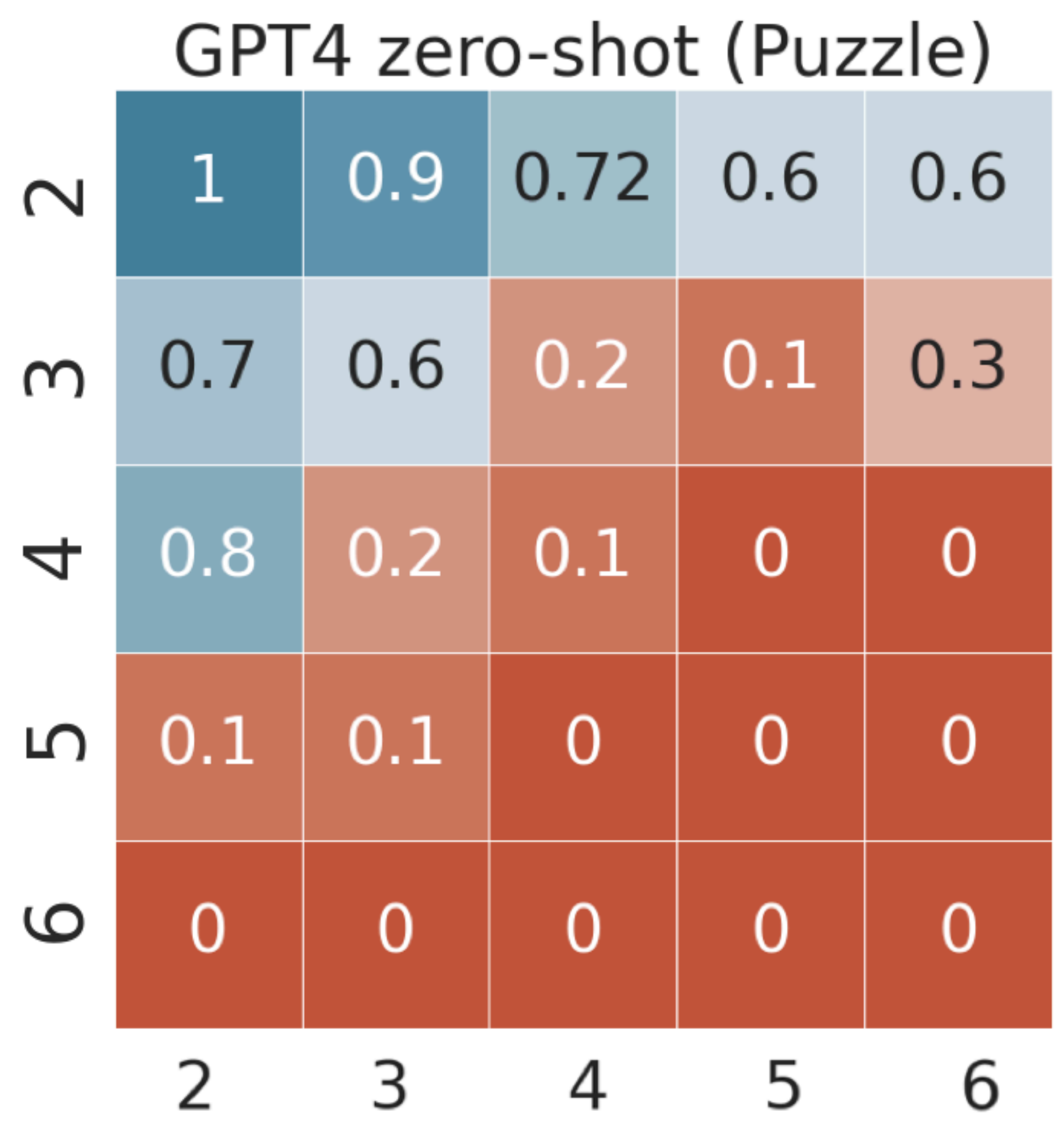



**Figure 1: Zero-shot accuracy.** Axes refer to problem sizes, number of houses and attributes in puzzle.

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

Dziri*, Lu*, Sclar*, **Li****, Jiang**, Lin**, West, Bhagavatula, Bras, Hwang, Sanyal, Welleck, Ren, Ettinger, Harchaoui, Choi. ``Fate and Faith: Limits of Transformers on Compositionality" NeurIPS 2023.

