Every Opinion Matters: Distributional and Long-tail Evaluation for LLMs

Xiang Lorraine Li

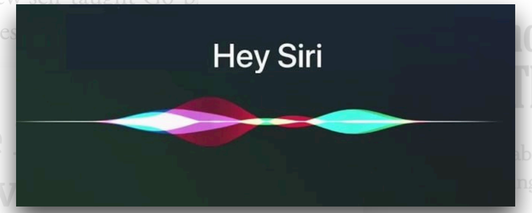
Assistant Professor at SCI Pitt

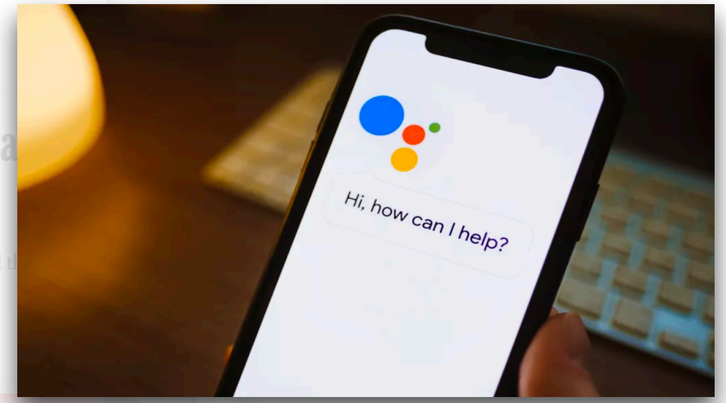
UMassAmherst
Robert and Donna Manning College of Information & Computer Sciences

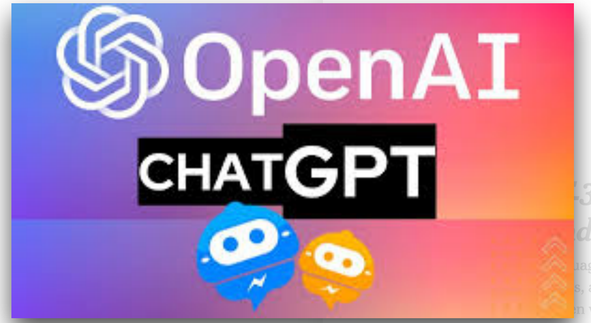
UNIVERSITY OF PITTSBURGH
VERITAS 1787 VIRTUS

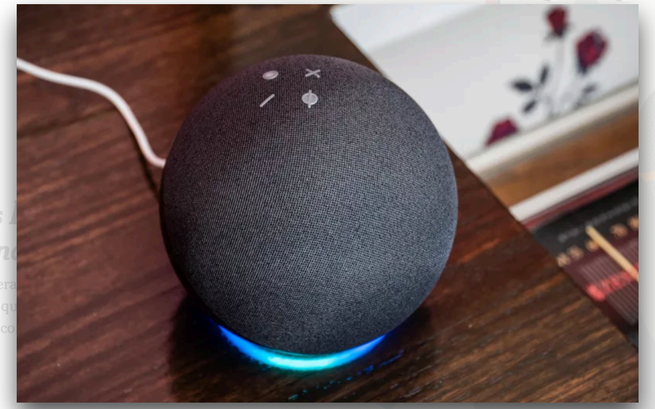
AI2 Allen Institute for AI

Impressive Progress in AI









Impressive Progress in AI

For a traditional Chinese wedd
specific customs, but here are

1. **Women:**
   - **Traditional Option:** I
     wearing a *cheongsan*
     collar. It can be embr
   - **Modern Option:** If th
     or a well-tailored suit
     reserved for the bride
   - **Colors to Consider:**
     and happiness, so we
     also favorable.

nAI

T

Diverse comm hat truly understand ar

Models Need Diverse Commonsense Knowledge with Different Views!

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 found in river.

Humans drink water.

Water can be used for cleaning.

Water can be used to wash clothes.

Water evaporates.

Water is wet.

They boiled the water.

Water needs to be held in a container.

Heat is needed to boil water.

Burner can provide heat.

Boiled water is too hot to drink.

Boiled water can cook food.

What is Common Sense?

Shared | **Implicit** | **Everyday Matters**

Water is liquid.

Water can be found in river.

Humans drink water.

Water evaporates.

Water can be used for cleaning.

Water can be used to wash clothes.

Water is wet.

They boiled the water.

Water needs to be held in a container.

Heat is needed to boil water.

Burner can provide heat.

Boiled water is too hot to drink.

Boiled water can cook food.

Why is Common Sense Challenging?

Water is liquid.

Water can be found in river.

Water can be used for cleaning.

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.

Heat is needed to boil water.

Burner can provide heat.

Boiled water is too hot to drink.

Boiled water can cook food.

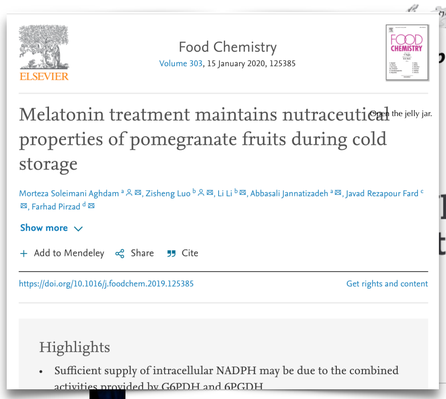
Why is Common Sense Challenging?

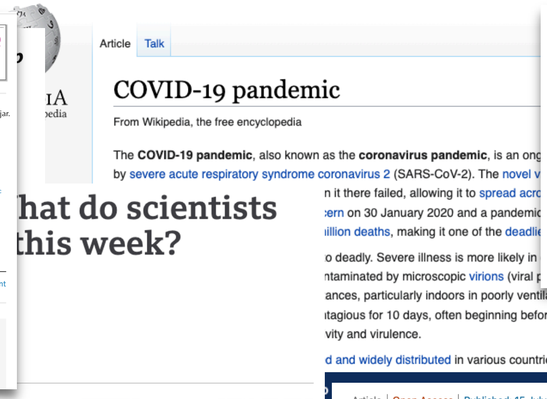
Massive

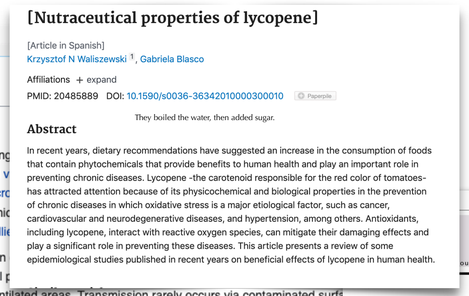
Humans drink water.
Boiled water can cook food.
Water can be found in river.
Heat is needed to boil water.
Water is liquid.
Open the jelly jar.
Water evaporates.
Some people love sugar.
can provide heat.
Human needs water to live.
People who wants to lose weight usually avoid peanut butter.
Humans drink water.
Water needs to be held in a container.
Heat is needed to boil water.
Water can be used for cleaning.
Heat is needed to boil water.
Water can be used to wash clothes.
Boiled water can cook food.
Boiled water is too hot to drink.
Water can be used for cleaning.
Boiled water is too hot to drink.
Sweet water tastes good
Boiled water can cook food.
Water is wet.
Human feel satisfied after having sweet stuff.
People needs tools to put peanut butter on the bread.
A knife with peanut butter could be the tool.
Human can put peanut butter on the bread
Spread the peanut butter on the bread.
Some people love sugar.
Peanut butter on the bread is usually breakfast.
People who wants to lose weight usually avoid peanut butter.
Sweet water tastes good
Peanut butter is high calorie food.
Some people hate sugar.
Human feel satisfied after having sweet stuff.
Peanut butter can be spread
Some people are allergic to peanut butter.
The kind of bread that can add peanut butter is flat.
Some people hate peanut butter.
Allergy reactions can be very serious, life-threatening.
Water can be used for cleaning.
Sugar can melt in water
They boiled the water, then added sugar.
Water evaporates.
Some people love sugar.
People who wants to lose weight usually avoid peanut butter.
Water needs to be held in a container.
Water can be used for cleaning.
Water can be used to wash clothes.
Boiled water is too hot to drink.
Water needs to be held in a container.
Water can be used to wash clothes.
They boiled the water.
Opening a jar needs tool
Heat is needed to boil water.
Human feel satisfied after having sweet stuff.
Water is wet.
Boiled water can cook food.
Peanut butter is high calorie food.
Water is wet.
Water needs to be held in a container.
Most bread is not sweet
There is water in the river
Peanut butter is sweet
A knife with peanut butter could be the tool.
Bread with peanut butter can be satisfying.
Boiled water can cook food.
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.
Sunset can be beautiful.
Sweet water tastes good
Sunset time is usually in the afternoon
Sugar water is wet.
There is water in the river
River water is not directly drinkable.
Water can be used for
Human feel satisfied after having sweet stuff.
Water needs to be h
They walked along the river at sunset time.
Most bread is not sweet
Sugar water is also liquid.
People are walking a
Water needs to be held in a container.
Humans drink water.
Water can be used for cleaning.
Water needs to be held in a container.
Water can be used for cleaning.
Humans drink water.
There are usually waiter helping you order food.
Water is wet.
Sometimes the ordering is automatic too.
Person can open jar, but not dogs
Order food means choosing dishes on the menu.
She walked into a restaurant and started ordering
Boiled water can cook food.
Heat is needed to boil water.
Sometimes the ordering is automatic
Human feel satisfied after having sweet stuff.
People walk into restaurant through door
Person can open jar, but not cats.
Water can be used for cleaning.
Ordering food needs menu
Restaurant serves food.
Boiled water can cook food.
Walking into a restaurant usually at breakfast/lunch/dinner time.
Water needs to be held in a container.
People walk into restaurant through door
Boiled water can cook food.
Ordering food needs menu
Tom asked me how to get to the library.
Water is wet.
Boiled water can cook food.
Walking into a restaurant usually at breakfast/lunch/dinner time.
Water needs to be held in a container.

Why is Common Sense Challenging?

Massive







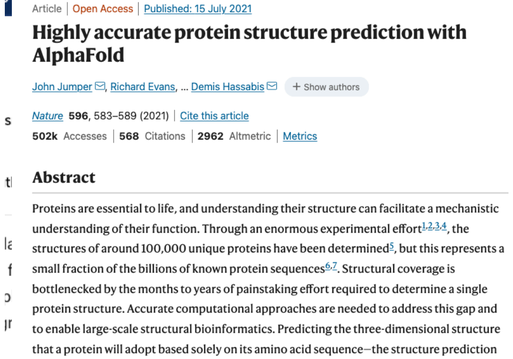
COP26 is seen as crucial if climate change is to be brought under control

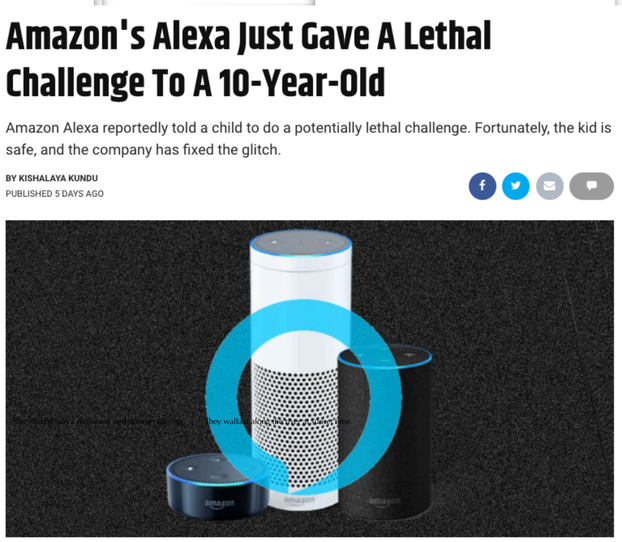
As the COP26 climate summit enters its second week, negotiations in Glasgow have hit a critical phase.

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.





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

Why is Common Sense Challenging?

Massive

Water is liquid.

Water can be found in ocean.

Humans drink water.

Water evaporates.

Water can be used for cleaning.

Water can be used to wash clothes.

Water is wet.

They boiled the water.

Water needs to be held in a container.

Heat is needed to boil water.

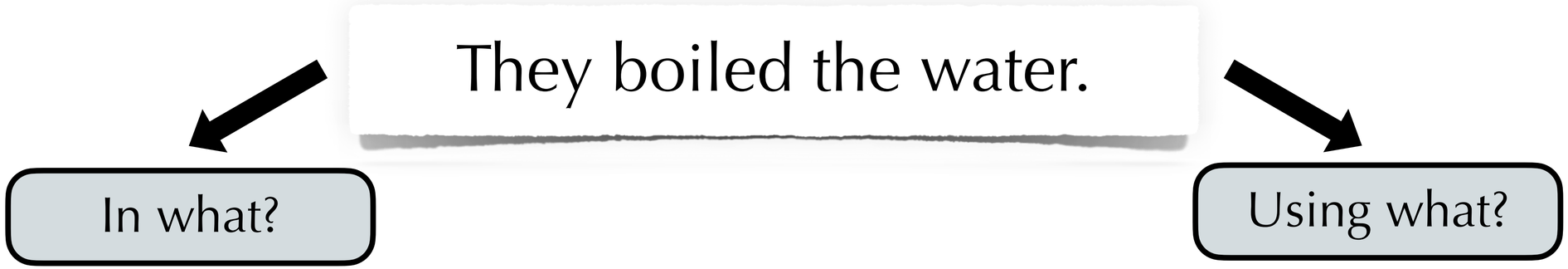
Burner can provide heat.

Boiled water is too hot to drink.

Boiled water can cook food.

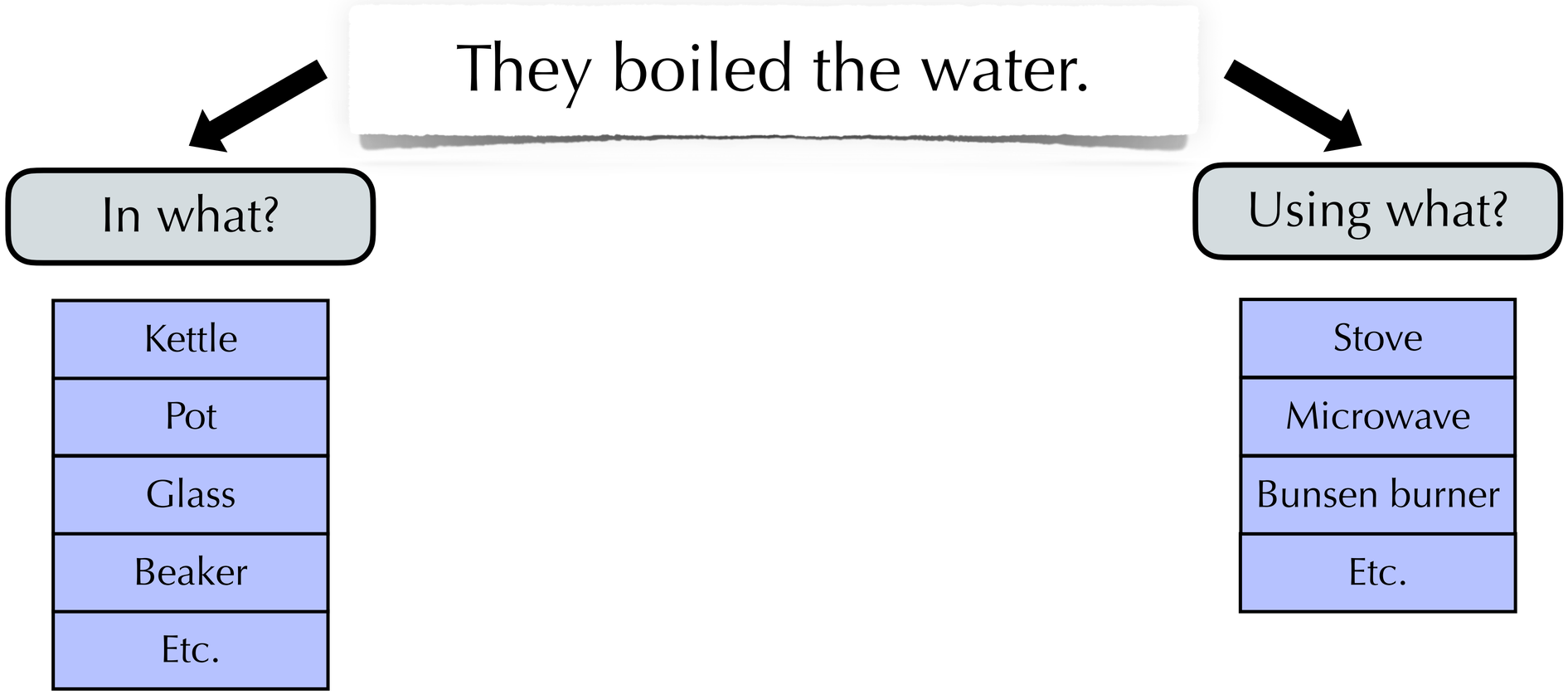
Why is Common Sense Challenging?

Massive



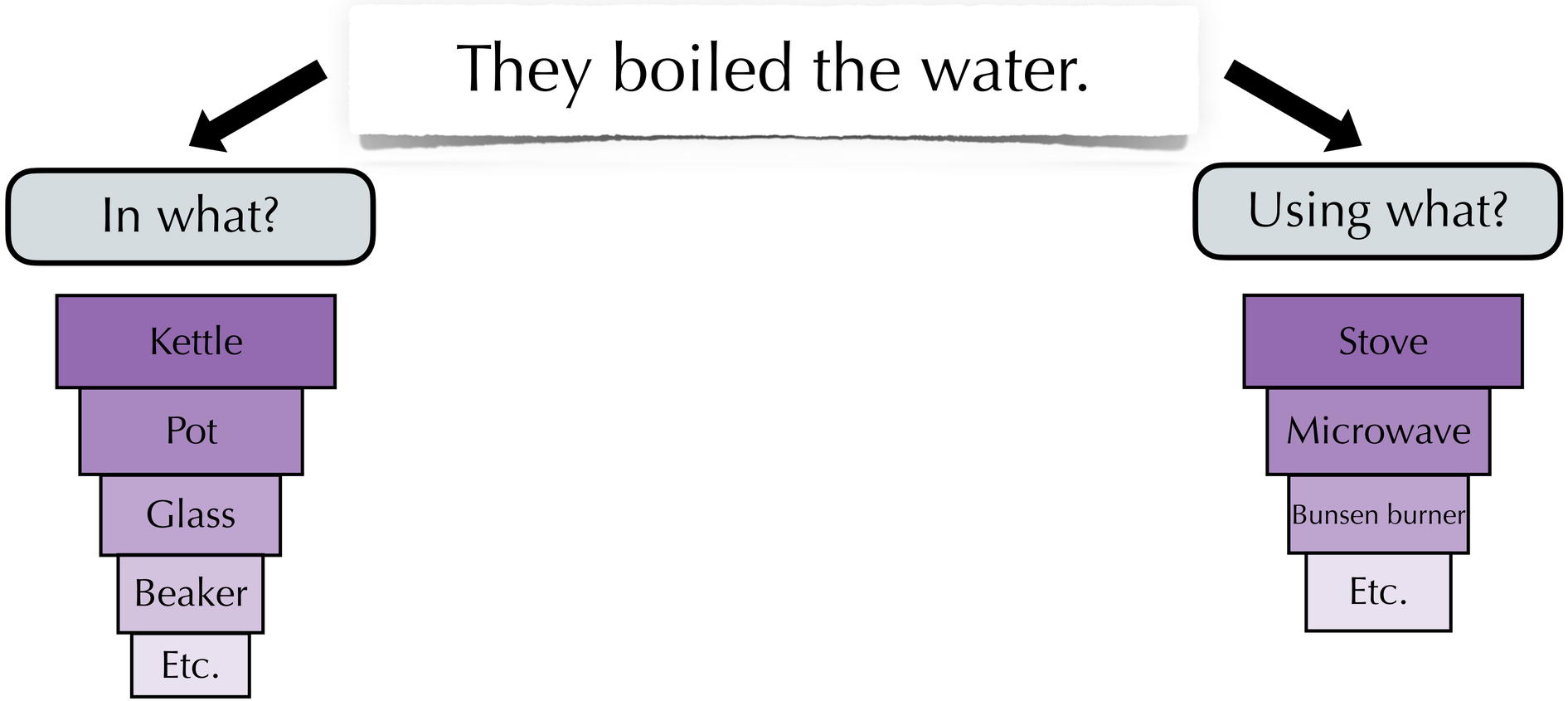
Why is Common Sense Challenging?