# 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.

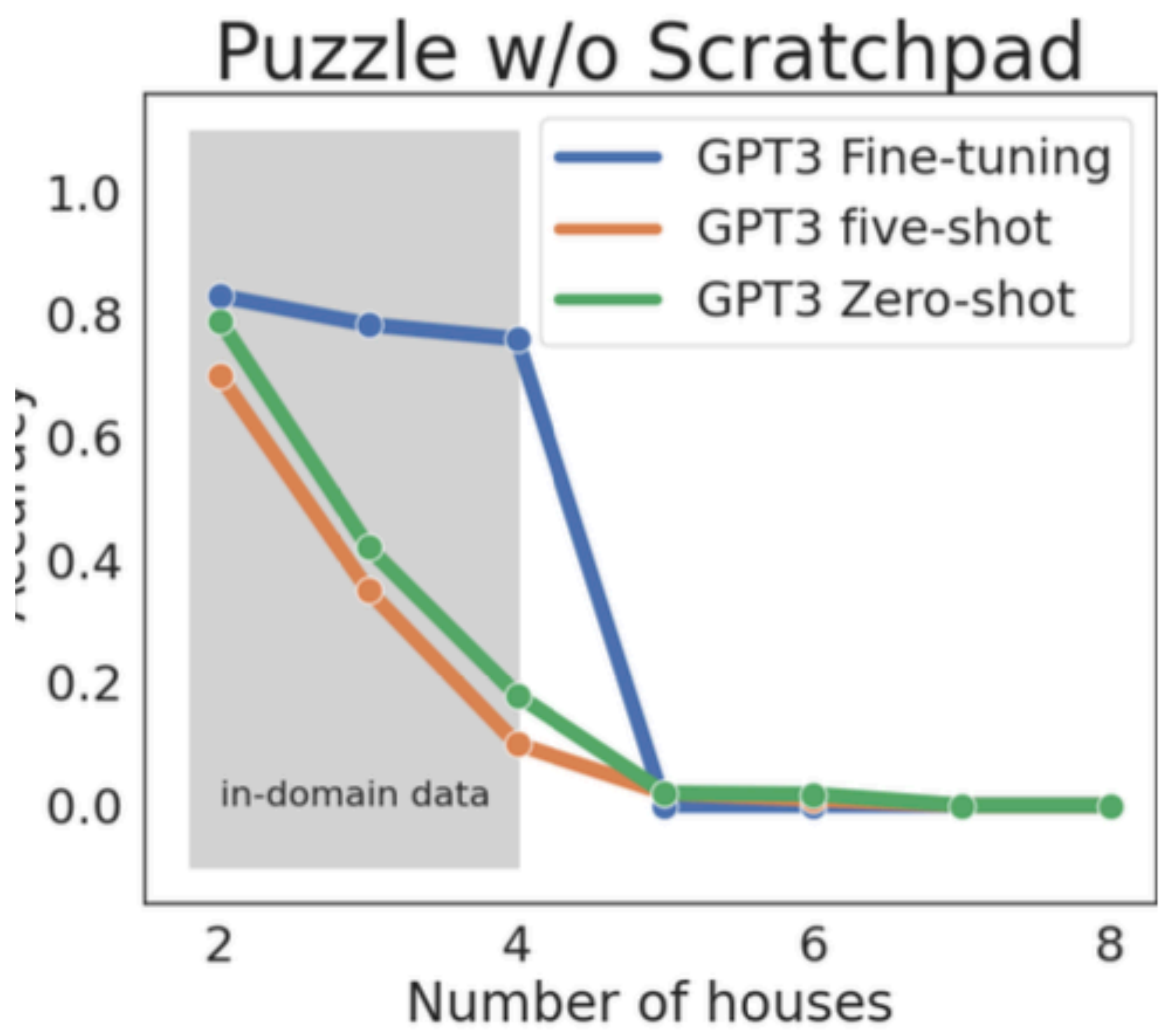



**Figure 3:** fine-tuning performance with in-domain data and out-of-domain data.

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

Dziri*, Lu*, Sclar*, **Li****, Jiang**, Lin**, West, Bhagavatula, Bras, Hwang, Sanyal, Welleck, Ren, Ettinger, Harchaoui, Choi. ``Fate and Faith: Limits of Transformers on Compositionality" NeurIPS 2023.

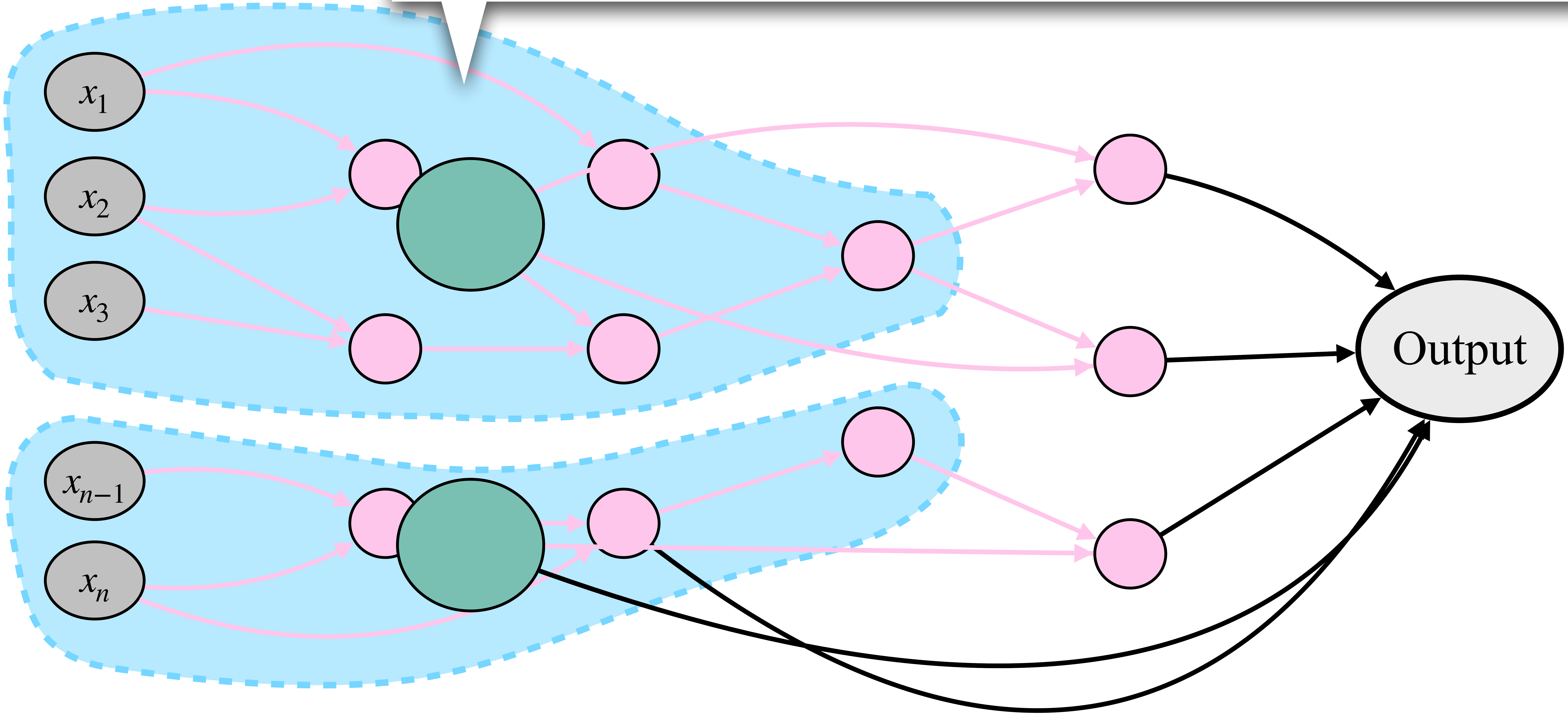
# One of the key findings.

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

Dziri*, Lu*, Sclar*, **Li****, Jiang**, Lin**, West, Bhagavatula, Bras, Hwang, Sanyal, Welleck, Ren, Ettinger, Harchaoui, Choi. ``Fate and Faith: Limits of Transformers on Compositionality" NeurIPS 2023.

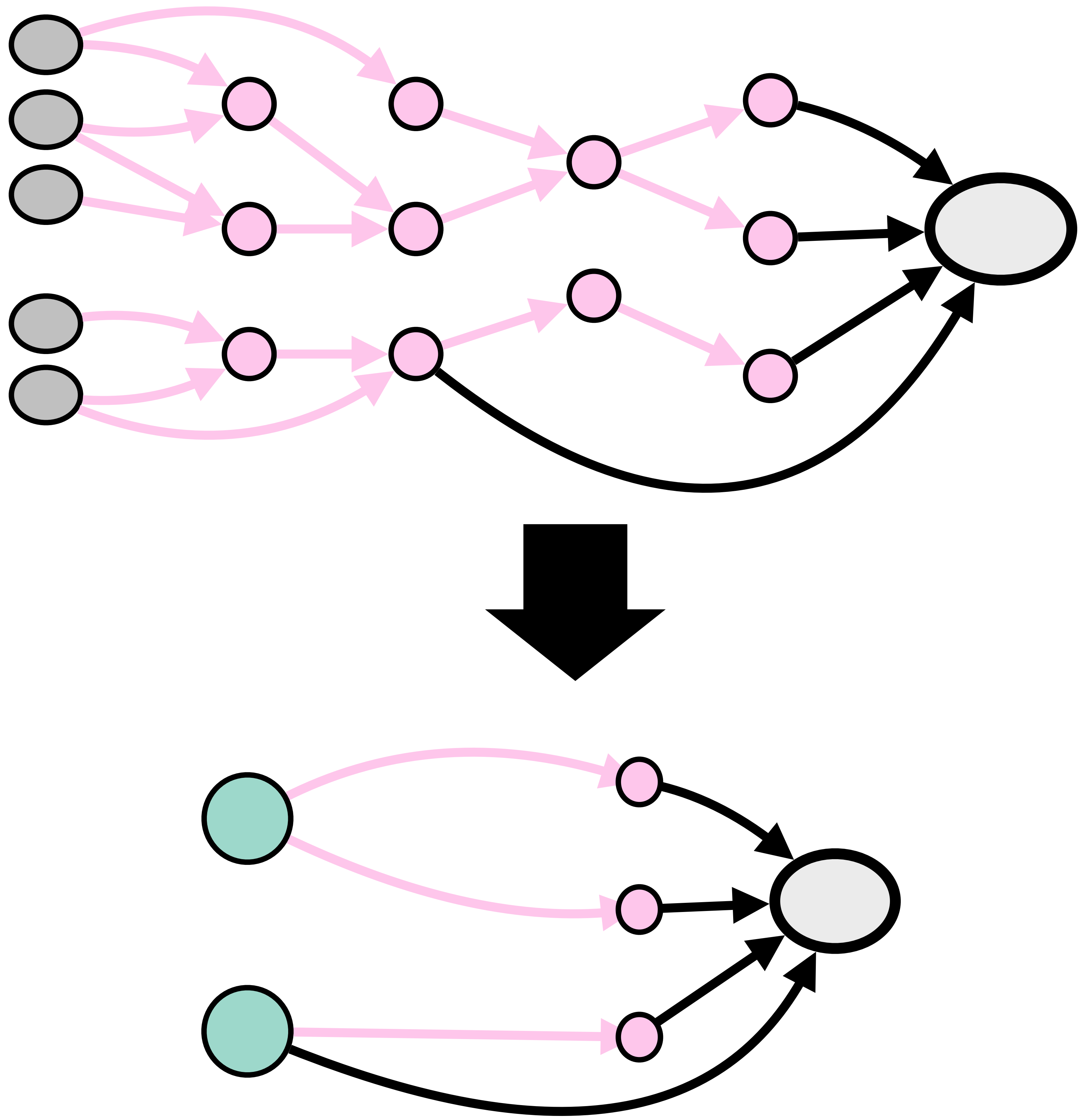
*Detect subgraphs already seen during training:*