Massive



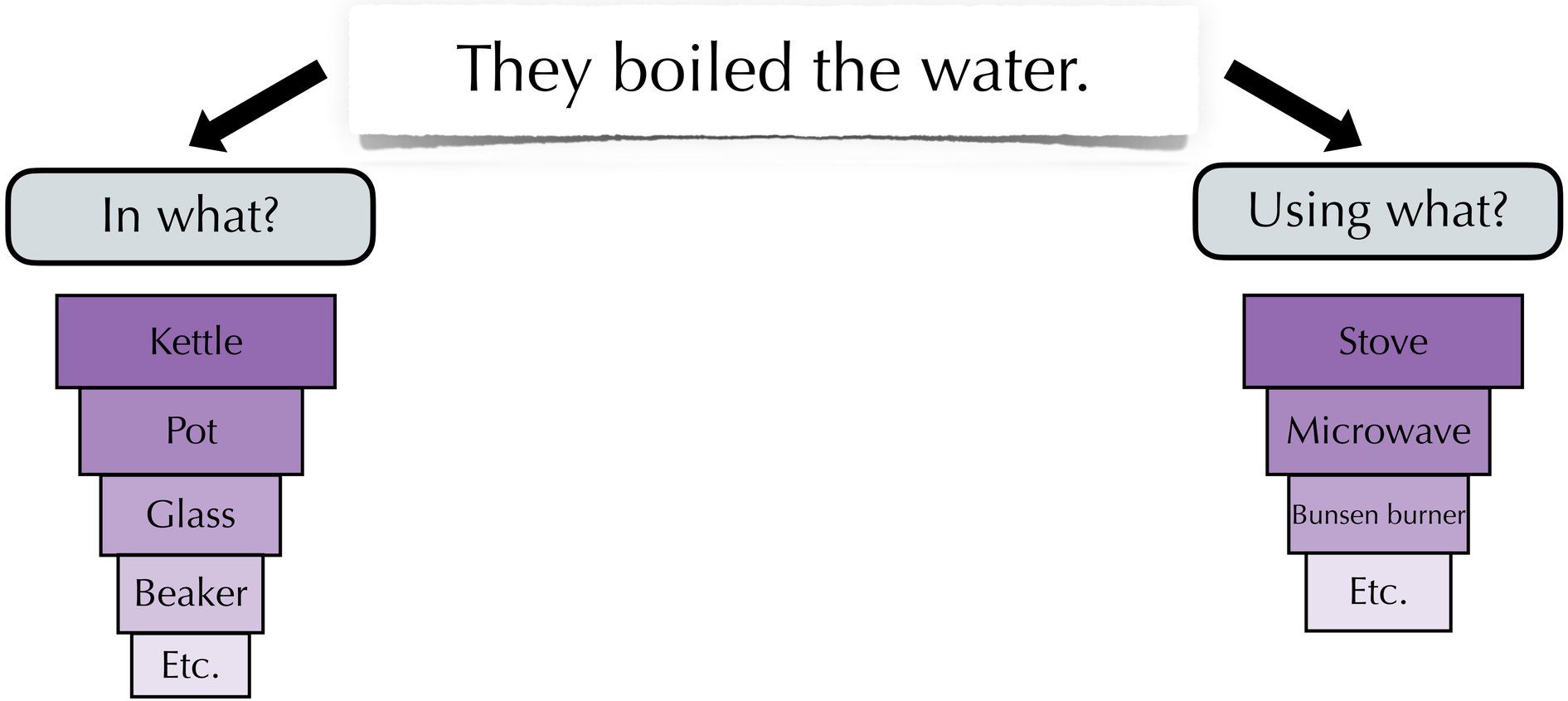
Why is Common Sense Challenging?

Massive



Why is Common Sense Challenging?

Massive **Probabilistic**



Why is Common Sense Challenging?

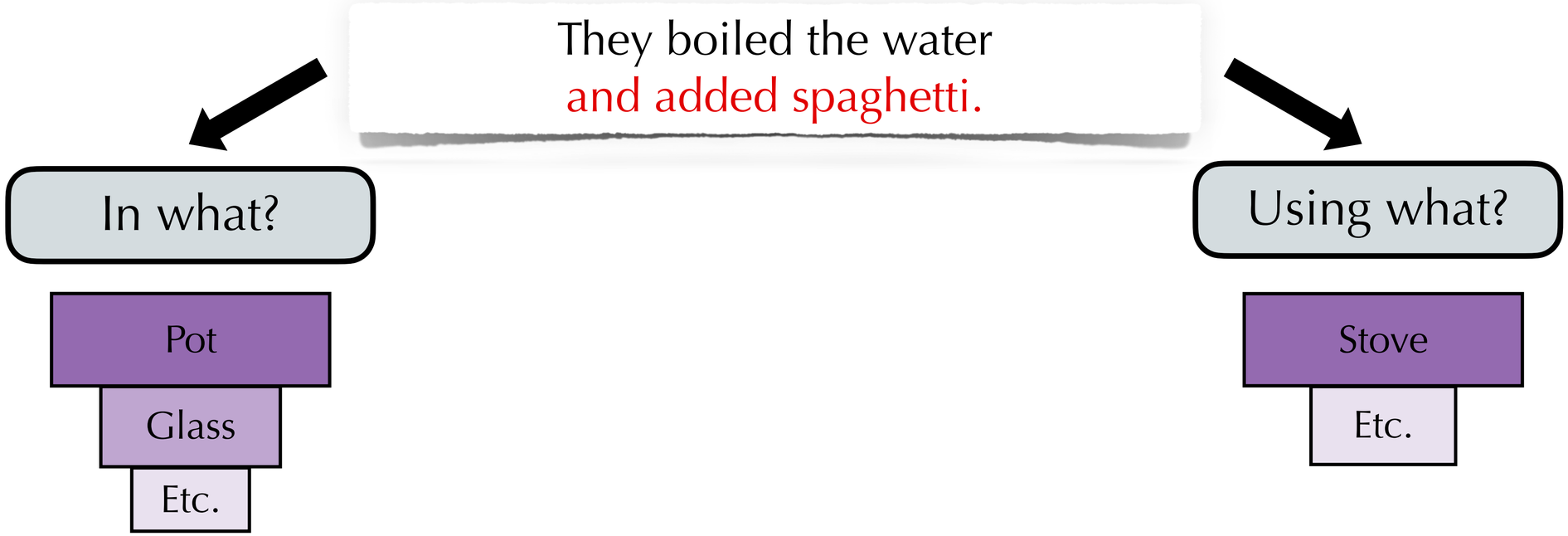
Massive **Probabilistic**

They boiled the water
and added spaghetti.

Why is Common Sense Challenging?

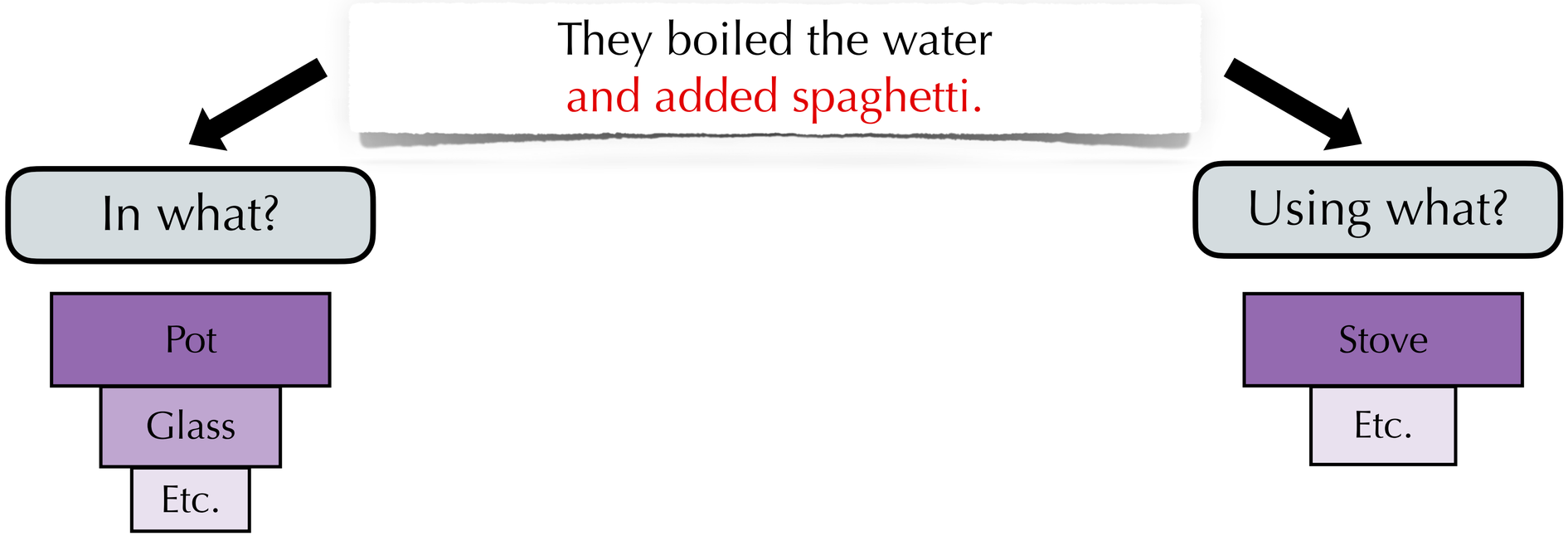
Massive

Probabilistic



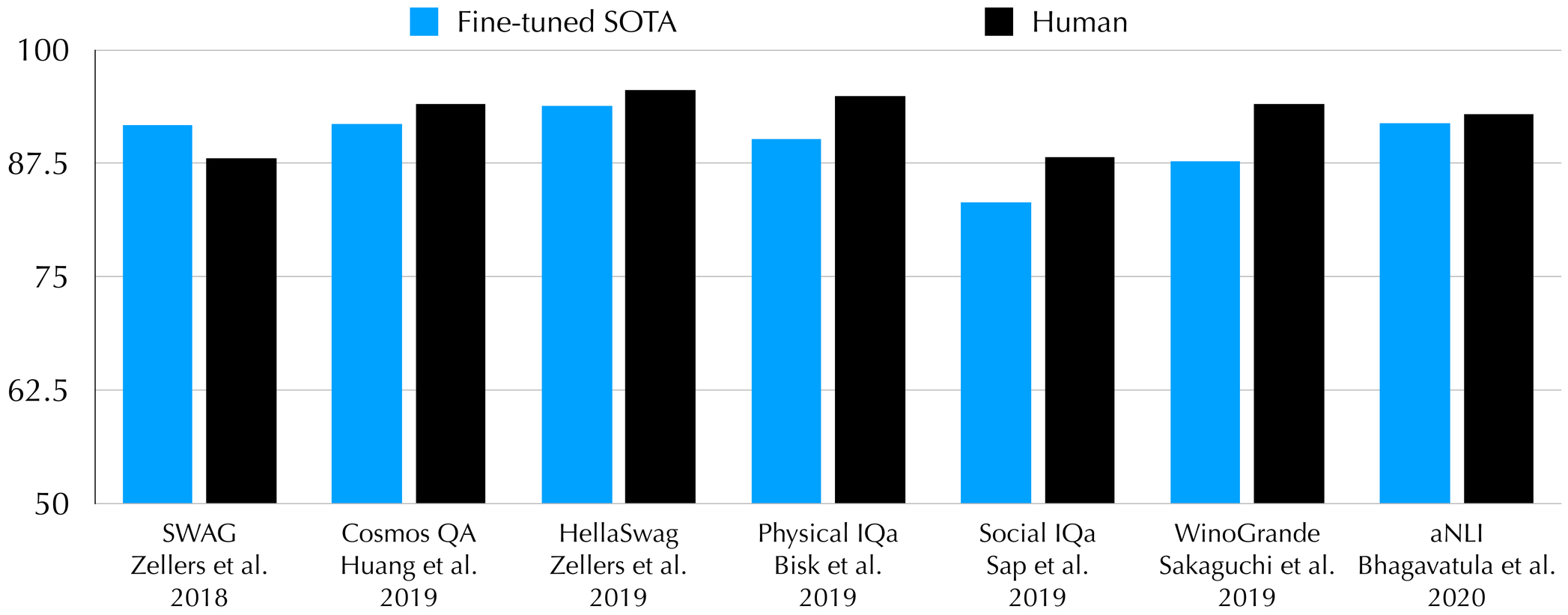
Why is Common Sense Challenging?

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



Common Sense in Language Model

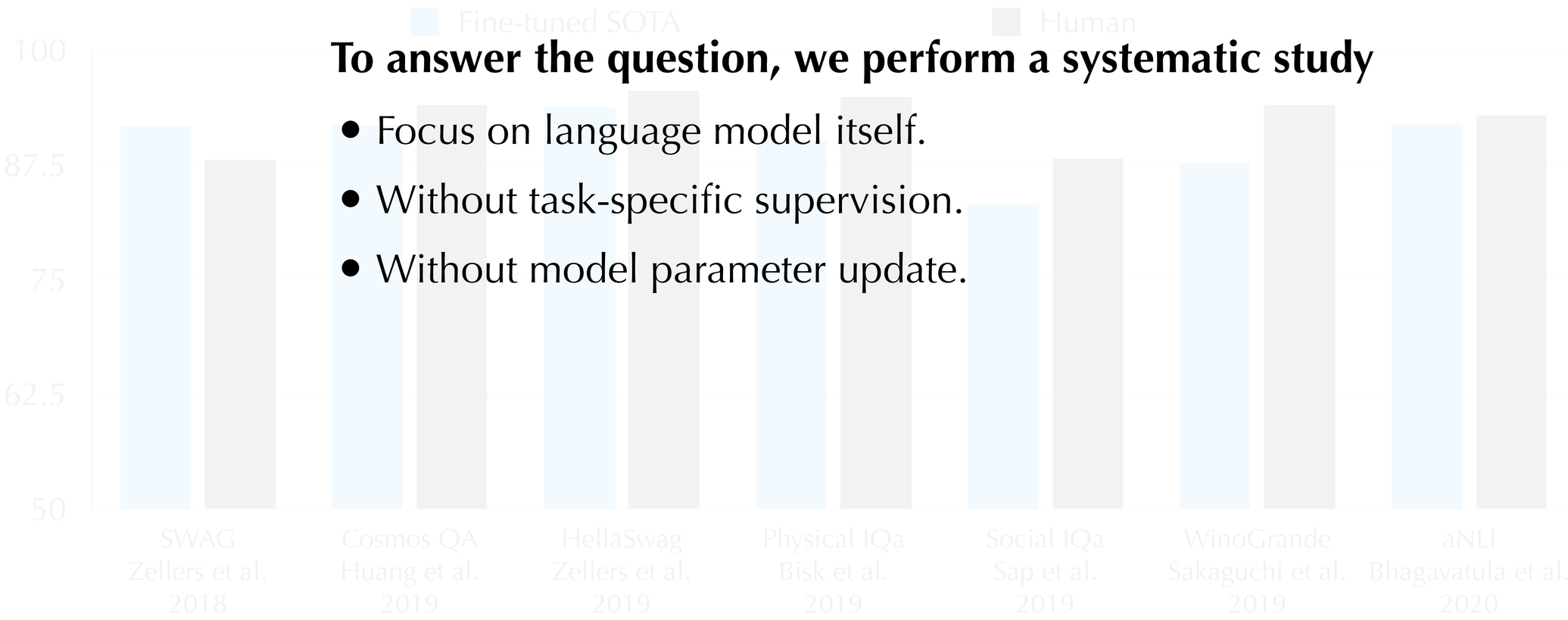
Models based on large language models show impressive performance on many **commonsense question answering** tasks.



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?

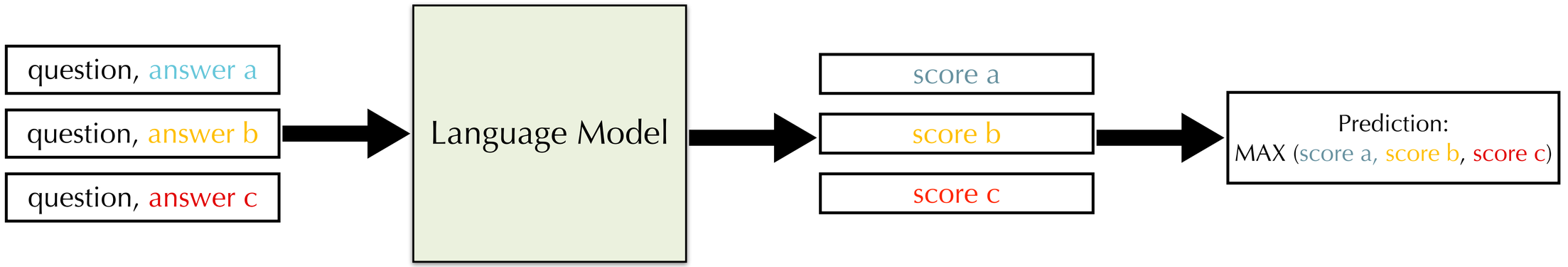
Dataset	Example	Number of Choices	Reasoning Type
Physical IQa (Bisk et al. 2019)	**Question:** To apply eyeshadow without a brush, should I use a cotton swag or a toothpick? **Answer:** Cotton swab.	2	Physical
Social IQa (Sap et al. 2019)	**Question:** Tracy had accidentally pressed upon Austin in the 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.

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

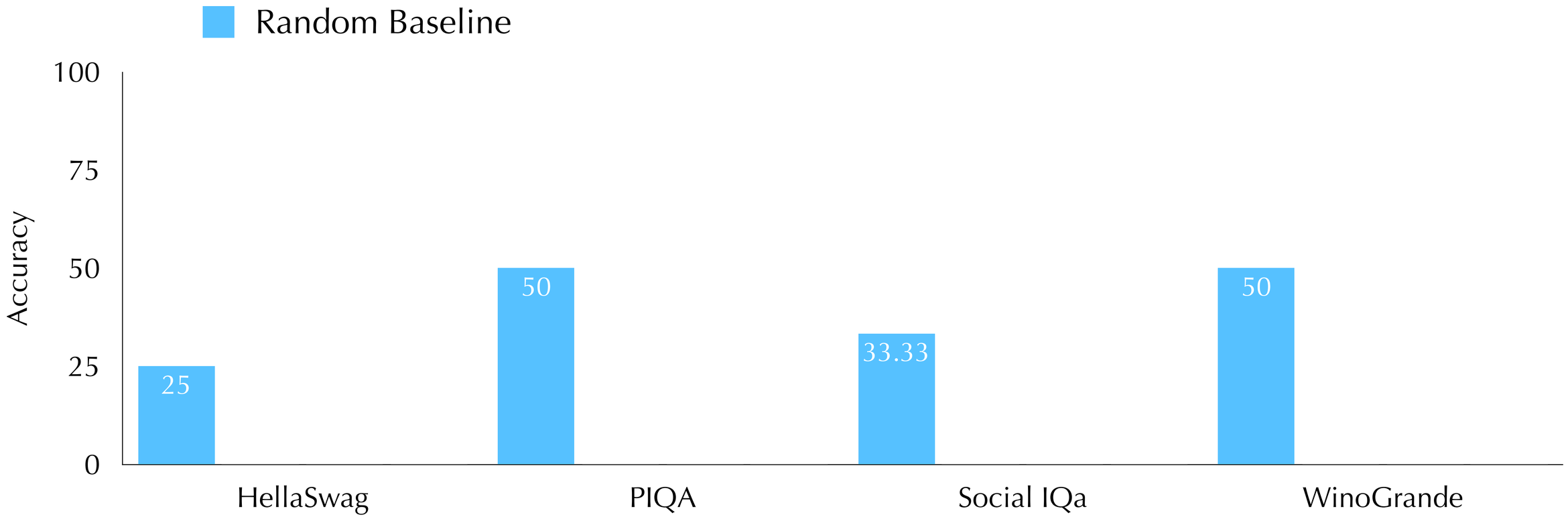


Figure: the dev accuracy for each dataset evaluated on Gopher.

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

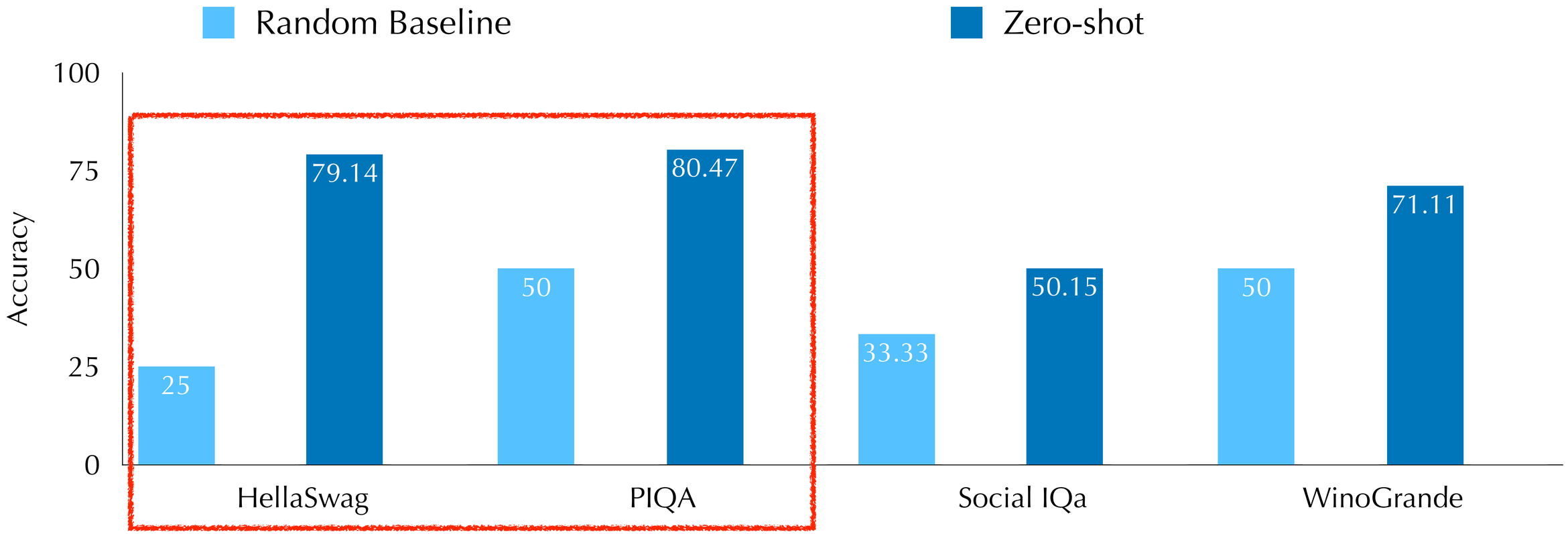


Figure: the dev accuracy for each dataset evaluated on Gopher.

How much of the performance comes only from answers?