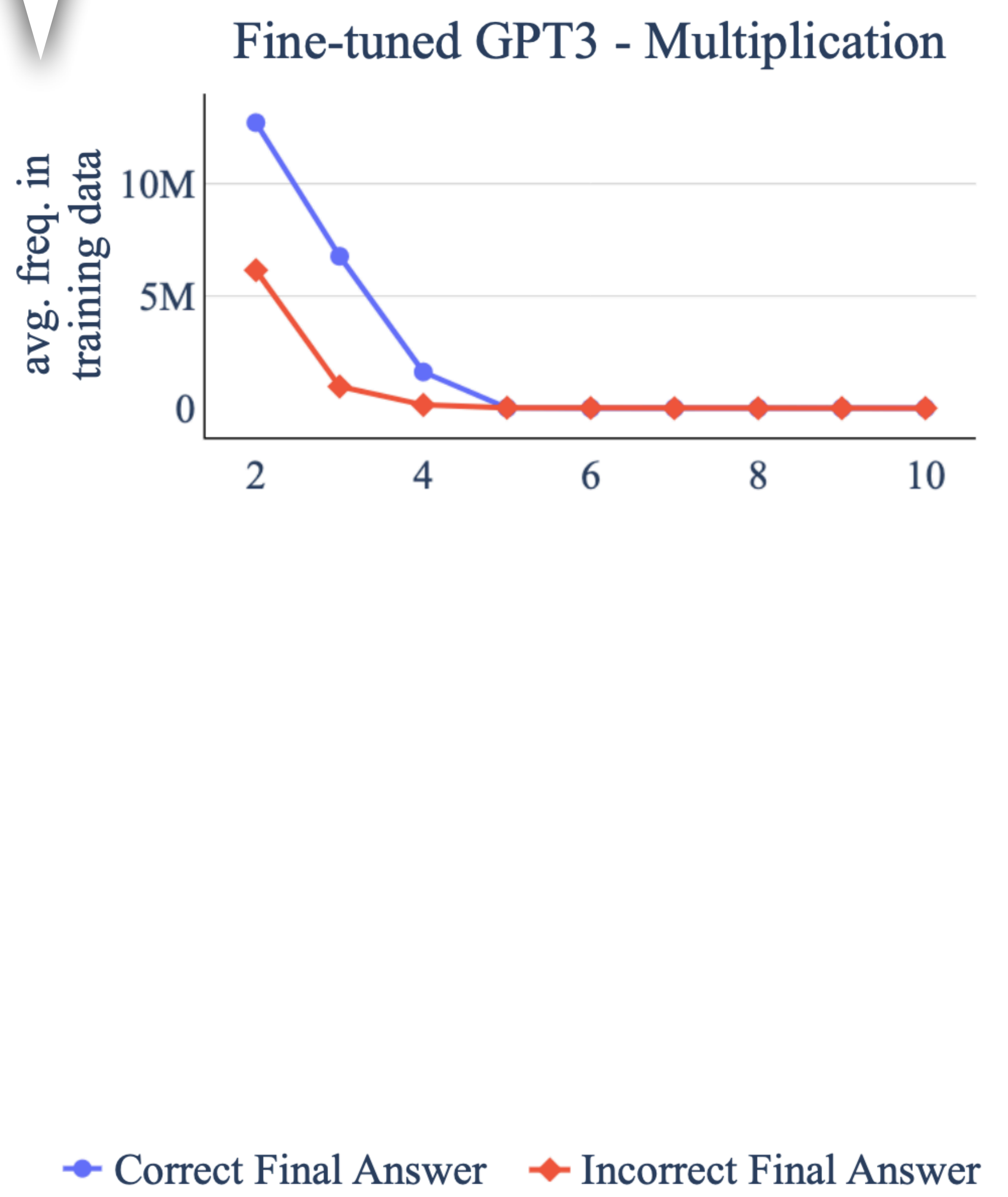
*…ant subgraphs during training, the inference is only seemingly highly compositional*