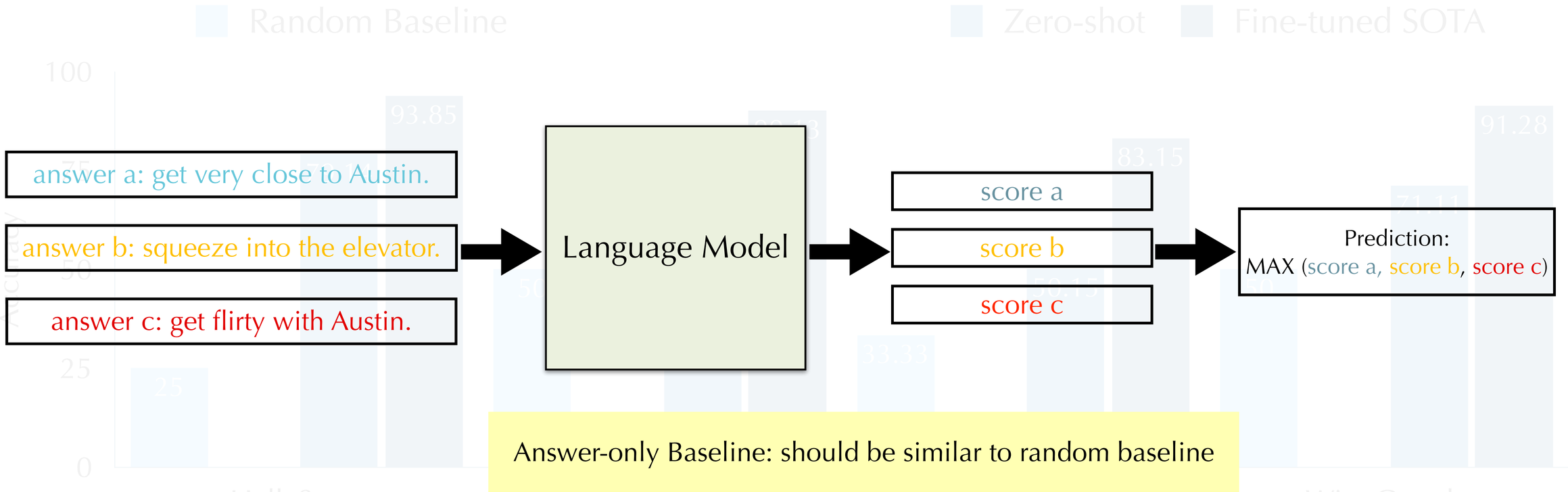


Figure: the dev accuracy for each dataset evaluated on Gopher.

Models pick the correct answer without seeing the question

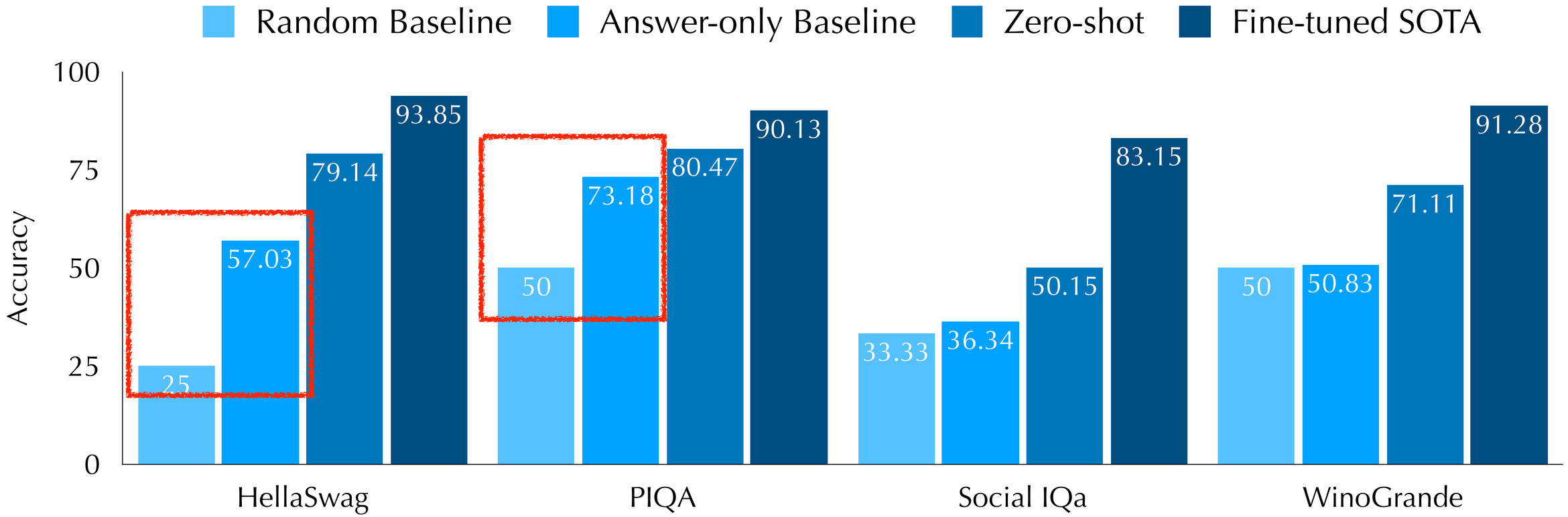


Figure: the dev accuracy for each dataset evaluated on Gopher.

We need better commonsense evaluation!

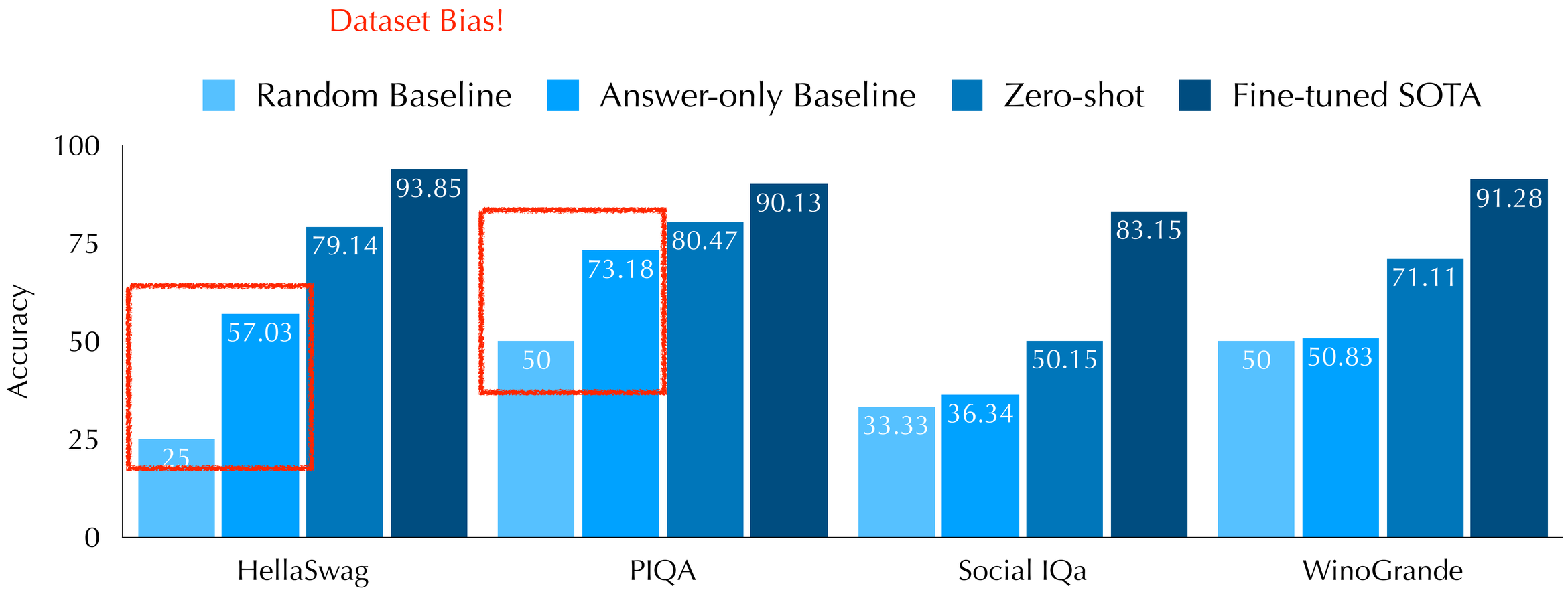


Figure: the dev accuracy for each dataset evaluated on Gopher.

Outline

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

- Every Answer Matters: Evaluating Commonsense with Probabilistic Measures. [ACL 2024]
- Leveraging Large Models for Evaluating Novel Content: A Case Study on Advertisement Creativity. [EMNLP 2025]

Benchmark: Long-tail Question: Commonsense Reasoning Evaluation

- UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations. [NAACL 2024]
- In search of the long-tail: systematic generation of long-tail knowledge via logical rule guided search [EMNLP 2024]
- Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games. [EMNLP 2025]

Outline

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

- Every Answer Matters: Evaluating Commonsense with Probabilistic Measures. [ACL 2024]
- Leveraging Large Models for Evaluating Novel Content: A Case Study on Advertisement Creativity. [EMNLP 2025]

Benchmark: Long-tail Question: Commonsense Reasoning Evaluation

- UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations. [NAACL 2024]
- In search of the long-tail: systematic generation of long-tail knowledge via logical rule guided search [EMNLP 2024]
- Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games. [EMNLP 2025]

Every Answer Matters: Evaluating Commonsense with Probabilistic Measures

Qi Cheng

Michael Boratko

Pranay Yelugam

Tim O' Gorman

Nalini Singh

Andrew McCallum

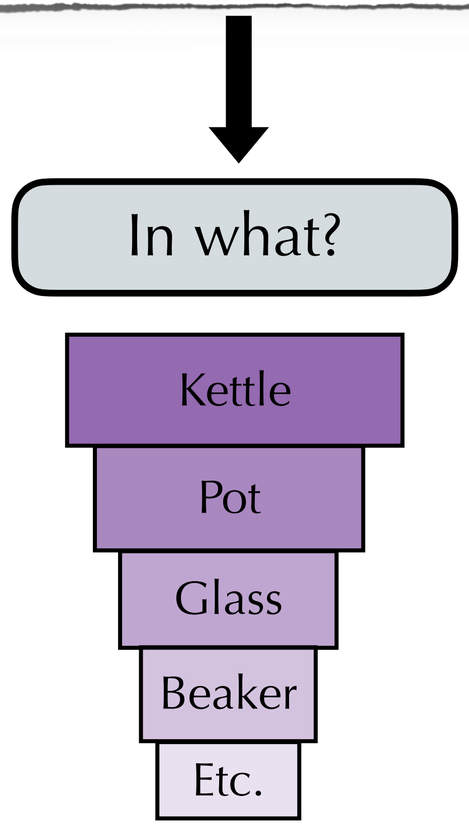
Xiang Lorraine Li

University of Pittsburgh

ACL 2024
Bangkok, Thailand

Probabilistic Evaluation of Commonsense

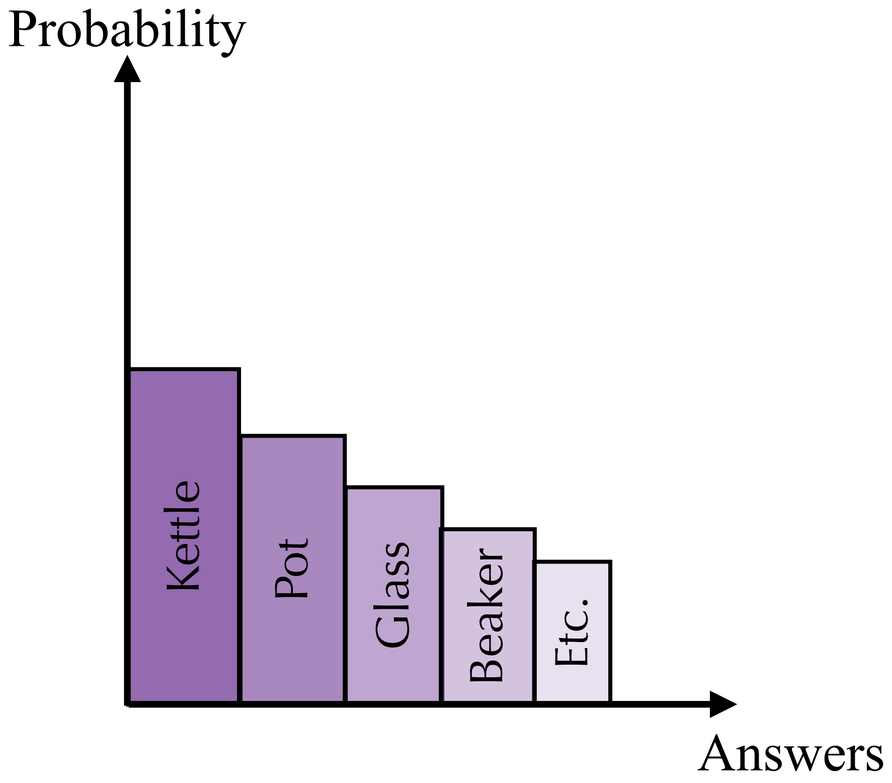
They boiled the water.



Probabilistic Evaluation of Commonsense

They boiled the water.

In what?



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

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

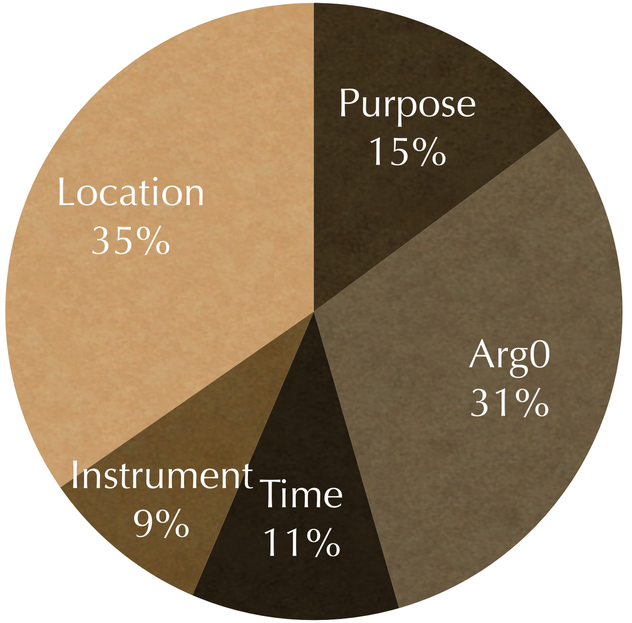
We crowd-sourced high-quality **101 questions (manual filtering)**

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)**

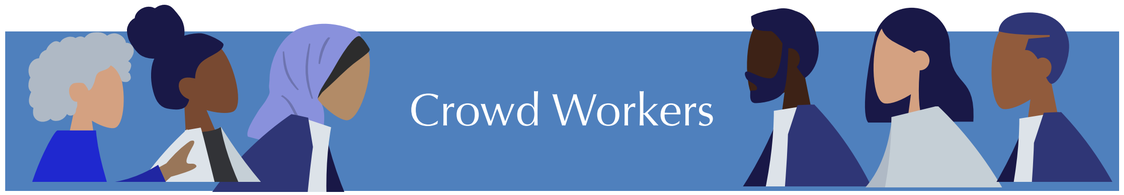
Missing Slot	Definition	Examples
Arg0	Who/what does the event?	Sentence: putting cheese on the pizza. Arg0? Answers: person, cook
Purpose	What is the goal for doing the event?	Sentence: putting cheese on the pizza. Purpose? Answers: get nutrition, stop being hungry
Instrument	What kind of tools are used 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
disinfect
cook
make tea
disinfecting
making dinner
for making tea
cleaning
for a hot drink
to cook
cook food
cleaning tools
cooking spaghetti
making pasta
kill bacteria
steaming vegetables
purify
purification
boiling potatoes
make safe to drink
boiling chicken
sterilization
for an experiment

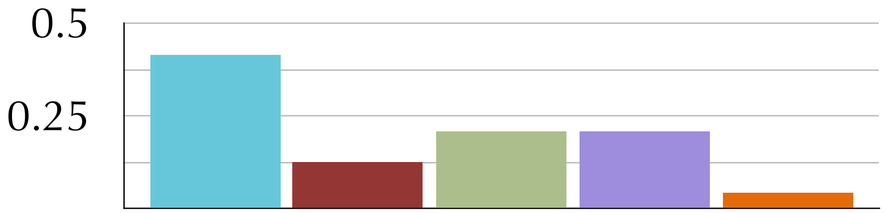


CFC Data Collection

“They boiled the water” Purpose?

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



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]

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

[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.
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]

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

  - 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?

[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]

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

  - 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?

[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

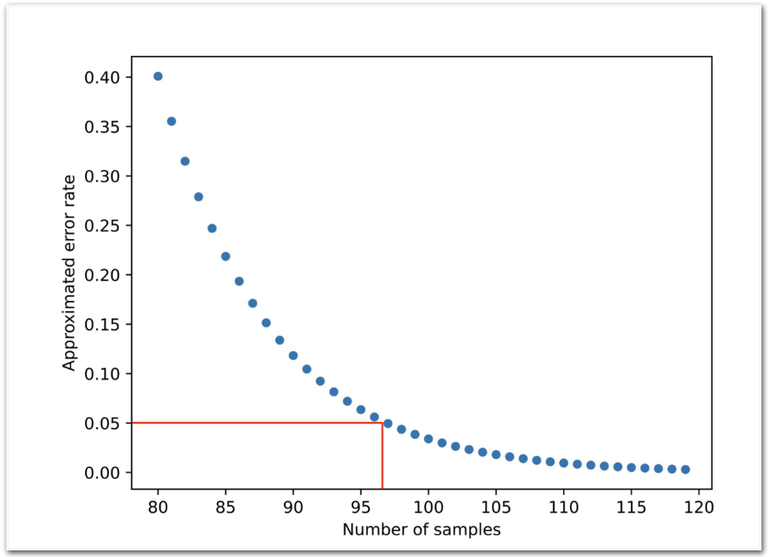
CFC Data Collection

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

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

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

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

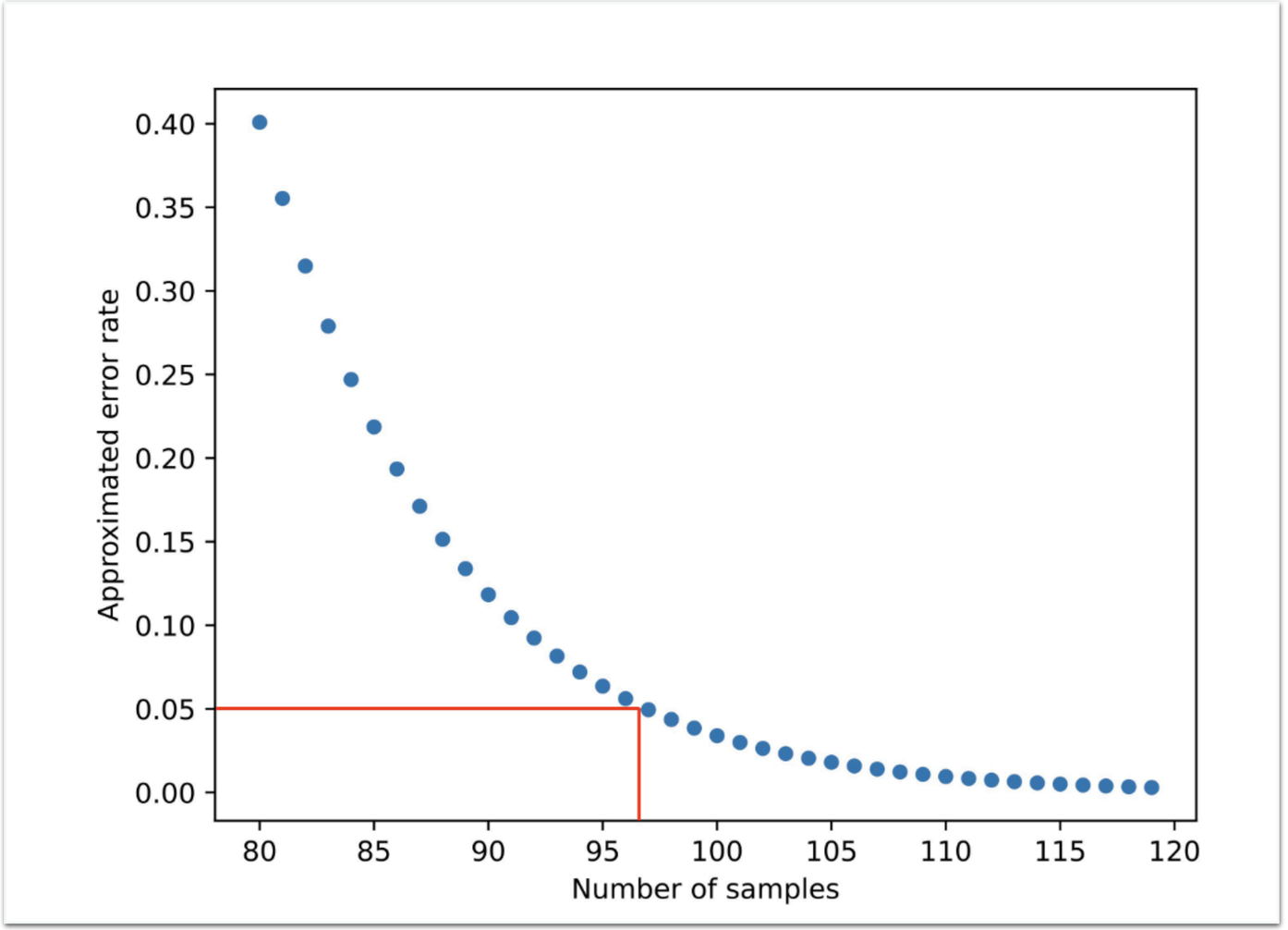


[1] Mardia, Jay, Jiantao Jiao, Ervin Tánczos, Robert D. Nowak, and Tsachy Weissman. "Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types." *Information and Inference: A Journal of the IMA* 9, no. 4 (2020): 813-850.
Qi, Boratko, Yelugam, O'Gorman,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

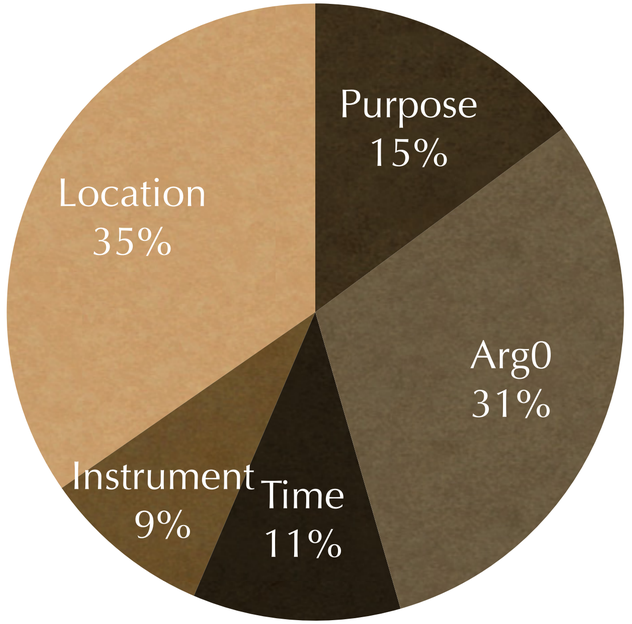
CFC Data Statistics

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**:

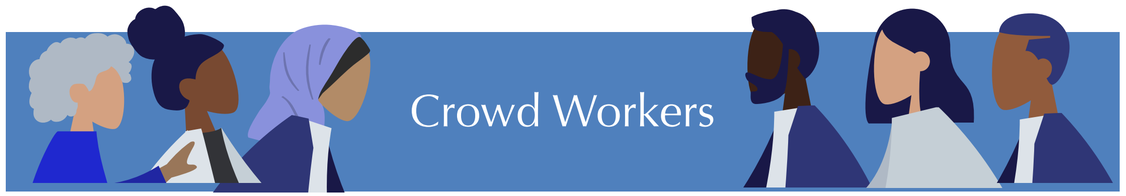


Question Slot Type

> 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 …

CFC Probabilistic Evaluation

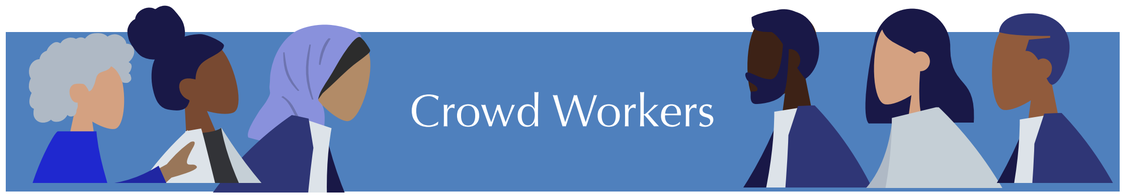
"They boiled the water" Purpose?



CFC Probabilistic Evaluation

"They boiled the water" Purpose?

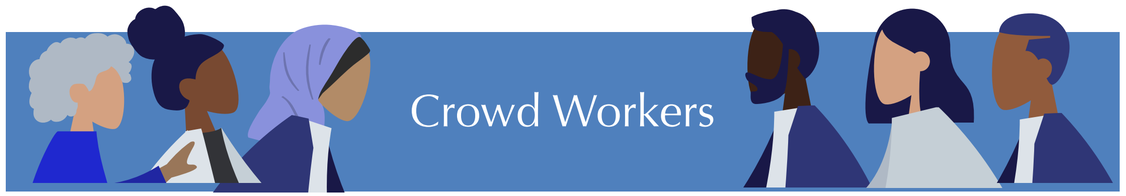
cooking
clean
disinfect
cook
make tea
disinfecting
making dinner
for making tea
cleaning
for a hot drink
to cook
cook food
cleaning tools
making pasta
cooking spaghetti
kill bacteria
steaming vegetables
purify
purification
boiling potatoes
make safe to drink
boiling chicken
sterilization
for an experiment



CFC Probabilistic Evaluation

“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



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 Probabilistic Evaluation

“They boiled the water” Cause?

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

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

CFC Probabilistic Evaluation

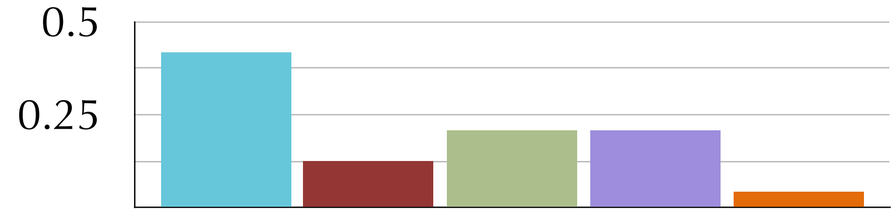
"They boiled the water" Purpose?

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

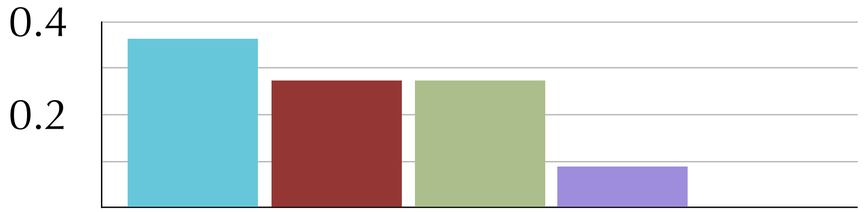
clean
disinfect
disinfecting
cleaning
cleaning tools
kill bacteria
purify
purification
make safe to drink
sterilization

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 =



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

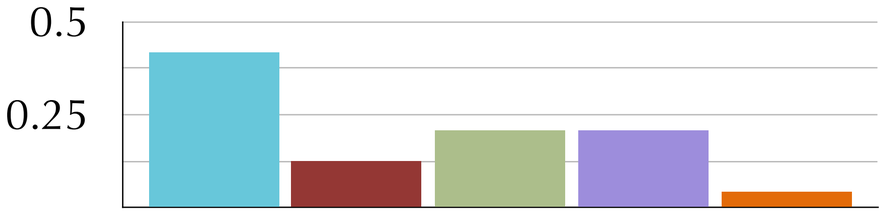
CFC Probabilistic Evaluation

"They boiled the water" Purpose?

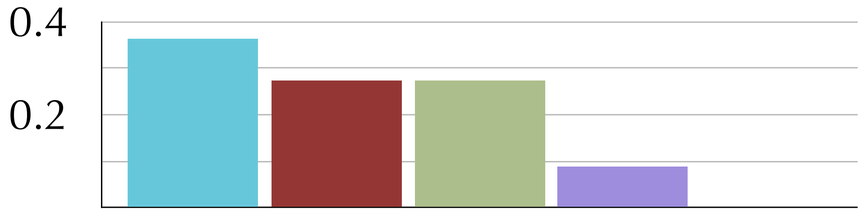
for making tea
for a hot drink
make tea
cooking spaghetti
making pasta
to cook
steaming vegetables
boiling potatoes
cook
boiling chicken
cook food
making dinner
cooking
for an experiment

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



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

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

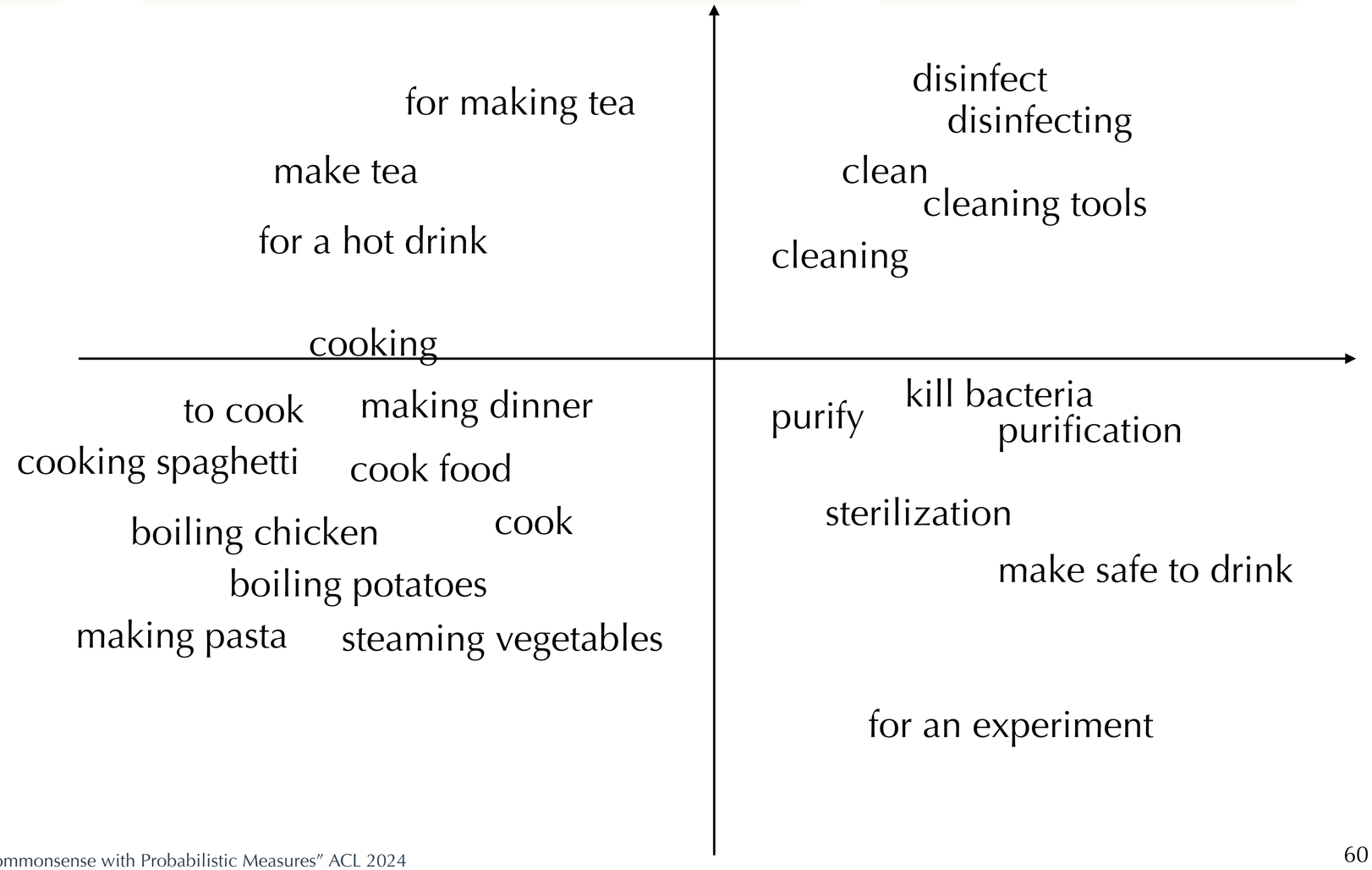
Model prediction: H

CFC Automatic Evaluation

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

With Context
- BERT
- RoBERTa

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

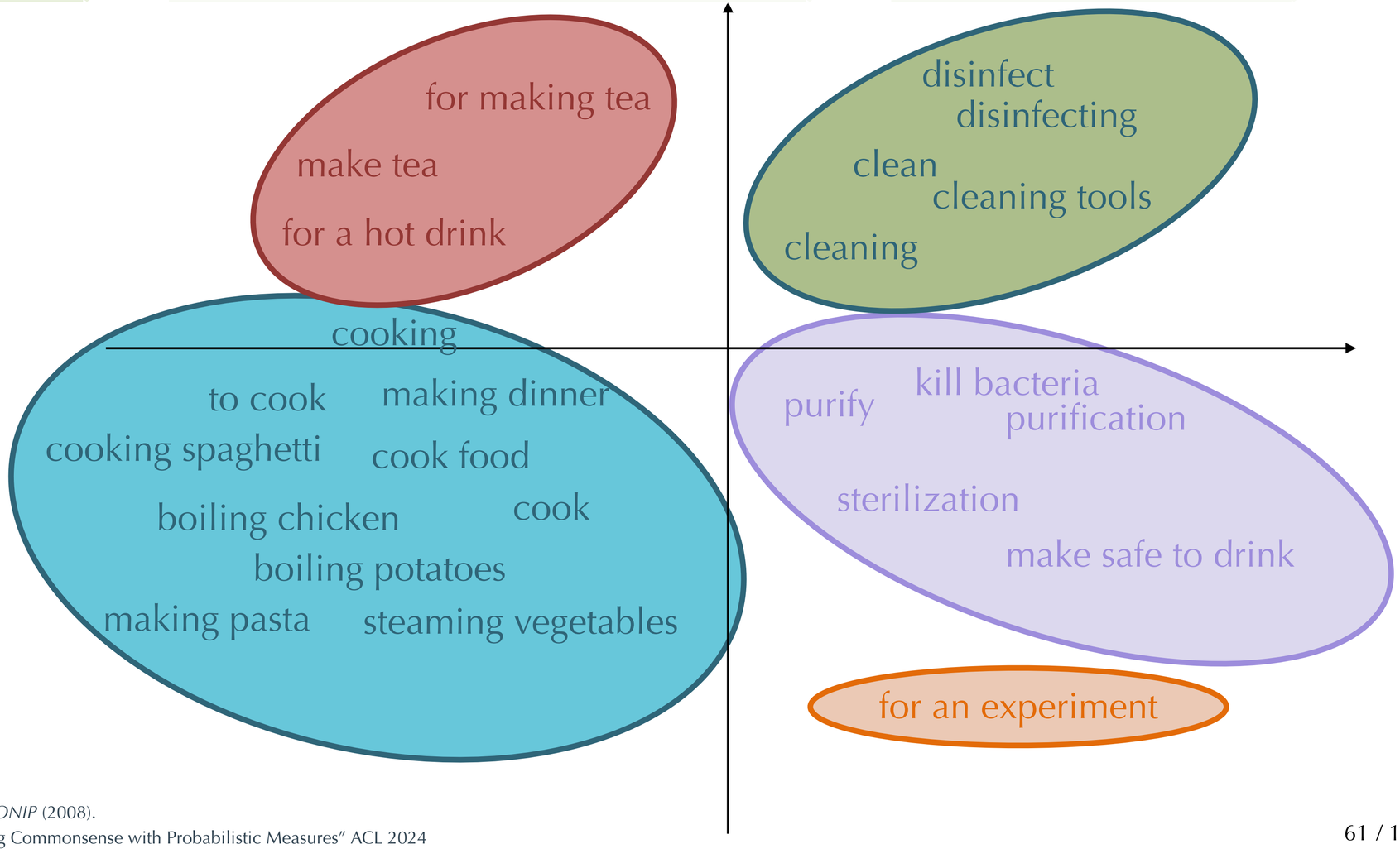


CFC Automatic Evaluation

Embed G | Cluster G | Match H to cluster of G | Calculate Score

Clustering Algorithm

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

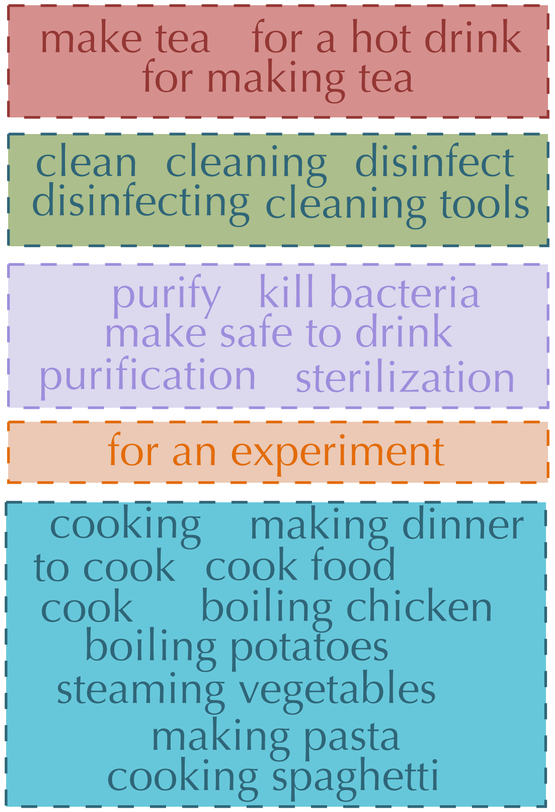


[1] Zhao, Zhonghua et al. "G-Means: A Clustering Algorithm for Intrusion Detection." *ICONIP* (2008).

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 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

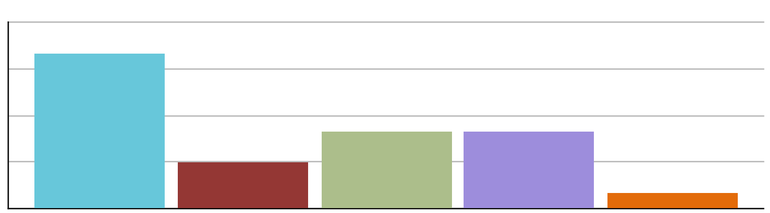
Embed G → Cluster G → Match H to cluster of G → Calculate Score

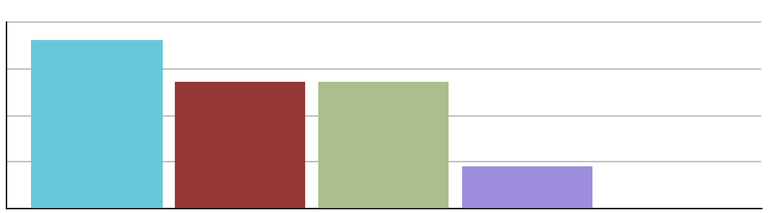
Embeddings Based
- FastText

Lexical Token Based
- WordNet

CFC Automatic Evaluation

Embed G | Cluster G | Match H to cluster of G | Calculate Score

P = 

Q = 

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

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**

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**

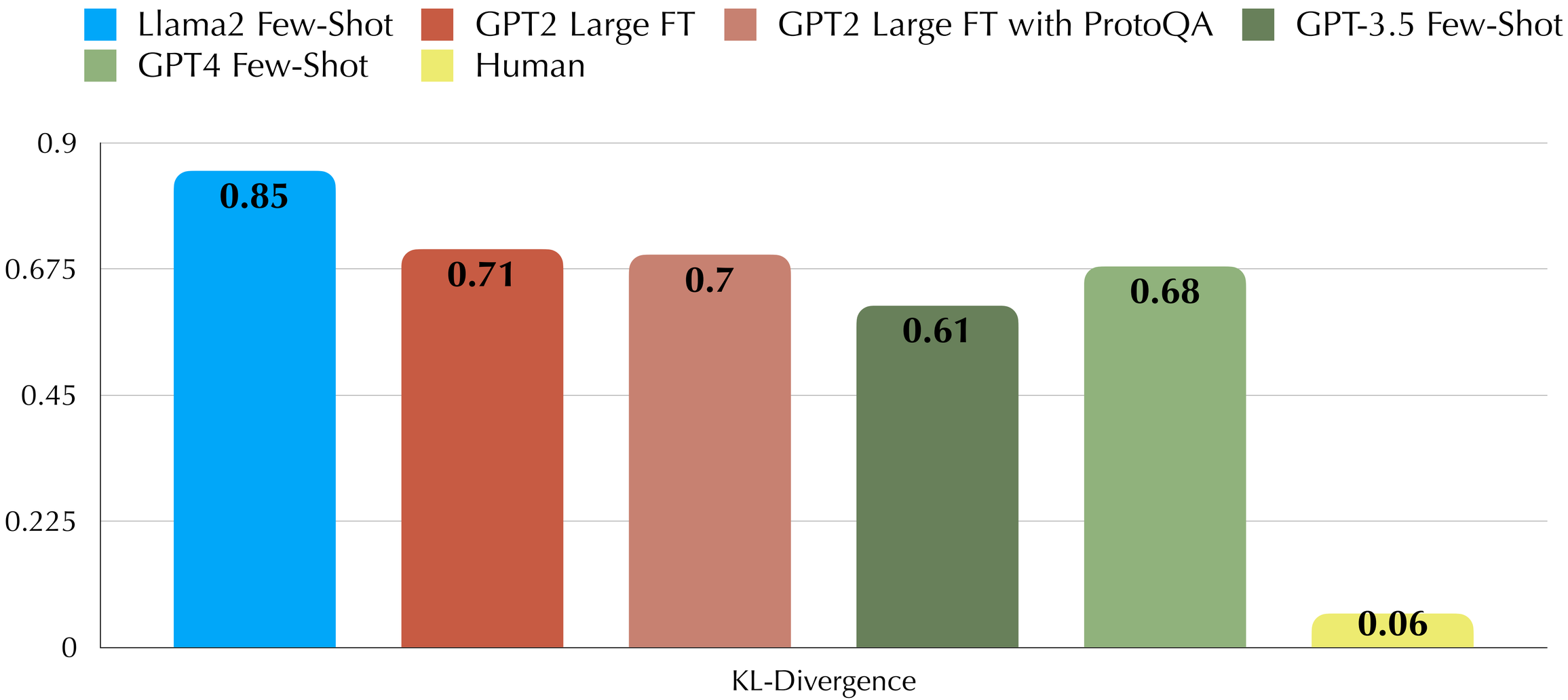
Evaluating Automatic Metric - PROBEVAL

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



Leveraging Large Models to Evaluate Novel Content A Case Study on Advertisement Creativity

Joey Hou

Adriana Kovashka

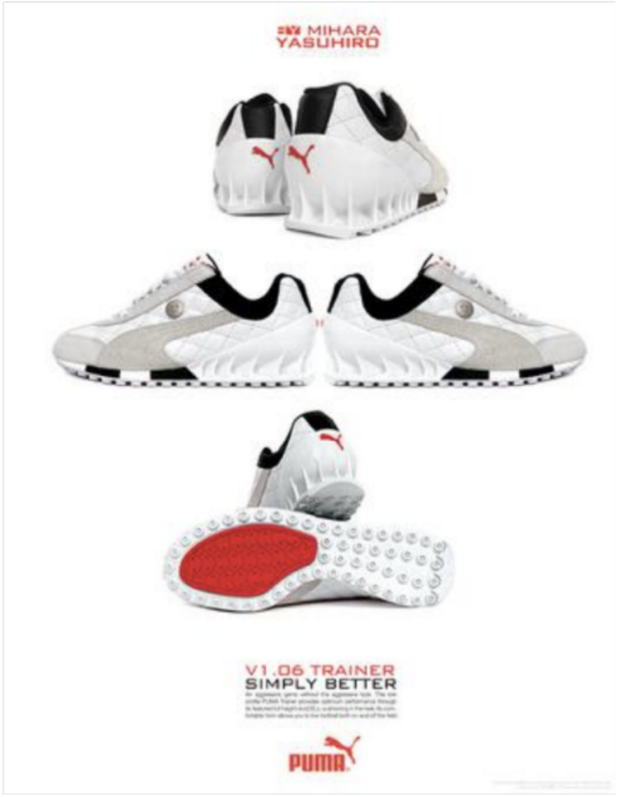
Xiang Lorraine Li

University of Pittsburgh

EMNLP 2025
Suzhou, China 中国苏州

Which image ads is more creative

A



B

On the scale from 1 to 3, What's their creativity level?