Dziri*, Lu*, Sclar*, **Li****, Jiang**, Lin**, West, Bhagavatula, Bras, Hwang, Sanyal, Welleck, Ren, Ettinger, Harchaoui, Choi. ``Fate and Faith: Limits of Transformers on Compositionality" NeurIPS 2023.

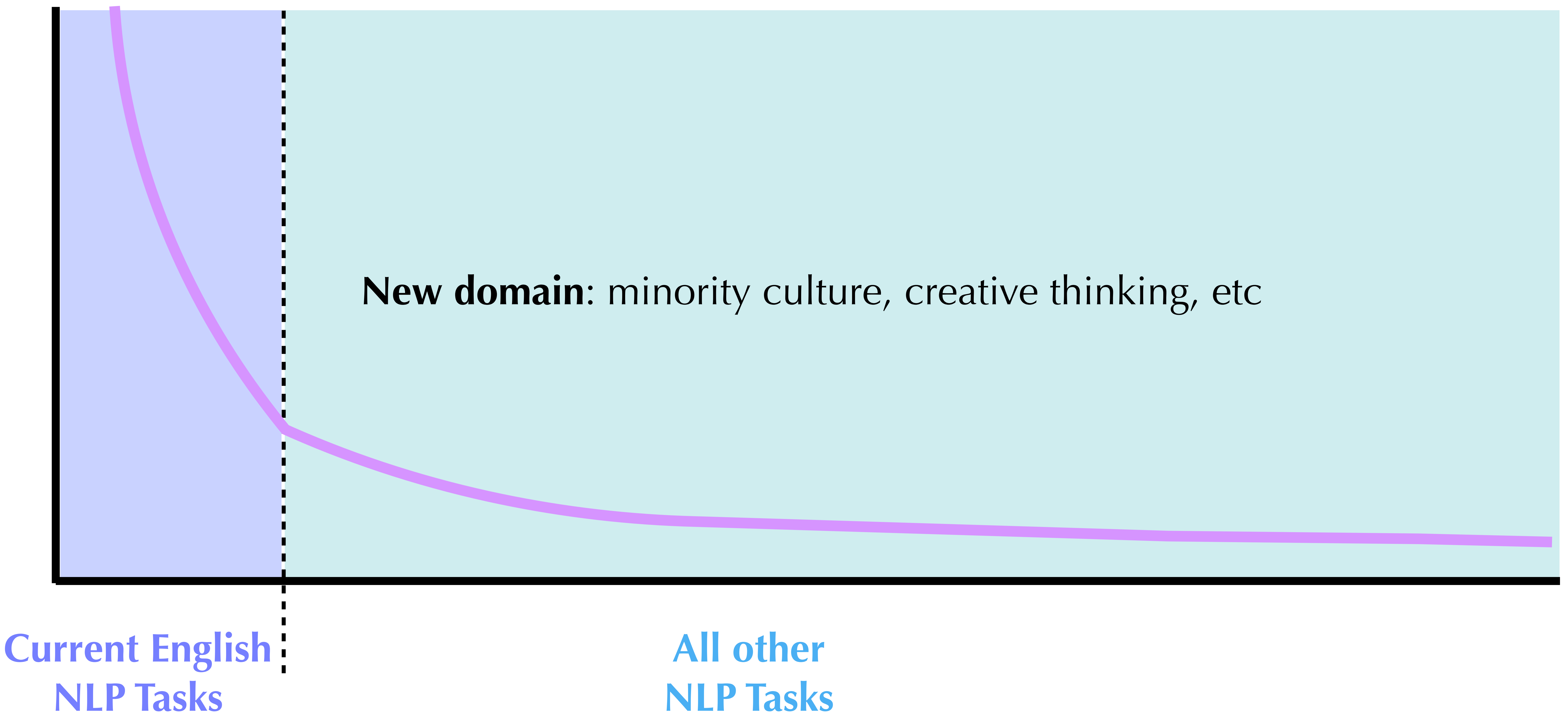
*Transformers' successes are heavily linked to having seen significant portions of the required computation graph during training*



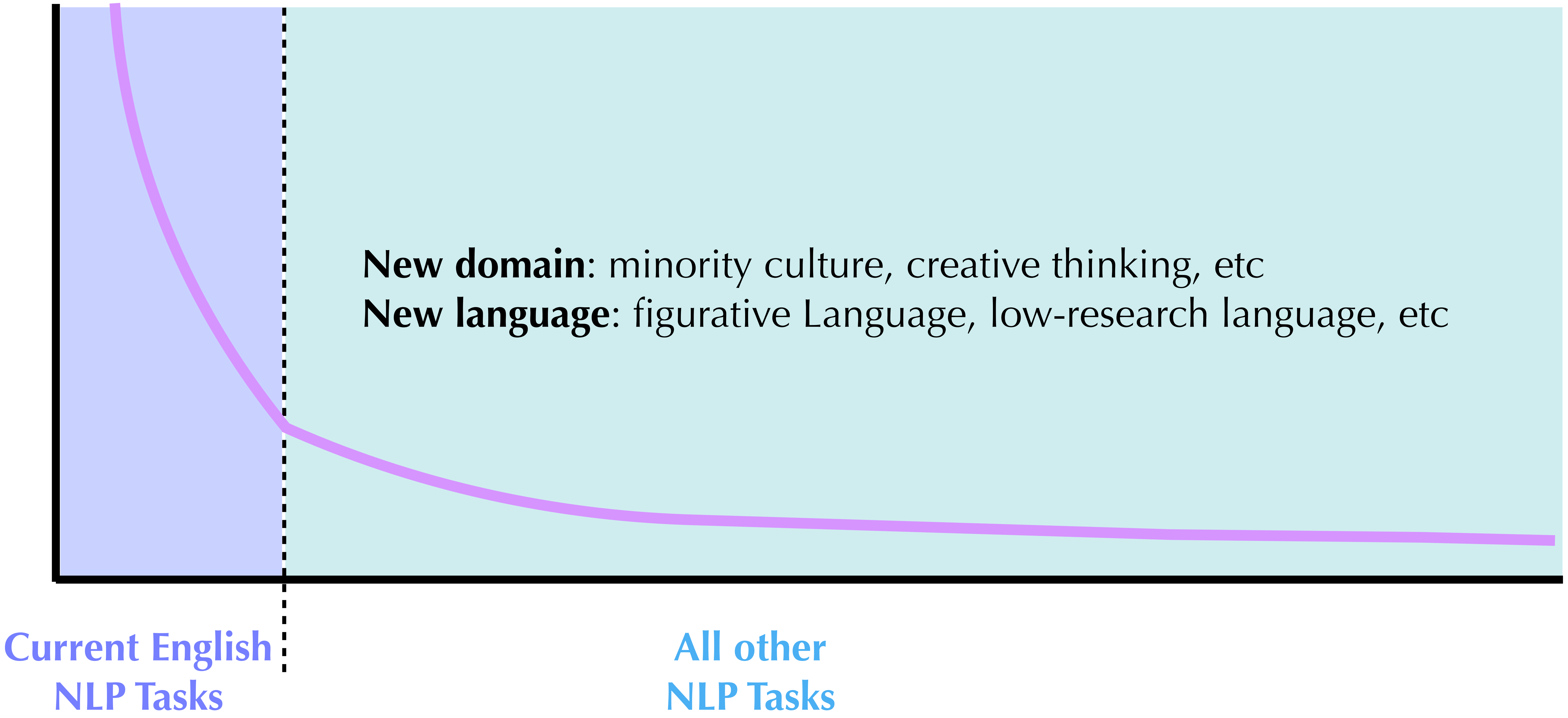



Dziri*, Lu*, Sclar*, **Li****, Jiang**, Lin**, West, Bhagavatula, Bras, Hwang, Sanyal, Welleck, Ren, Ettinger, Harchaoui, Choi. ``Fate and Faith: Limits of Transformers on Compositionality" NeurIPS 2023.