A

B

Advertisement Creativity

It’s a subjective task

How will VLMs perform on subjective tasks like this?

Ads Creativity Dataset Construction

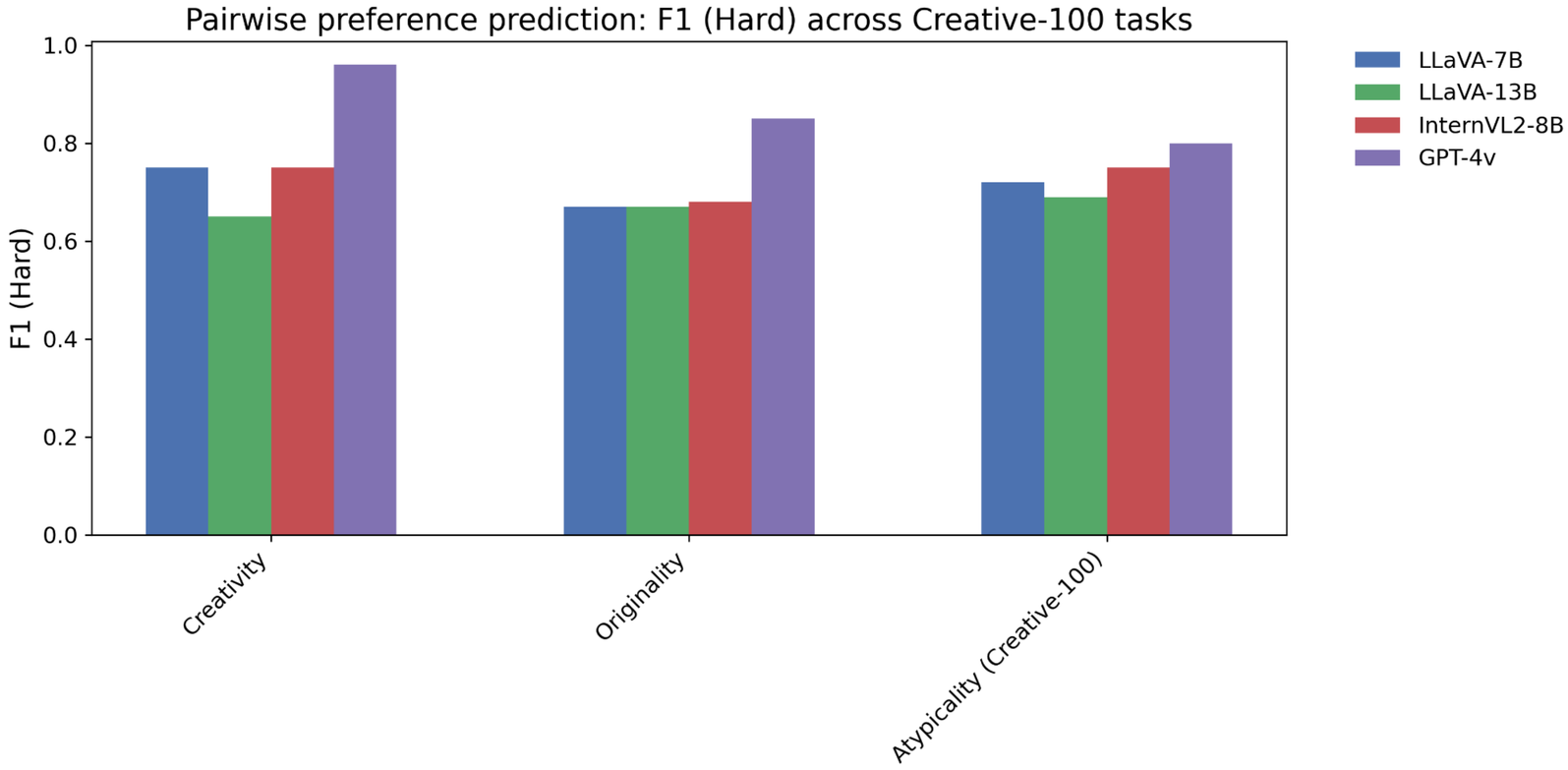
- **Ads Images**
  - Subsets: *Creative-100* and *Atypical-300*
  - Topics: clothing, food, automobile, beauty, leisure, electronics, drinks, service, non-commercial, and healthcare

- **Fine-grained Annotation (*Creative-100*)**
  - **Atypicality, Originality, Overall Creativity [3]**
  - Annotators: Amazon Mechanical Turk
  - 25 annotations per data point

Ads Creativity Tasks

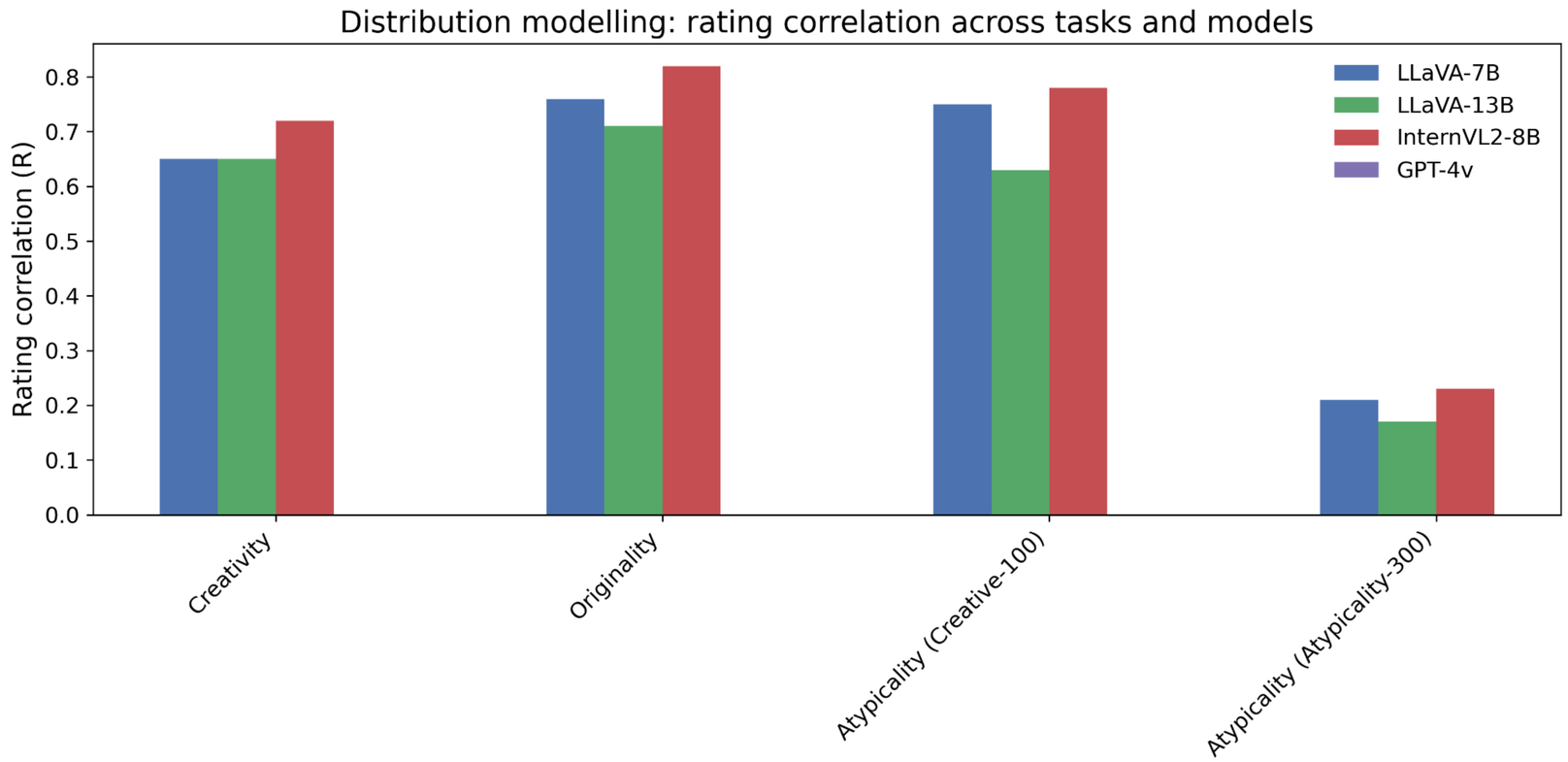
- **Distribution Modeling**: Simulate human group behavior
  - Task Setup: VLM simulates scores distribution by repeated prompting (for creativity, atypicality, originality)
  - Metrics
    - KL Divergence (human scores V.S. VLM scores)
    - Spearman's Correlation (human scores V.S. VLM scores)
- **Pairwise Evaluation**
  - Task Setup: VLM predicts which advertisement is more creative/atypical/original
  - Metric: Accuracy (binary)
- **Disagreement Prediction**: estimate the level of disagreement and ambiguity
  - Task Setup: VLM directly predicts disagreement level (3 scale)
  - Metric: Spearman's Correlation (human std. V.S. VLM prediction)

Ads Creativity: Findings



- Great performance for the pair-wise task.

Ads Creativity: Findings



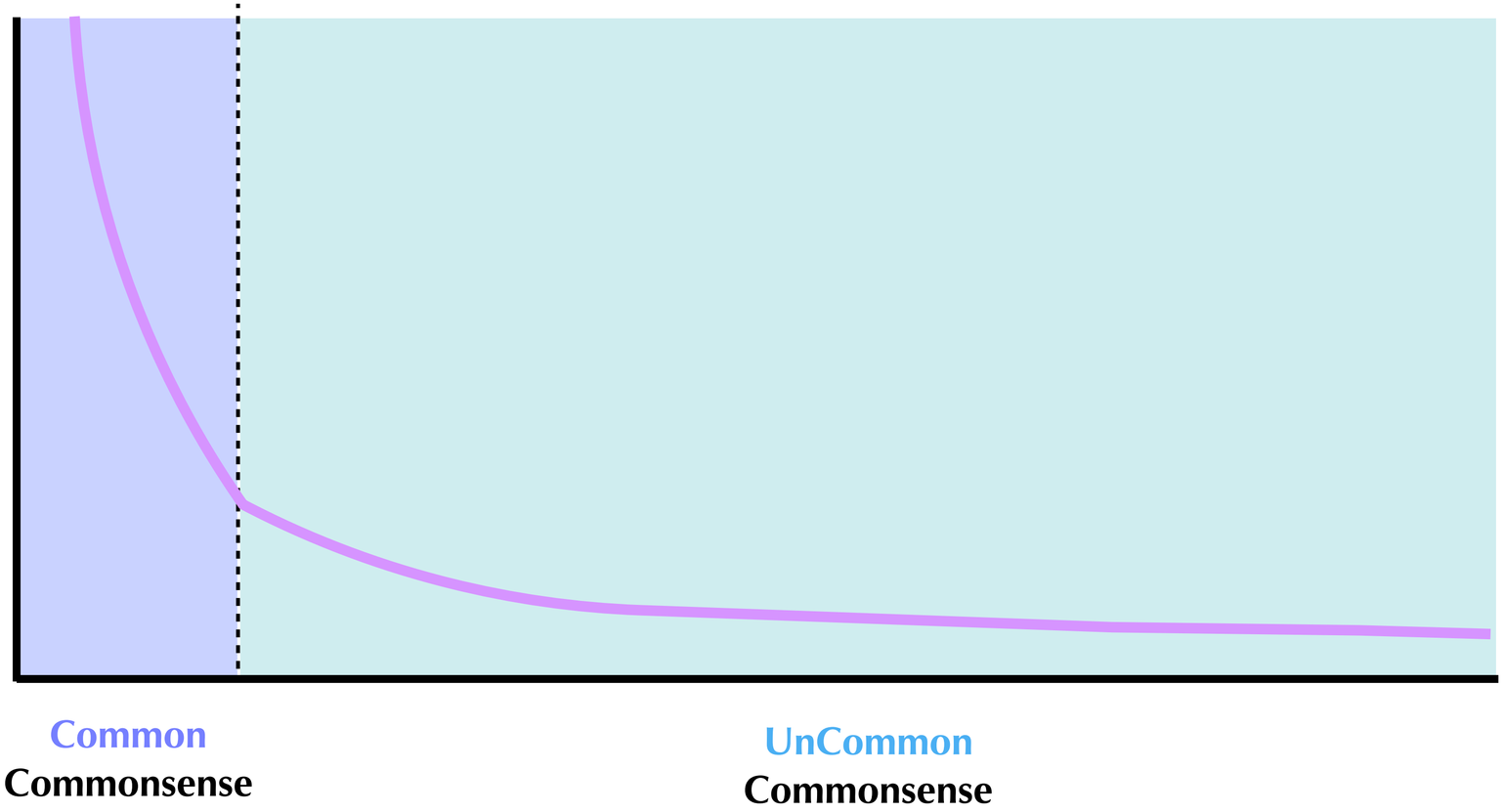
- Okay results for distributional modeling tasks, similar to CFC.

Ads Creativity: Findings

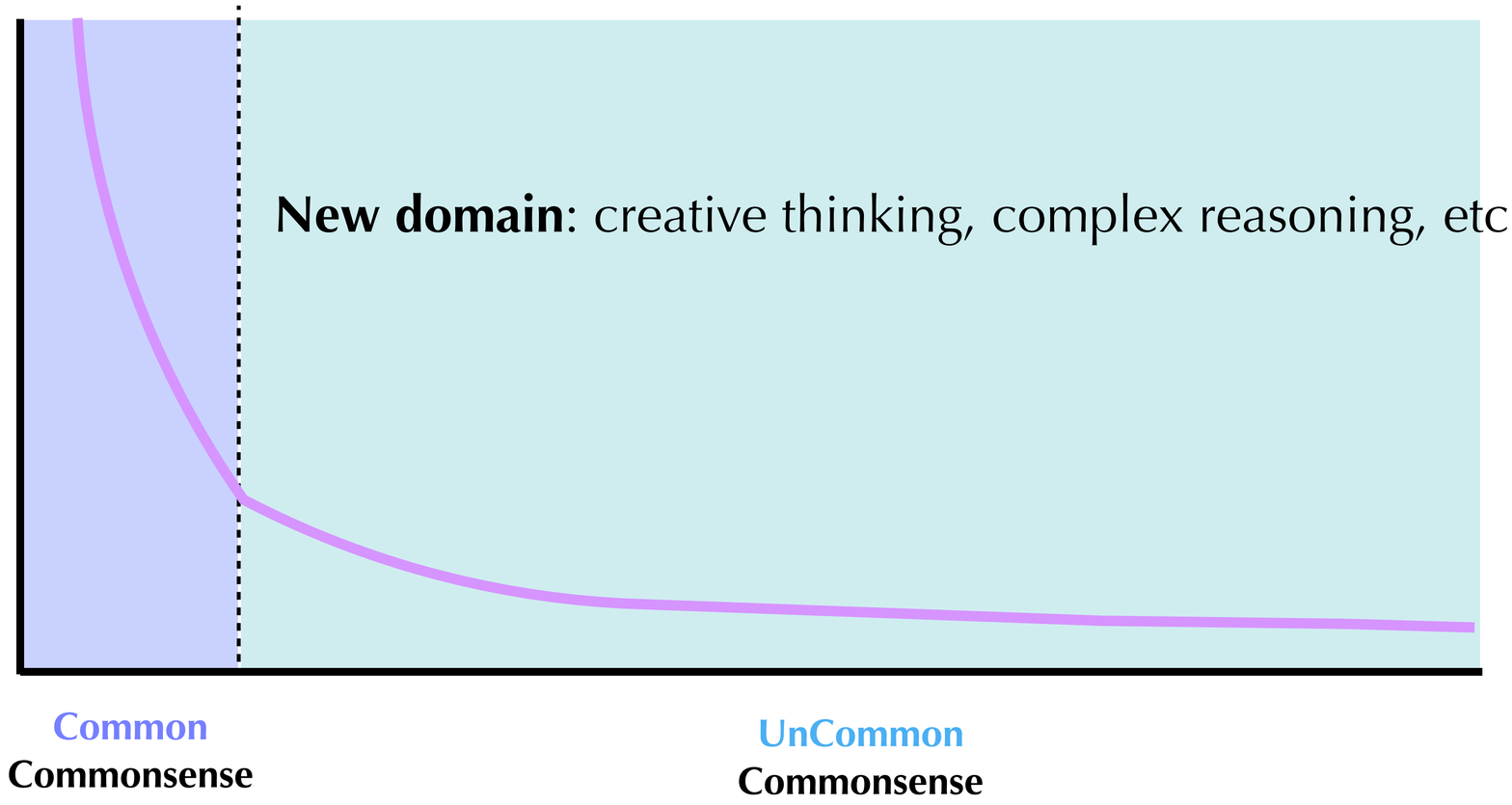
- The disagreement task is the hardest.
- The model can't predict when there is greater disagreement/subjectivity for specific instances.

Dimension	Model	Disagreement ↑ *R (p-value)*
Creativity (Creative-100)	LLaVA-7B	0.06 (.52)
	LLaVA-13B	*nan*
	InternVL2-8B	**0.20 (.05)**
	GPT-4v	-0.04 (.72)
Originality (Creative-100)	LLaVA-7B	0.07 (.49)
	LLaVA-13B	*nan*
	InternVL2-8B	0.11 (.27)
	GPT-4v	**0.15 (.13)**
Atypicality (Creative-100)	LLaVA-7B	0.17 (.08)
	LLaVA-13B	*nan*
	InternVL2-8B	**0.24 (.02*)**
	GPT-4v	-0.07 (.47)
Atypicality (Atypicality-300)	LLaVA-7B	0.01 (.92)
	LLaVA-13B	-0.05 (.43)
	InternVL2-8B	0.02 (.75)
	GPT-4v	-0.00 (.96)

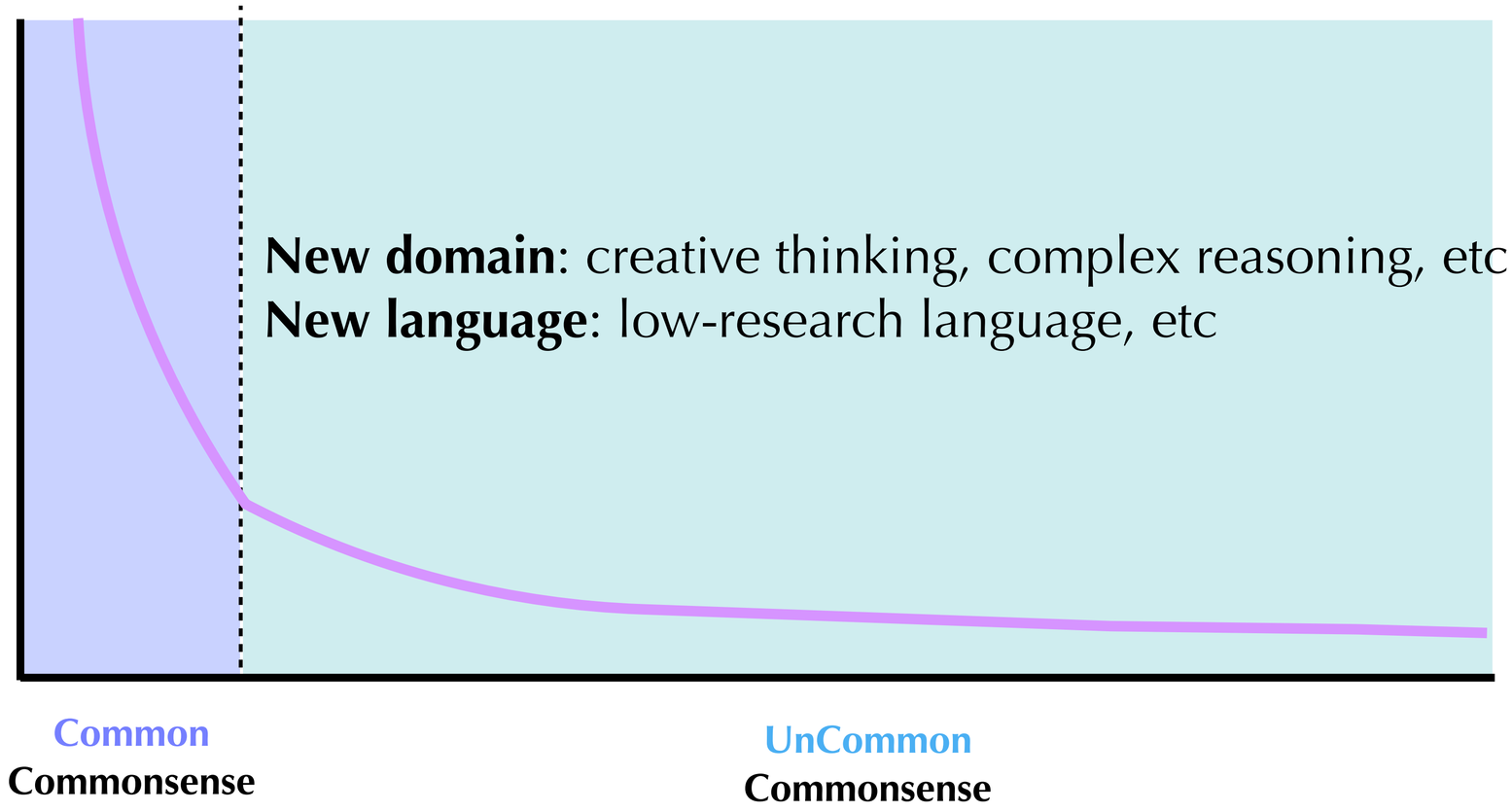
Why is performance so bad?



Why is performance so bad?

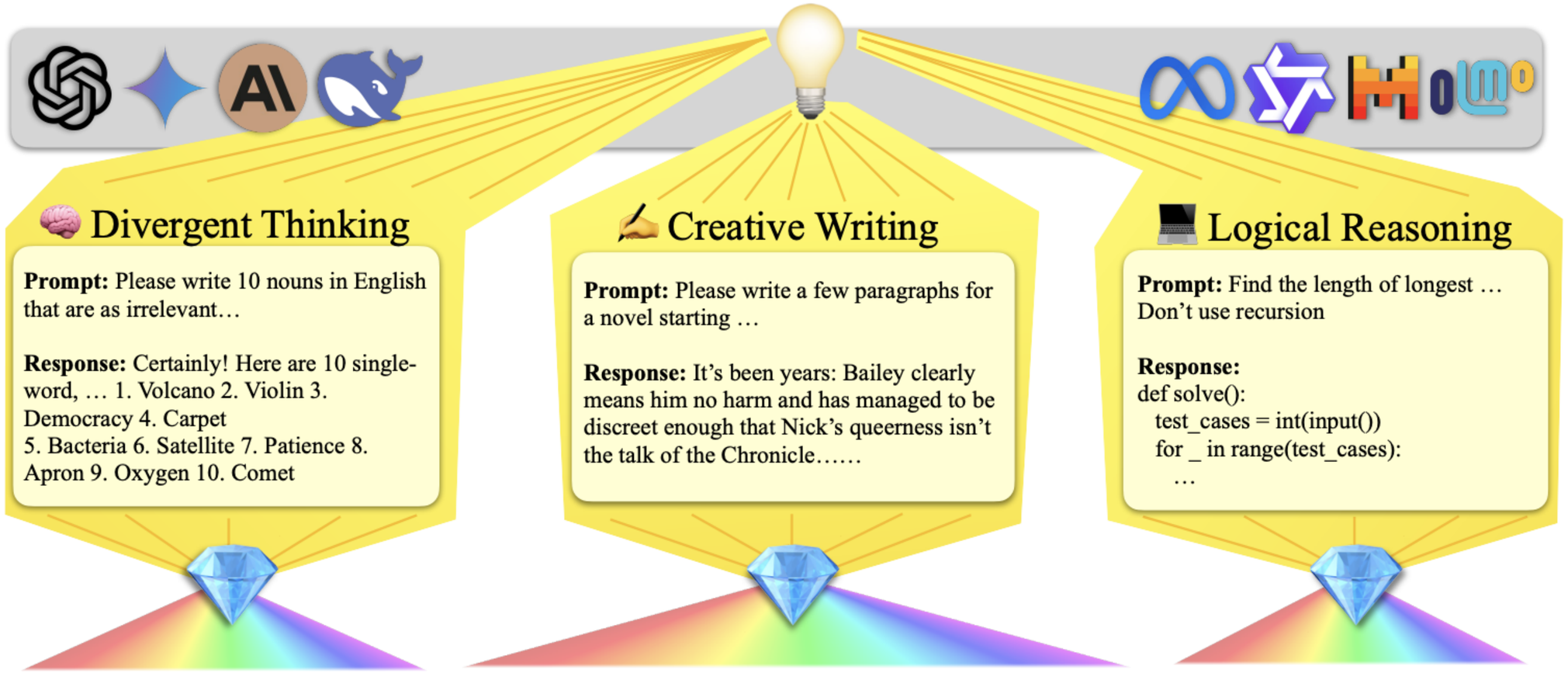


Why is performance so bad?



Creative Thinking

- New pre-prints!
- CreativityPrism: A Holistic Benchmark for Large Language Model Creativity



Outline

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

- Every Answer Matters: Evaluating Commonsense with Probabilistic Measures. [ACL 2024]
- Leveraging Large Models for Evaluating Novel Content: A Case Study on Advertisement Creativity. [EMNLP 2025]

Benchmark: Long-tail Question: Commonsense Reasoning Evaluation

- UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations. [NAACL 2024]
- In search of the long-tail: systematic generation of long-tail knowledge via logical rule guided search [EMNLP 2024]
- Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games. [EMNLP 2025]

New domain: creative thinking, complex reasoning, etc
New language: low-research language, etc

Outline

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

- Every Answer Matters: Evaluating Commonsense with Probabilistic Measures. [ACL 2024]
- Leveraging Large Models for Evaluating Novel Content: A Case Study on Advertisement Creativity. [EMNLP 2025]

Benchmark: Long-tail Question: Commonsense Reasoning Evaluation

Complex reasoning

- UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations. [NAACL 2024]
- In search of the long-tail: systematic generation of long-tail knowledge via logical rule guided search [EMNLP 2024]
- Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games. [EMNLP 2025]

Multi-lingual

Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games

César Guerra-Solano[1], Zhuochun Li[2], Xiang Lorraine Li[1]

CS Undergrad!

[1]Department of Computer Science,
[2]Department of Informatics and Networked Systems
University of Pittsburgh, PA, USA

University of Pittsburgh

EMNLP 2025
Suzhou, China 中国苏州

The New York Times *Connections*

Create four groups of four!

PINOT	CARD	SHOT	GLASS
ROSE	QUARTZ	DRIVE	THRUST
CLOUD	BOUNCE	PLATE	DISK
NAPKIN	FLEW	GALLEON	FORK

Mistakes Remaining: ● ● ● ●

Shuffle | Deselect All | Submit

Create four groups of four!

PARTS OF A TABLE SETTING
FORK, GLASS, NAPKIN, PLATE

KINDS OF DIGITAL STORAGE
CARD, CLOUD, DISK, DRIVE

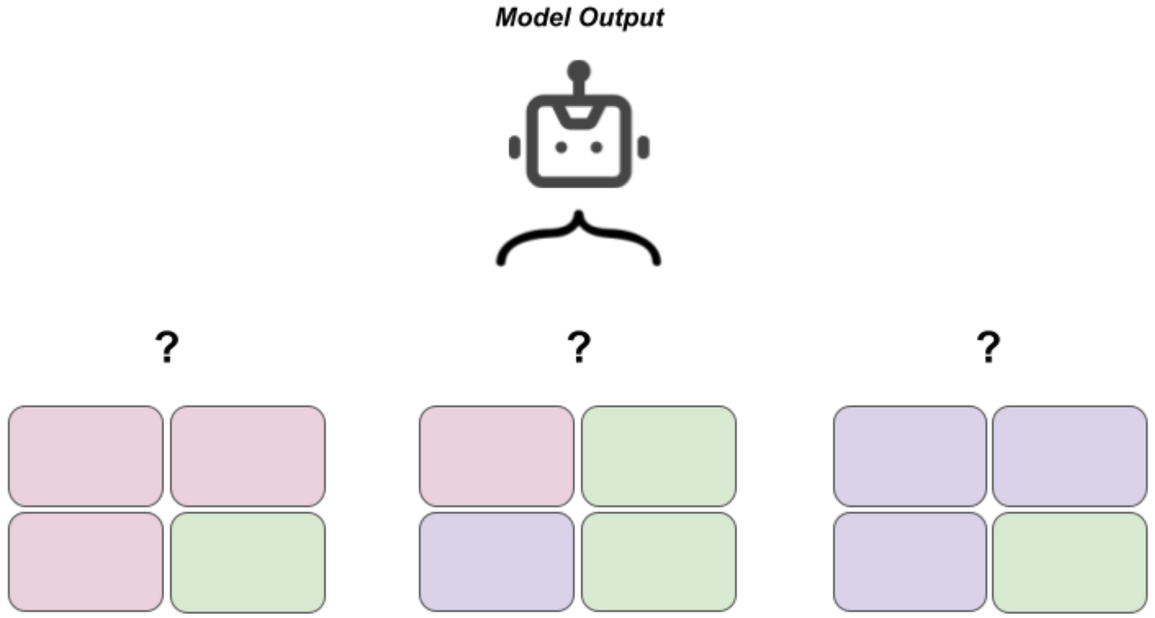
UNITS OF VOLUME PLUS LETTER
BOUNCE, GALLEON, PINOT, QUARTZ

INCREASED, WITH “UP”
FLEW, ROSE, SHOT, THRUST

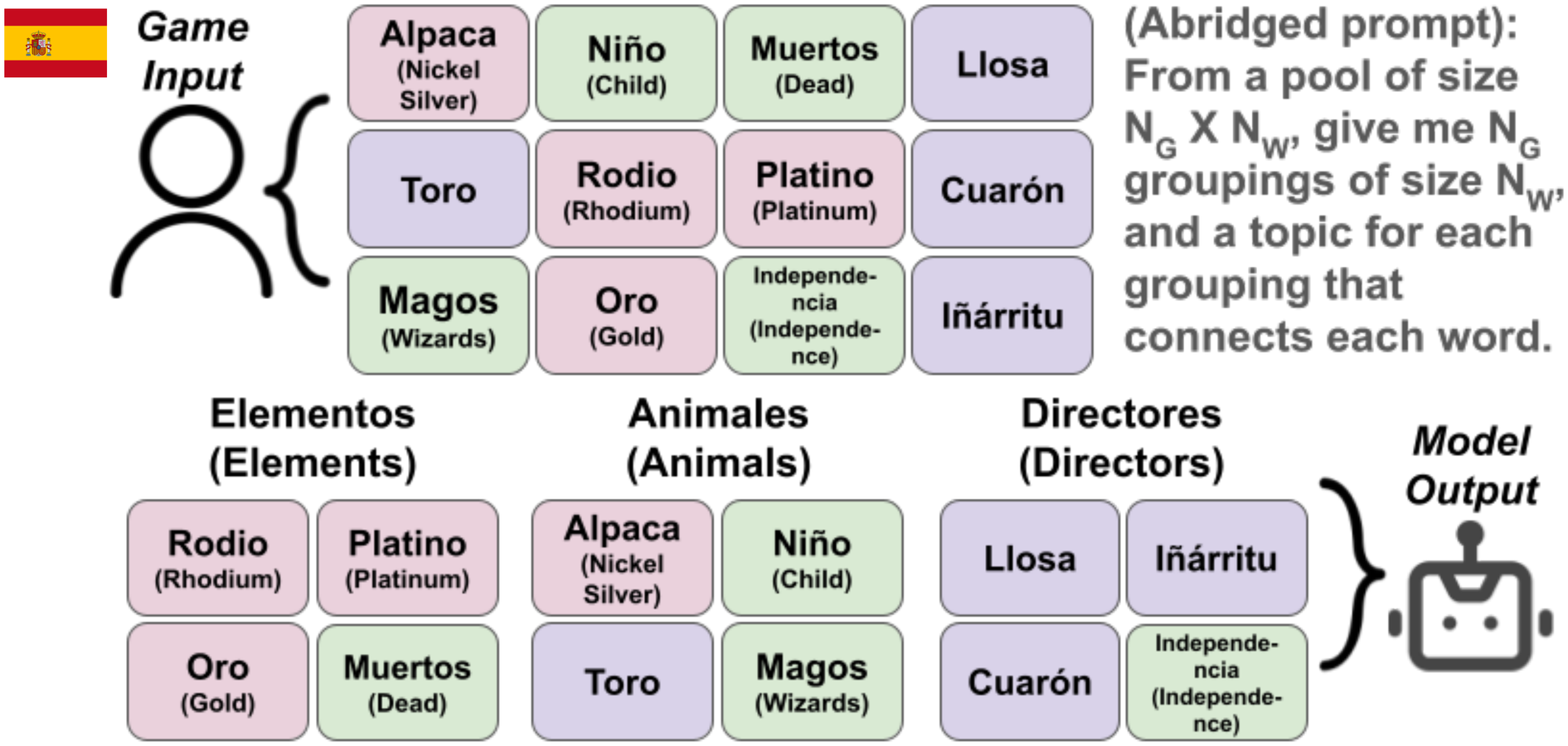
View Results

Connections as an Evaluation

- Work has explored *Connections* as an abstract reasoning evaluation
  - Lack of a clear strategy
  - Identifying/connecting properties of words
  - (Todd et al., 2024; Samadarshi et al., 2024)
- Lack of explainability with model performance
- Bias towards the English language && Western/North American culture



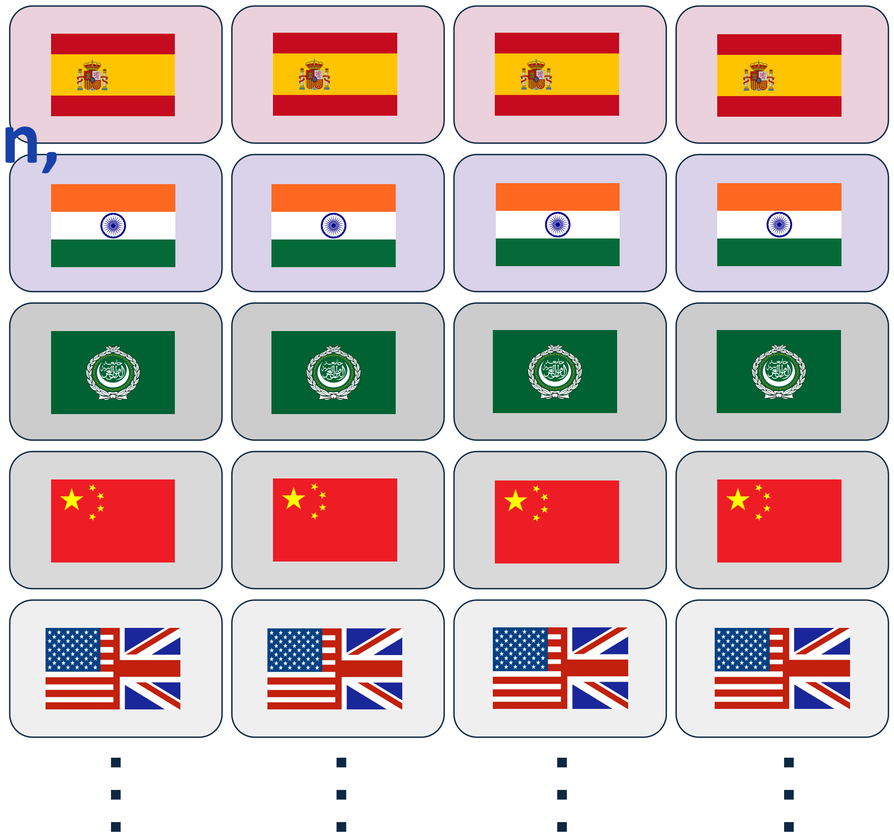
Global Group



Global Group

Native speaking annotators

Grouping creation, translation, tagging



Global Group

Sampling N_G groups, N_W words from each group

Game Input

Alpaca (Nickel Silver)	Niño (Child)	Muertos (Dead)	Llosa
Toro	Rodio (Rhodium)	Platino (Platinum)	Cuarón
Magos (Wizards)	Oro (Gold)	Independencia (Independence)	Iñárritu

(Abridged prompt): From a pool of size N_G X N_W, give me N_G groupings of size N_W, and a topic for each grouping that connects each word.

Elementos (Elements)		Animales (Animals)		Directores (Directors)	
Rodio (Rhodium)	Platino (Platinum)	Alpaca (Nickel Silver)	Niño (Child)	Llosa	Iñárritu
Oro (Gold)	Muertos (Dead)	Toro	Magos (Wizards)	Cuarón	Independencia (Independence)

Model Output

Global Group

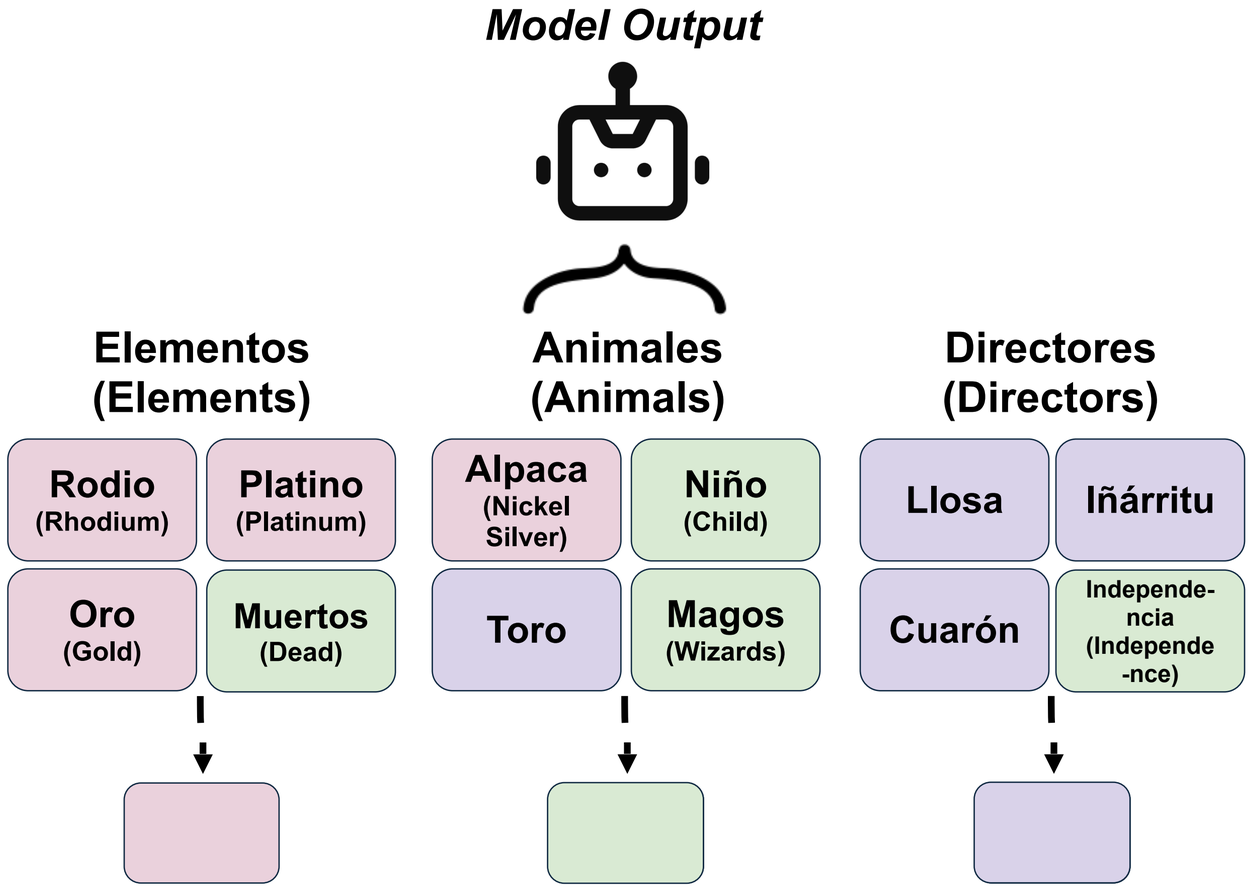
Dataset Background Subsets	Available Languages	Non-Culturally-Related Group Topic (in en)	Culturally-Related (C-R) Group Topic (in en)	C-R Group Amount	Total Group Amount	Game Settings ($N_{words} \times N_{groups}$)
en	en	Professions	American Holidays	22	48	$\{2,3,4\} \times \{2,3,4\}$
es	es, en	Aquatic Animals	Characters From Don Quixote	27	48	$\{2,3,4\} \times \{2,3,4\}$
zh	zh, en	Tree Composition	The Five Classics	35	80	$\{2,3,4\} \times \{2,3,4\}$
hi	hi, en	Parts of a Train	Fairy Tales	32	49	$\{2,3,4\} \times \{2,3,4\}$
ar	ar, en	Things That Spin	Traditional Food	19	40	$\{2,3,4\} \times \{2,3,4\}$
nyt-seq	en	Basic Directions	Starts of U.S. Presidents	N/A	2044	$4 \times 2, 4 \times 3, 4 \times 4$
nyt-shuf	en	Basic Directions	Starts of U.S. Presidents	N/A	2044	$\{2,3,4\} \times \{2,3,4\}$

Model Evaluation

- Evaluated 2 closed-source, 4 open-source models
  - GPT-3.5-Turbo, GPT-4
  - Llama3-8B, Llama3.1-70B
    - Size comparison
  - Mistral-7B
  - Aya-8B
    - Multilingual training paradigm