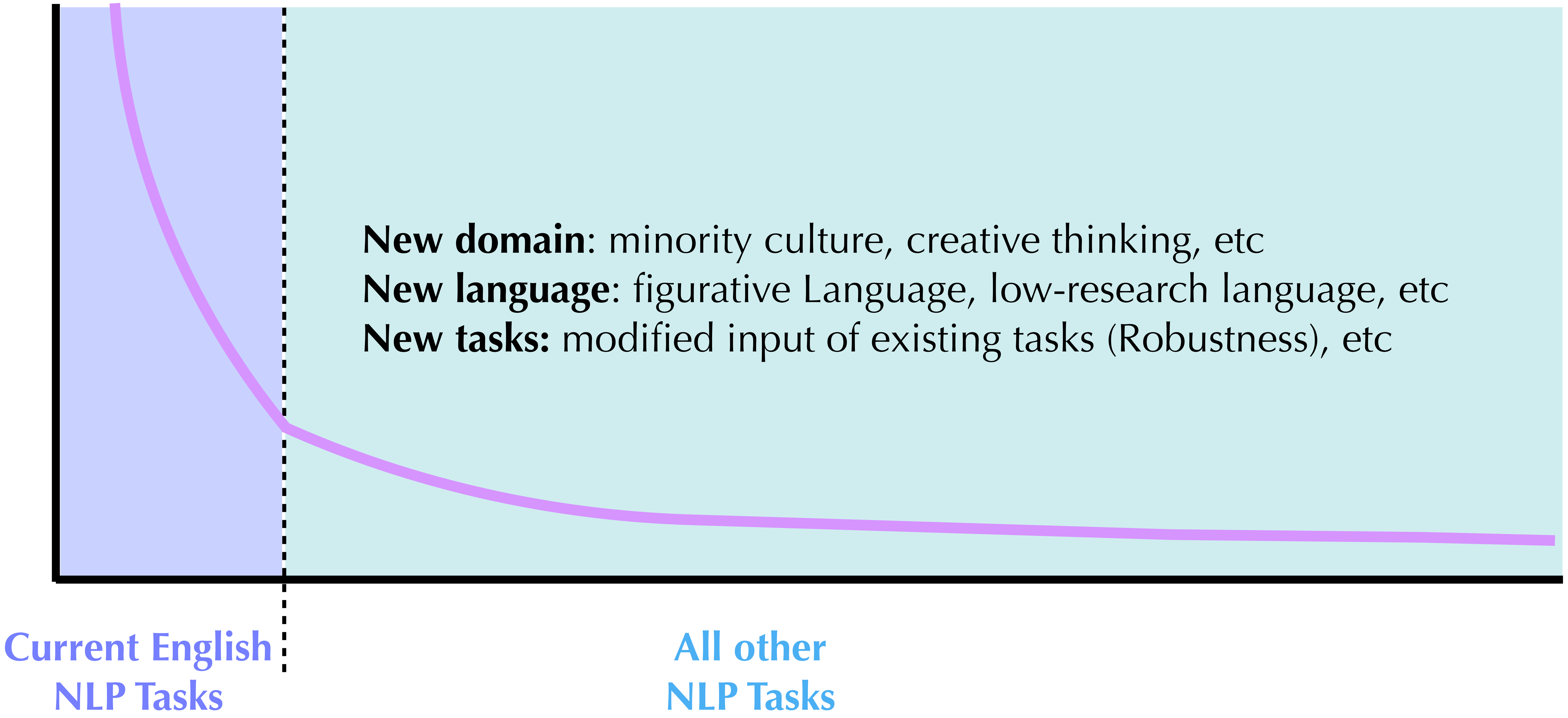
# Takeaways: evaluation



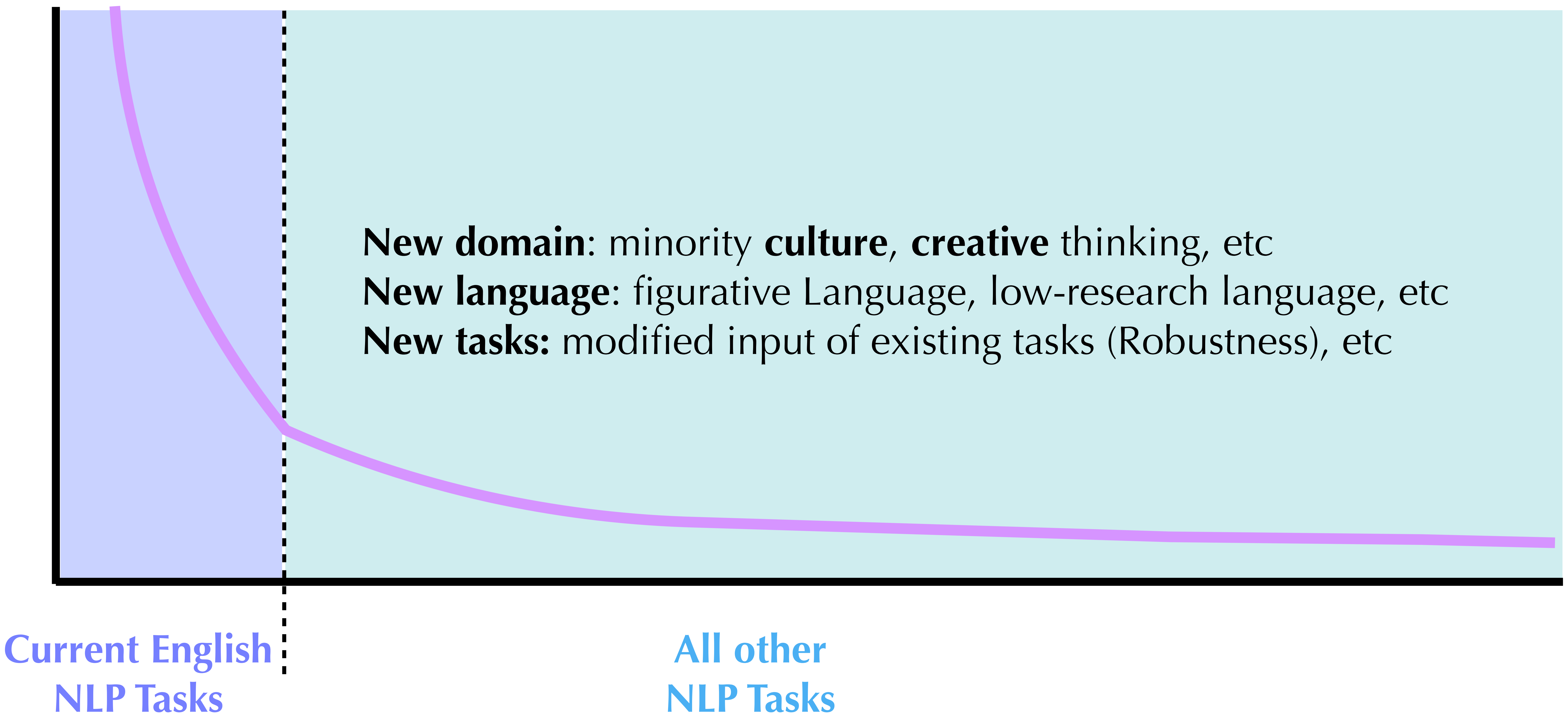
# Takeaways: evaluation



# Takeaways: evaluation



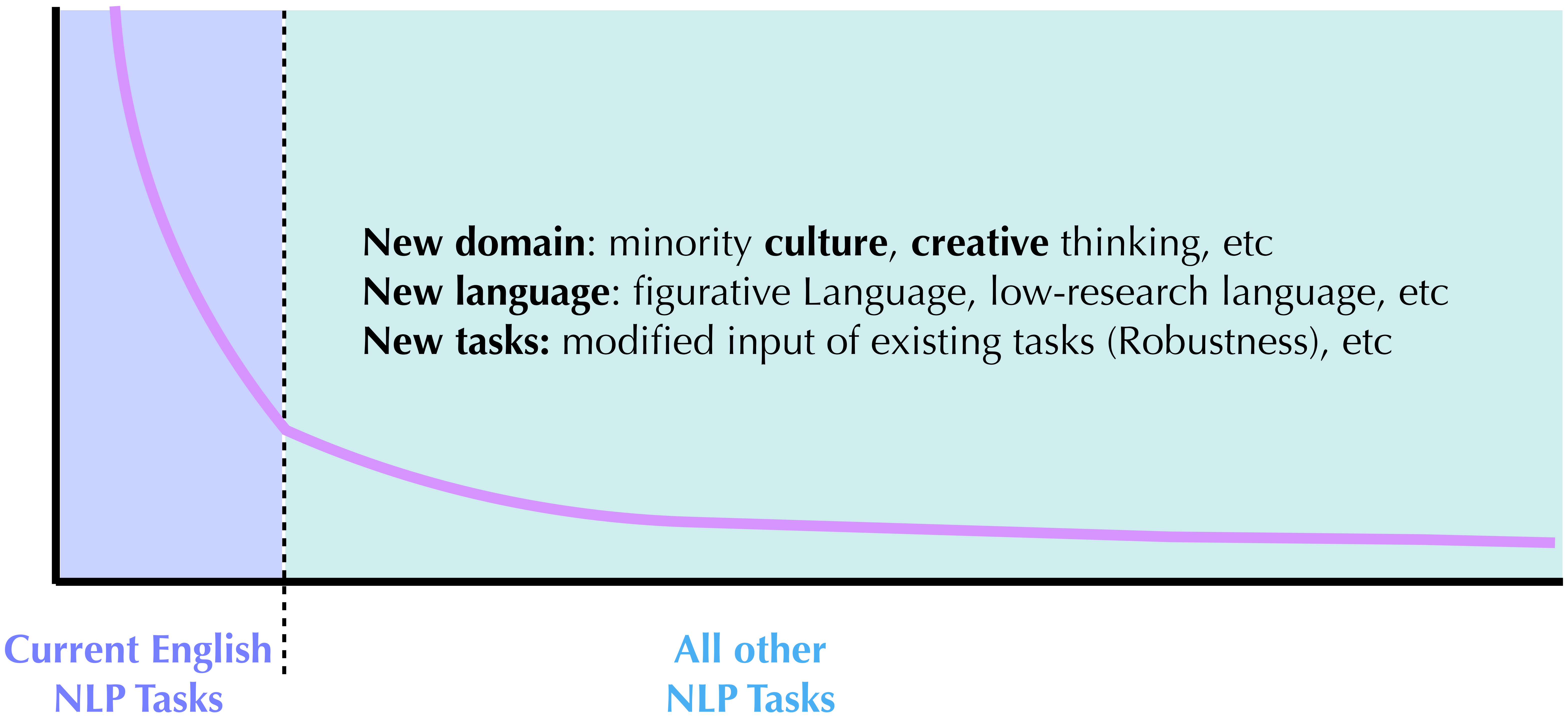
# Takeaways: evaluation




**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.)**