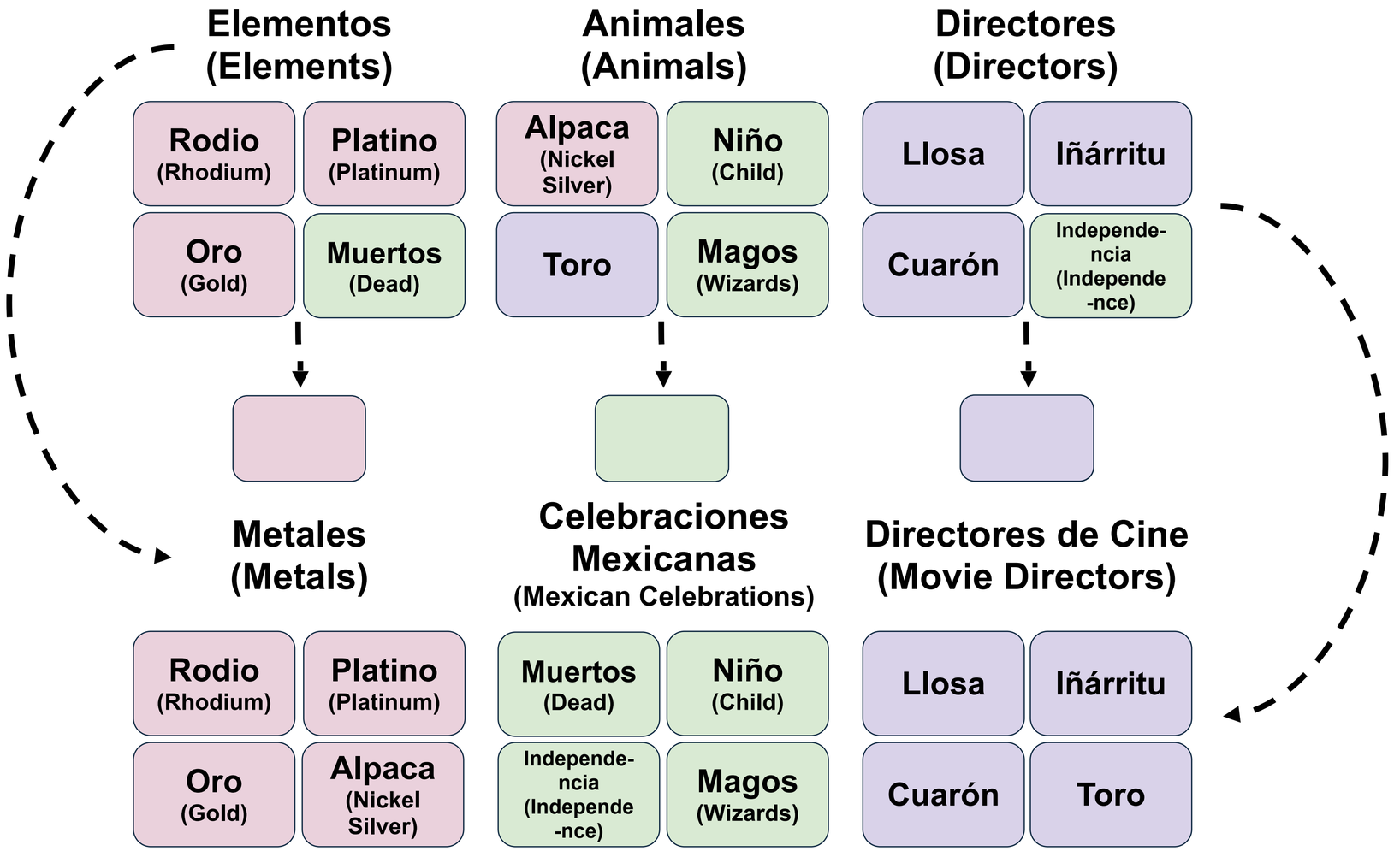
Model Evaluation

- Match attempts to true groups based on majority



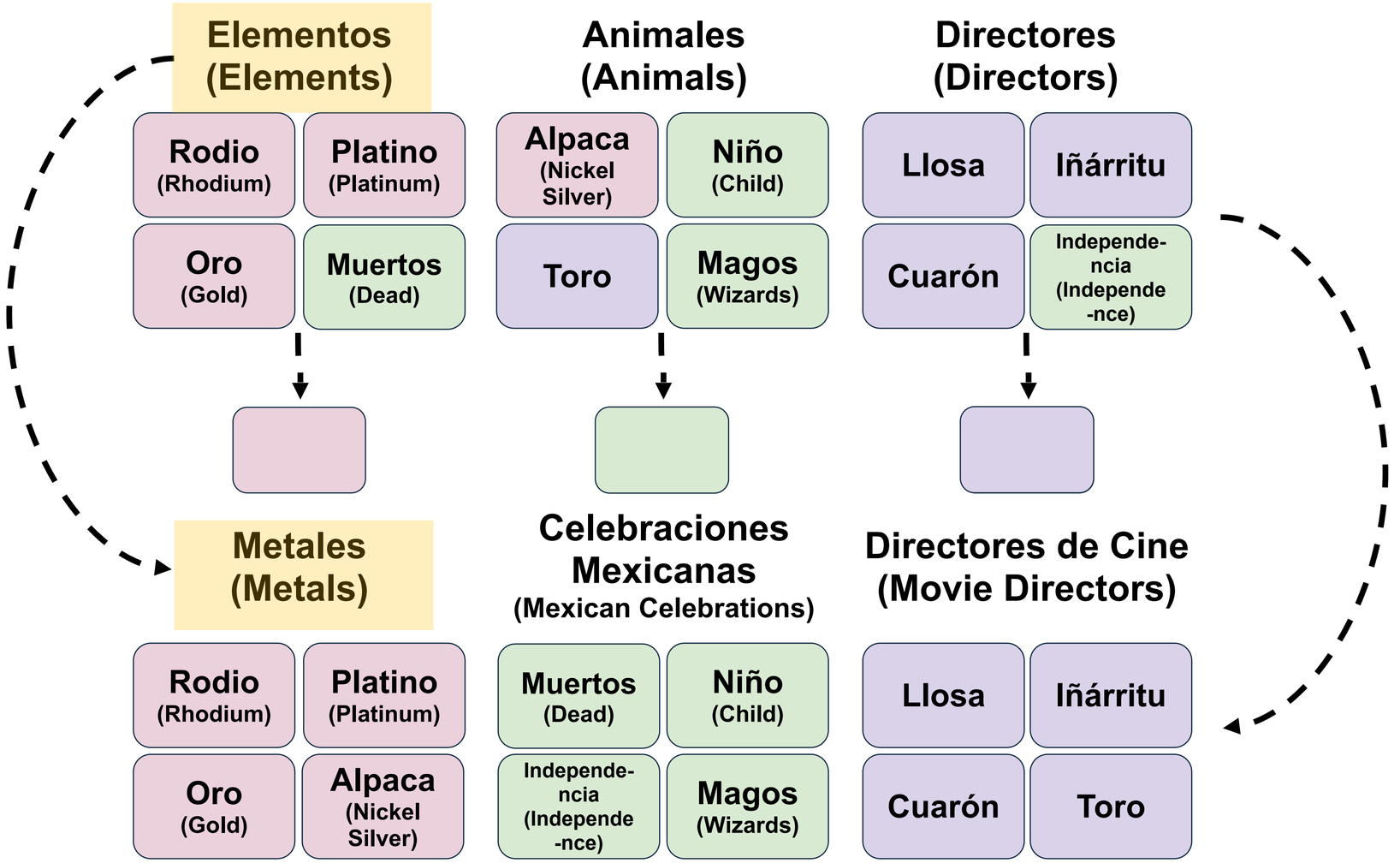
Model Evaluation

- Match attempts to true groups based on majority
- Scoring topics and groups:



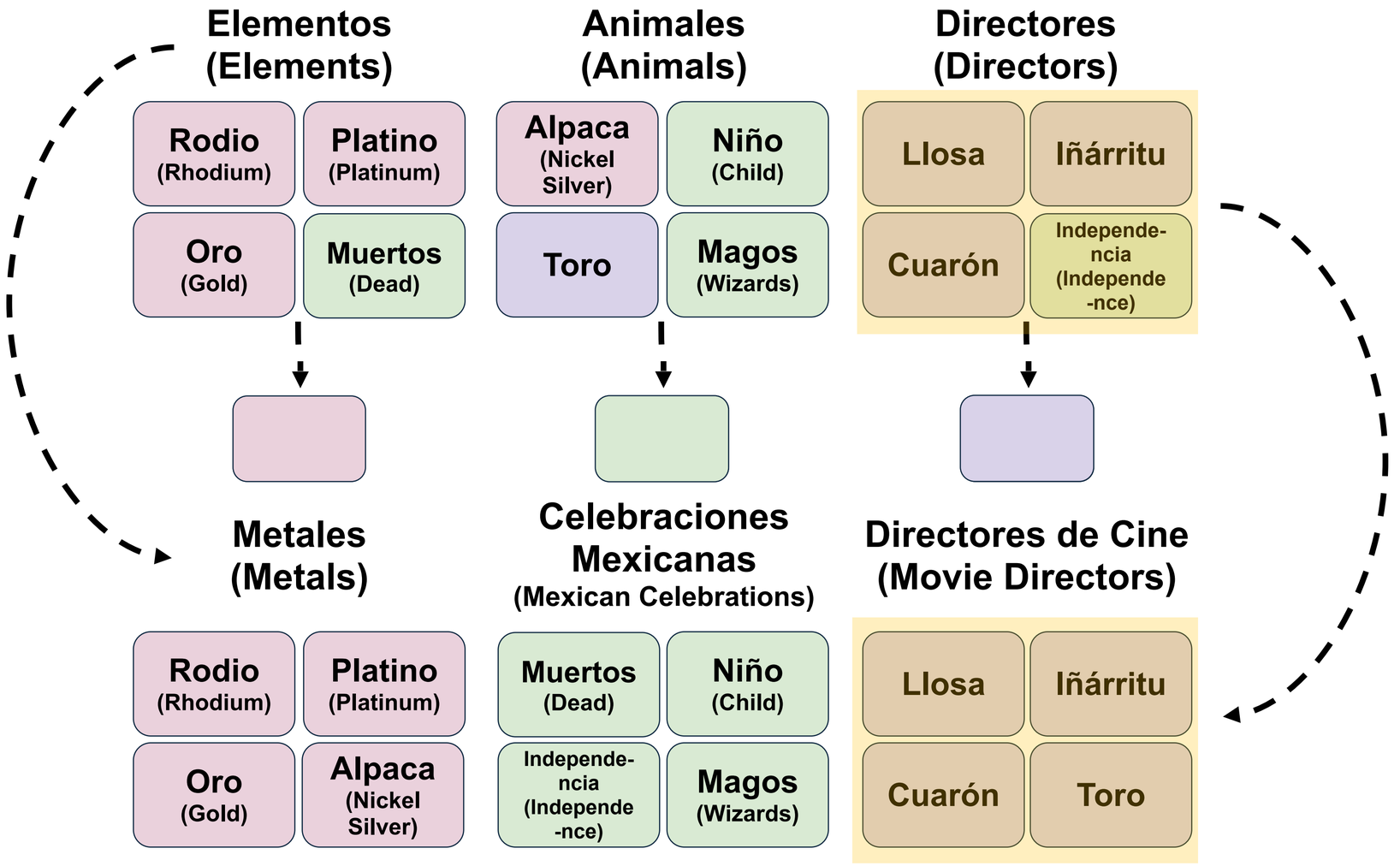
Model Evaluation

- Match attempts to true groups based on majority
- Scoring topics and groups:
  - **Topic Achieved Score**

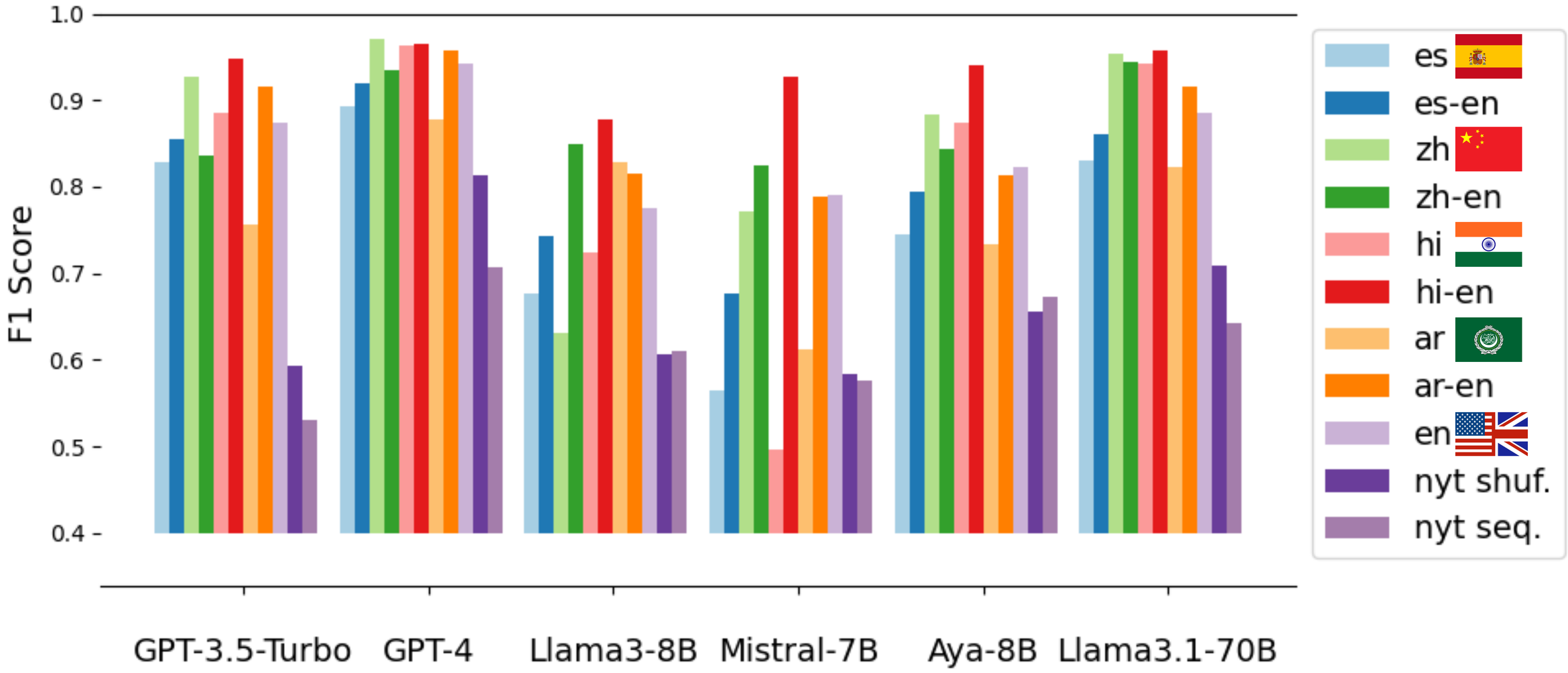


Model Evaluation

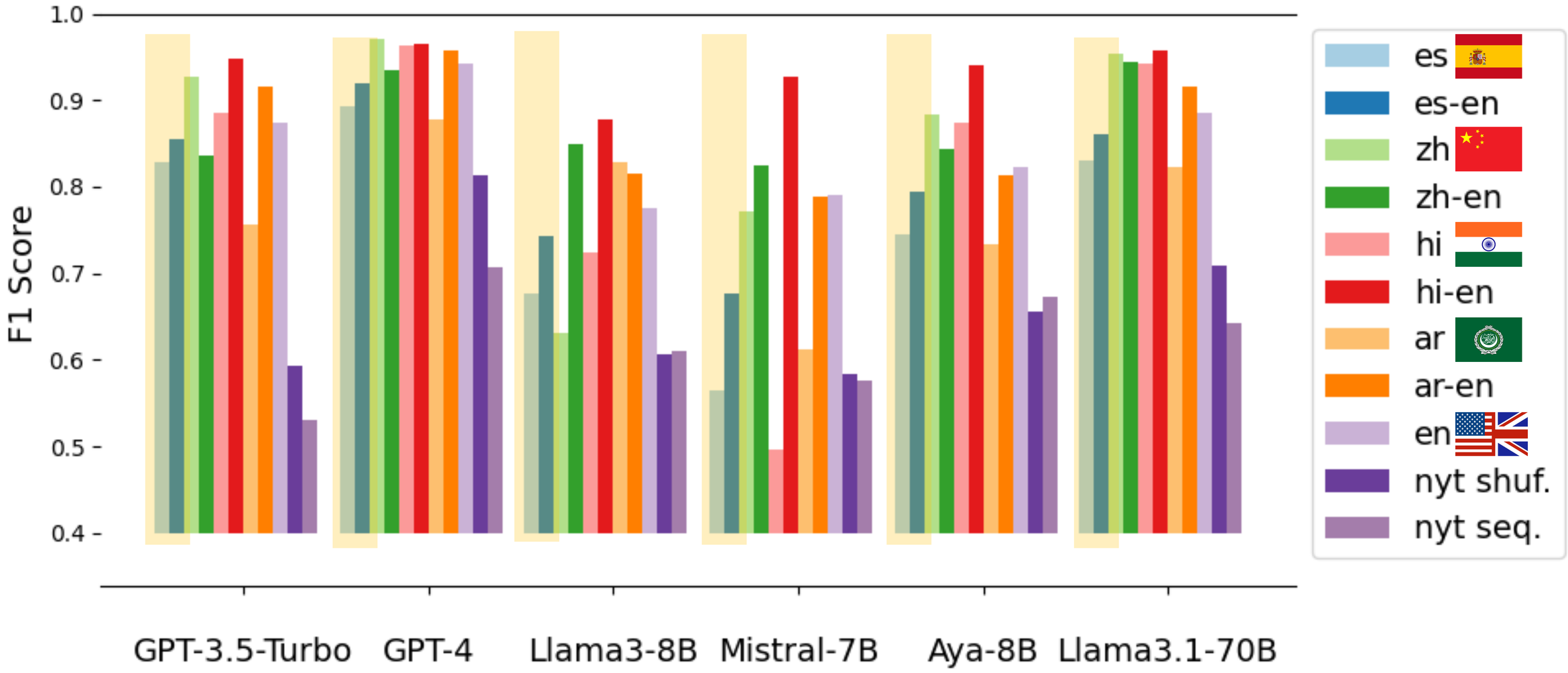
- Match attempts to true groups based on majority
- Scoring topics and groups:
  - **Topic Achieved Score**
  - **F1 Score**



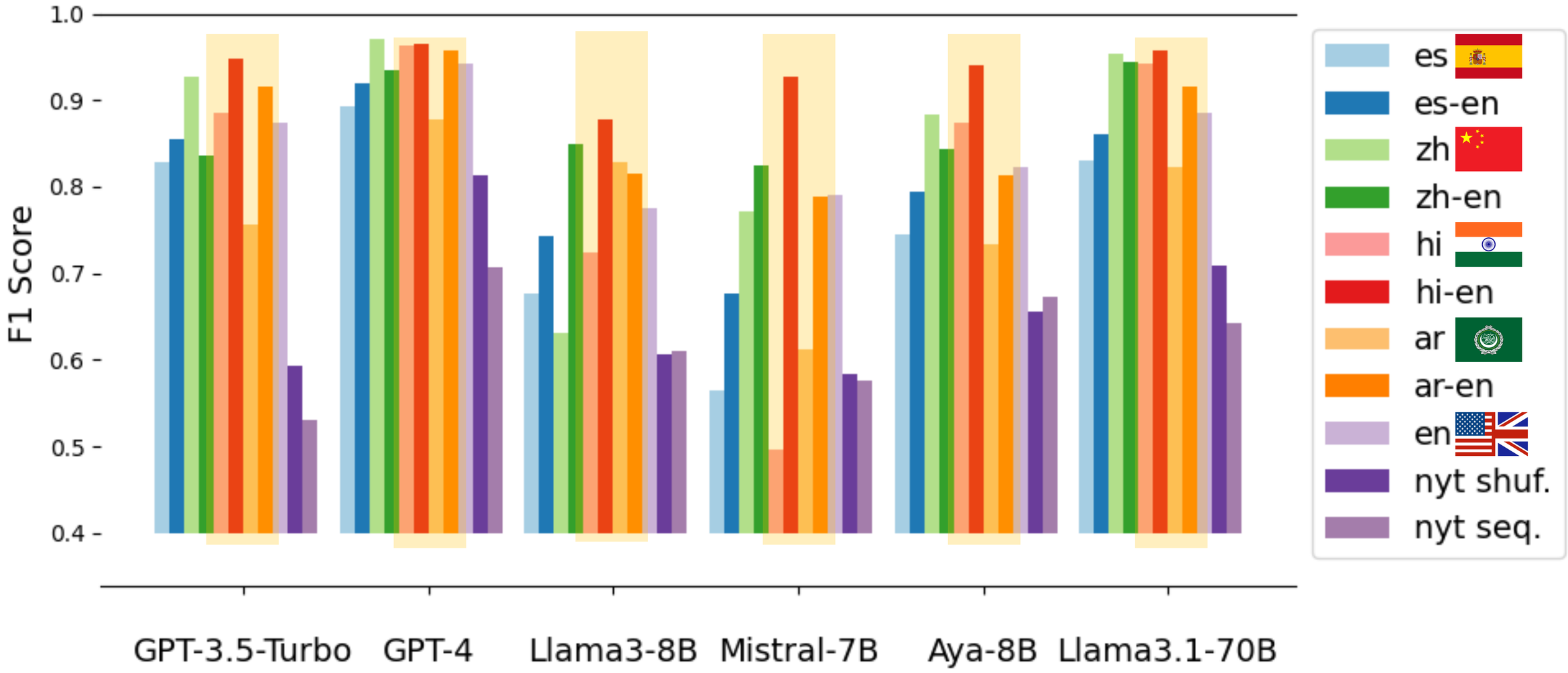
English Representations Improve Performance



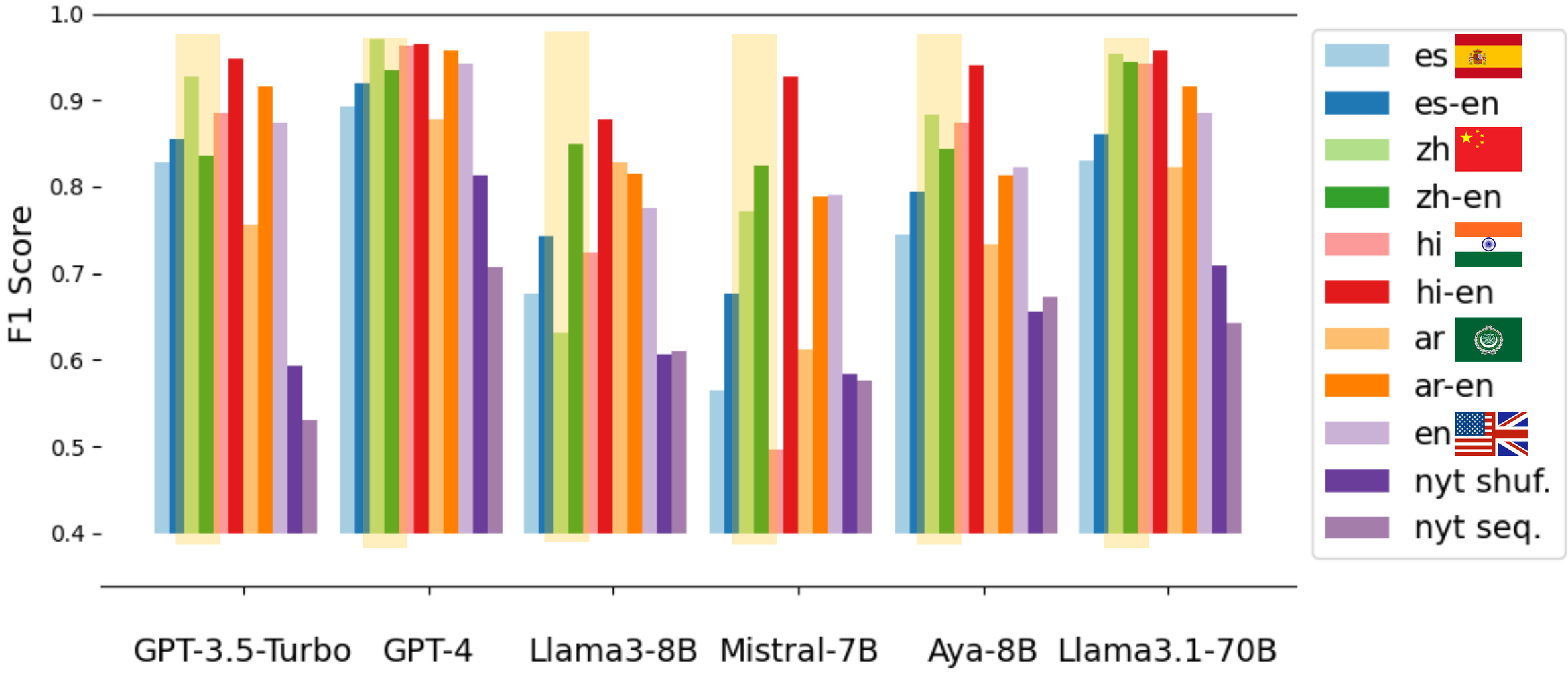
English Representations Improve Performance



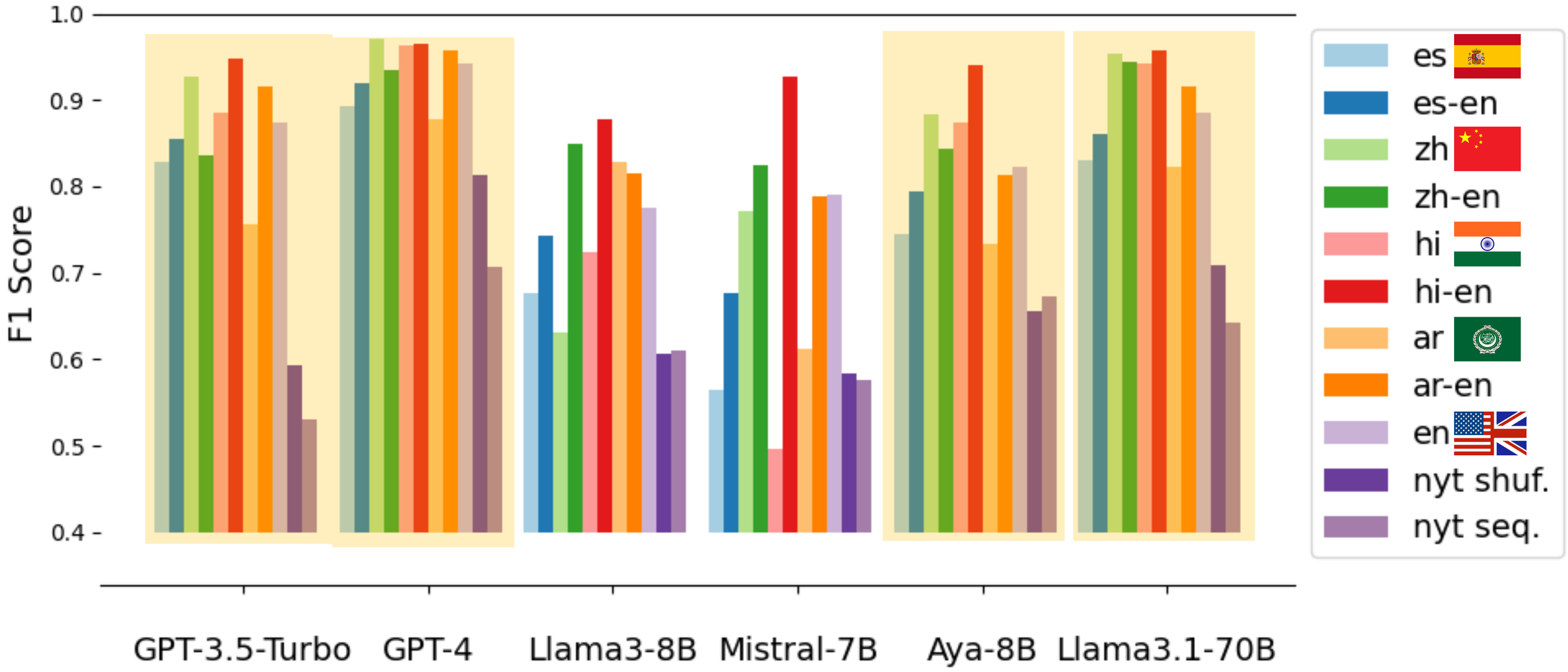
English Representations Improve Performance



English Representations Improve Performance

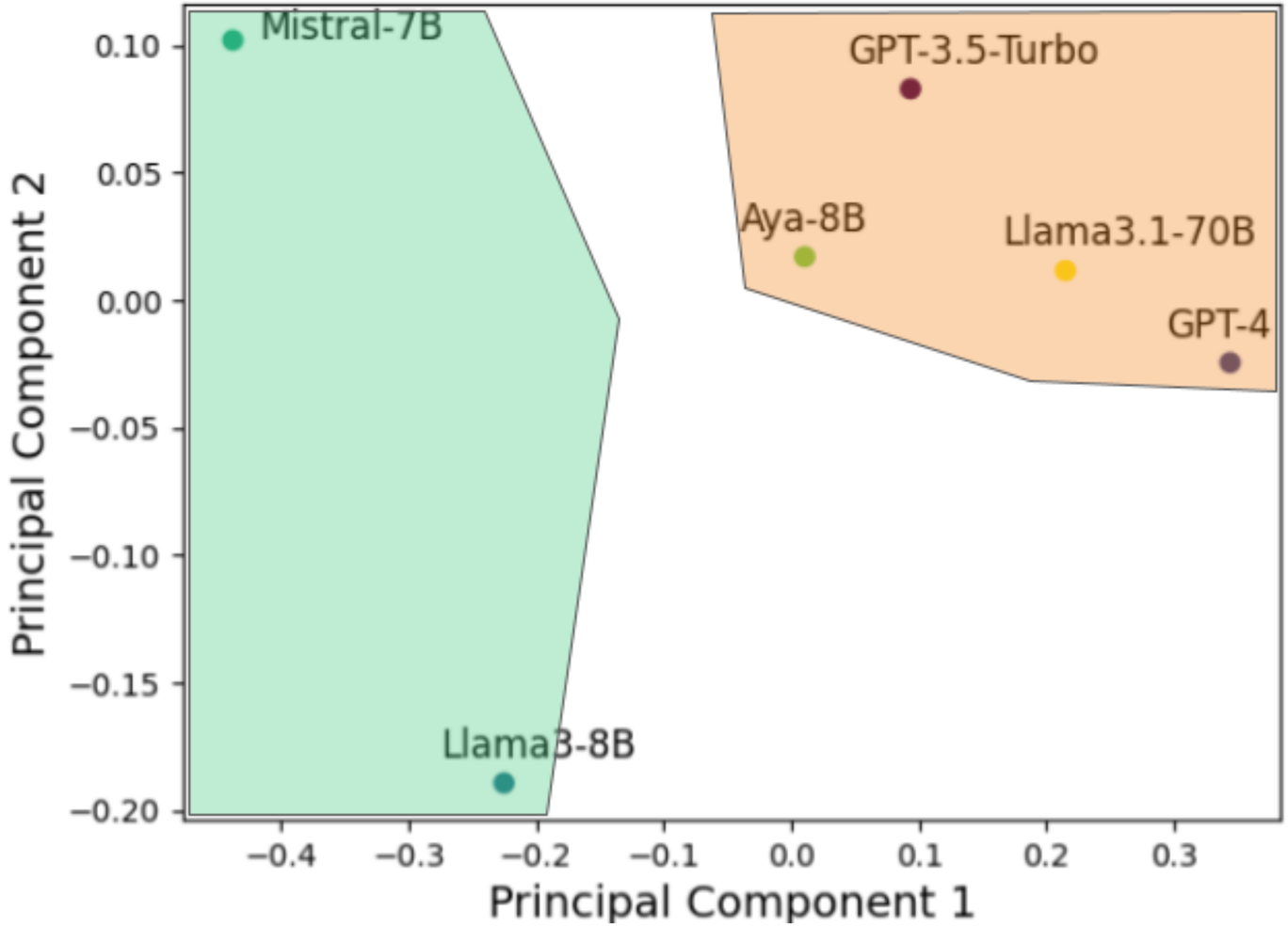


English Representations Improve Performance



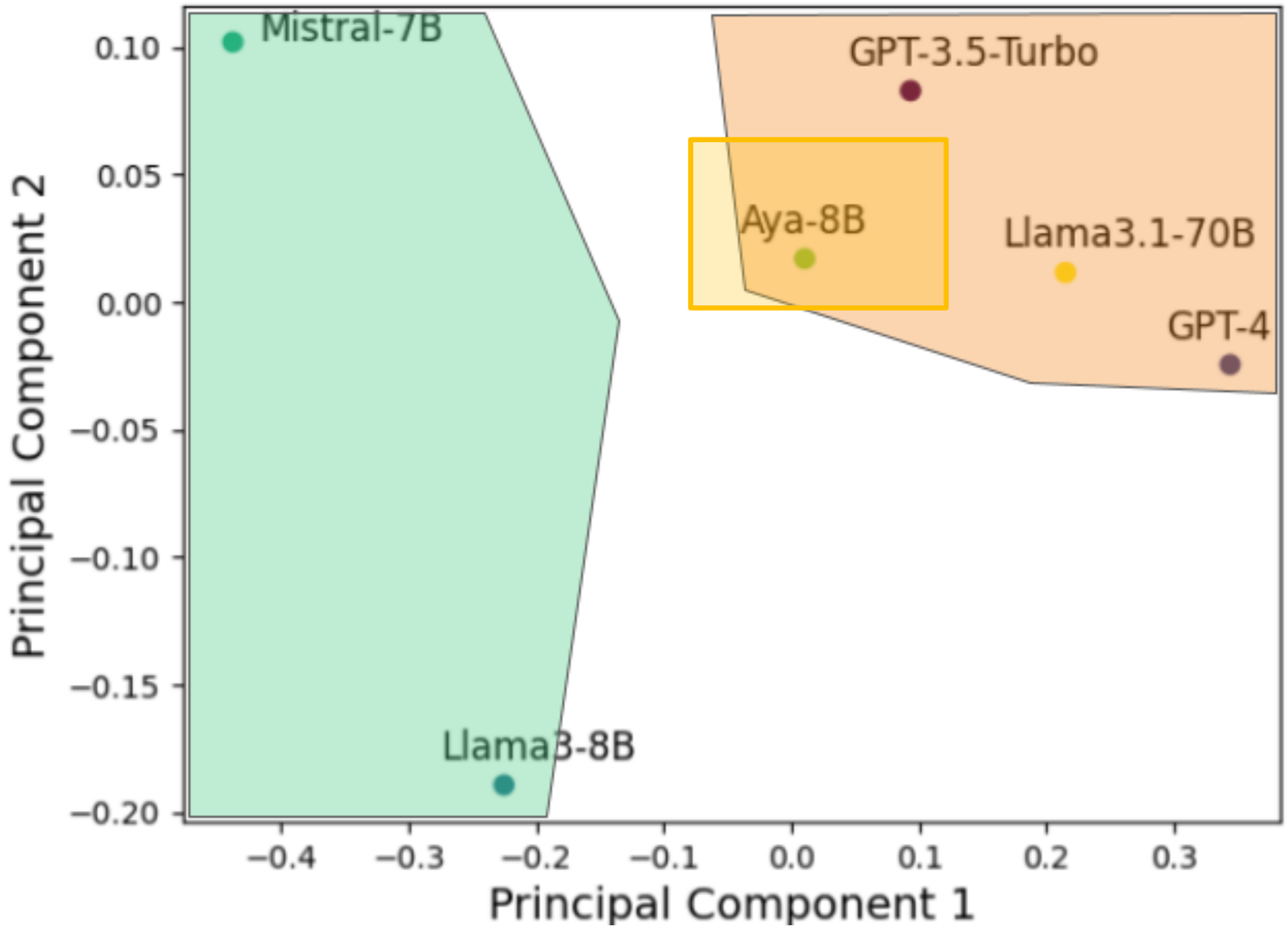
Multilingual Training Greatly Improves Small Models

- Represent models as vectors of dataset performance
  - Plot first two PCs after PCA



Multilingual Training Greatly Improves Small Models

- Represent models as vectors of dataset performance
  - Plot first two PCs after PCA
- Aya-8B **far smaller** than GPTs and Llama 70B
  - Yet performs very similarly!
  - Multilingual training



Key Game Parameters Correlate with Difficulty

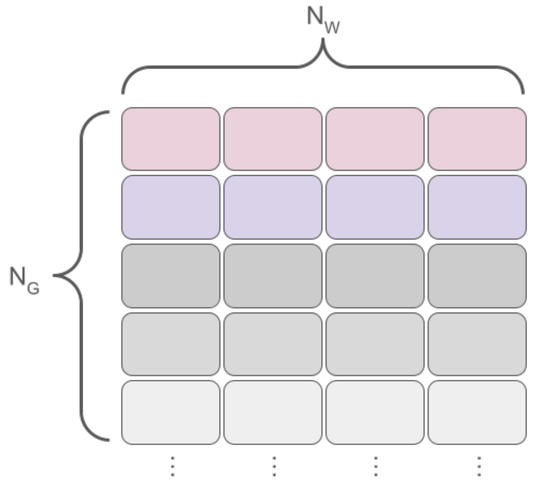
- We hypothesize, outside of language, **key game parameters are associated with difficulty**
  - Need for a controlled setting

Key Game Parameters Correlate with Difficulty

- We hypothesize, outside of language, **key game parameters are associated with difficulty**
  - Need for a controlled setting
- We consider three potential game parameters

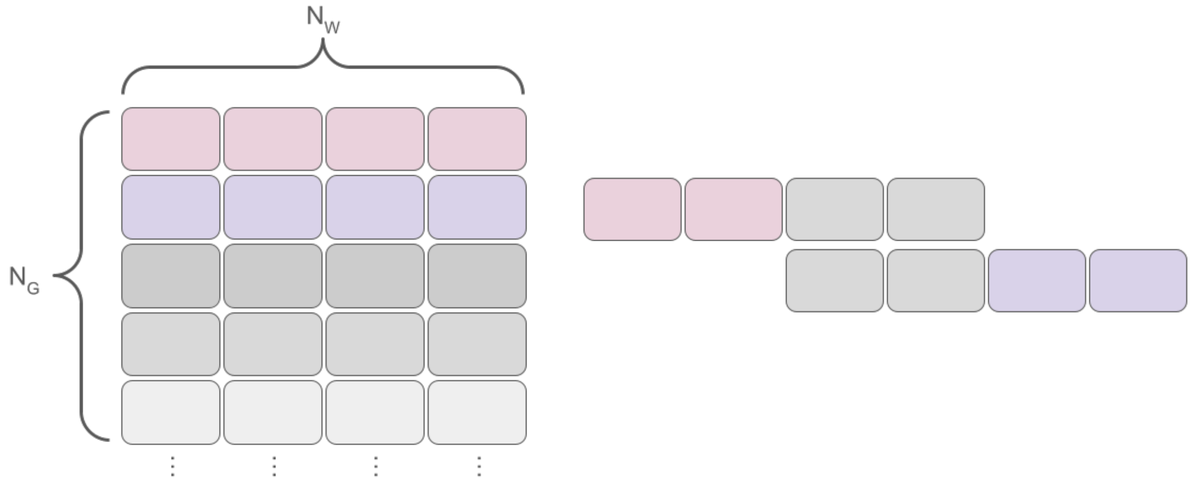
Key Game Parameters Correlate with Difficulty

- We hypothesize, outside of language, **key game parameters are associated with difficulty**
  - Need for a controlled setting
- We consider three potential game parameters
  - Game size



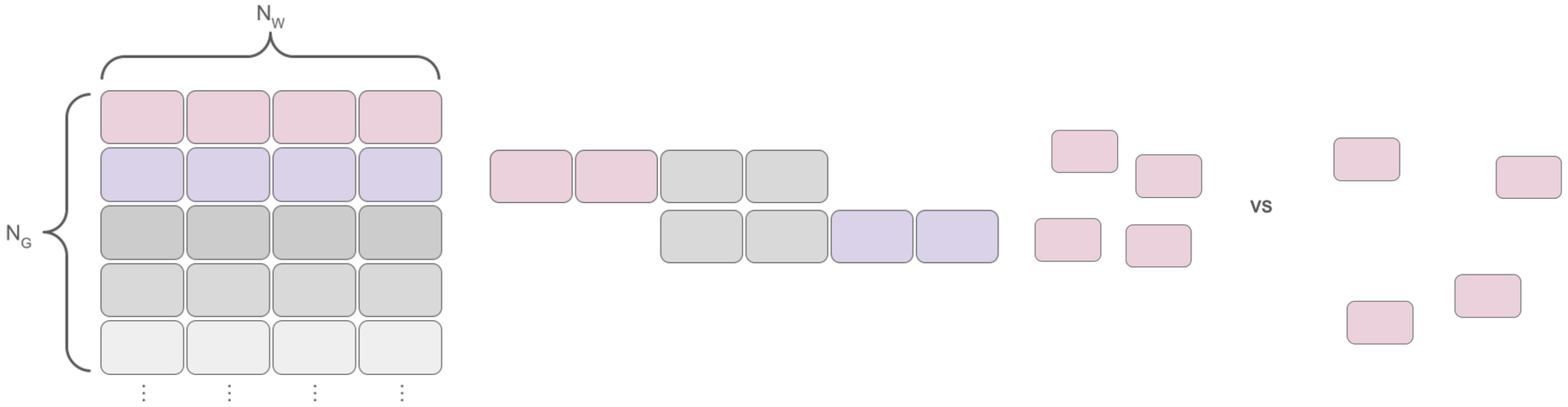
Key Game Parameters Correlate with Difficulty

- We hypothesize, outside of language, **key game parameters are associated with difficulty**
  - Need for a controlled setting
- We consider three potential game parameters
  - Game size
  - Word overlap between groupings



Key Game Parameters Correlate with Difficulty

- We hypothesize, outside of language, **key game parameters are associated with difficulty**
  - Need for a controlled setting
- We consider three potential game parameters
  - Game size
  - Word overlap between groupings
  - Semantic similarity of groupings



Key Game Parameters Correlate with Difficulty

Models	Average F1 Score			
Group Size	2	3	4	
GPT-3.5-Turbo	0.881	**0.890**	0.884	
GPT-4	0.925	0.940	**0.949**	
Group Count	2	3	4	
GPT-3.5-Turbo	**0.930**	0.883	0.842	
GPT-4	**0.962**	0.940	0.912	
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]	
GPT-3.5-Turbo	0.924	0.927	**0.951**	
GPT-4	0.969	0.974	**0.990**	
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831
GPT-4	**0.977**	0.923	0.863	0.859

Key Game Parameters Correlate with Difficulty

- Inconsistent trend with group size

Models	Average F1 Score			
Group Size	2	3	4	
GPT-3.5-Turbo	0.881	**0.890**	0.884	
GPT-4	0.925	0.940	**0.949**	
Group Count	2	3	4	
GPT-3.5-Turbo	**0.930**	0.883	0.842	
GPT-4	**0.962**	0.940	0.912	
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]	
GPT-3.5-Turbo	0.924	0.927	**0.951**	
GPT-4	0.969	0.974	**0.990**	
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831
GPT-4	**0.977**	0.923	0.863	0.859

Key Game Parameters Correlate with Difficulty

- Inconsistent trend with group size
- Great consistency with GC, ARI, Overlap

Models	**Average F1 Score**			
Group Size	2	3	4	
GPT-3.5-Turbo	0.881	**0.890**	0.884	
GPT-4	0.925	0.940	**0.949**	
Group Count	2	3	4	
GPT-3.5-Turbo	**0.930**	0.883	0.842	
GPT-4	**0.962**	0.940	0.912	
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]	
GPT-3.5-Turbo	0.924	0.927	**0.951**	
GPT-4	0.969	0.974	**0.990**	
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831
GPT-4	**0.977**	0.923	0.863	0.859

Key Game Parameters Correlate with Difficulty

- Inconsistent trend with group size
- Great consistency with GC, ARI, Overlap
  - Greater GC = lower performance

Models	Average F1 Score			
Group Size	2	3	4	
GPT-3.5-Turbo	0.881	**0.890**	0.884	
GPT-4	0.925	0.940	**0.949**	
Group Count	2	3	4	
GPT-3.5-Turbo	**0.930**	0.883	0.842	
GPT-4	**0.962**	0.940	0.912	
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]	
GPT-3.5-Turbo	0.924	0.927	**0.951**	
GPT-4	0.969	0.974	**0.990**	
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831
GPT-4	**0.977**	0.923	0.863	0.859

Key Game Parameters Correlate with Difficulty

- Inconsistent trend with group size
- Great consistency with GC, ARI, Overlap
  - Greater GC = lower performance
  - Greater ARI = higher performance

Models	Average F1 Score			
Group Size	2	3	4	
GPT-3.5-Turbo	0.881	**0.890**	0.884	
GPT-4	0.925	0.940	**0.949**	
Group Count	2	3	4	
GPT-3.5-Turbo	**0.930**	0.883	0.842	
GPT-4	**0.962**	0.940	0.912	
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]	
GPT-3.5-Turbo	0.924	0.927	**0.951**	
GPT-4	0.969	0.974	**0.990**	
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831
GPT-4	**0.977**	0.923	0.863	0.859

Key Game Parameters Correlate with Difficulty

- Inconsistent trend with group size
- Great consistency with GC, ARI, Overlap
  - Greater GC = lower performance
  - Greater ARI = higher performance
  - Greater Overlap = lower performance

Models	Average F1 Score			
Group Size	2	3	4	
GPT-3.5-Turbo	0.881	**0.890**	0.884	
GPT-4	0.925	0.940	**0.949**	
Group Count	2	3	4	
GPT-3.5-Turbo	**0.930**	0.883	0.842	
GPT-4	**0.962**	0.940	0.912	
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]	
GPT-3.5-Turbo	0.924	0.927	**0.951**	
GPT-4	0.969	0.974	**0.990**	
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831
GPT-4	**0.977**	0.923	0.863	0.859

Performance is Related to Group Content

Models	**Mean F1 Score in Groups of Tag**		
Group Tag	General Knowledge	Cultural and Pop Culture Knowledge	Linguistic
GPT-3.5-Turbo	**0.902**	0.871	0.826
GPT-4	**0.954**	0.937	0.916
Llama3-8B	**0.804**	0.767	0.724
Llama3.1-70B	**0.935**	0.900	0.859
Mistral-7B	**0.779**	0.694	0.669
Aya-8B	**0.866**	0.818	0.789

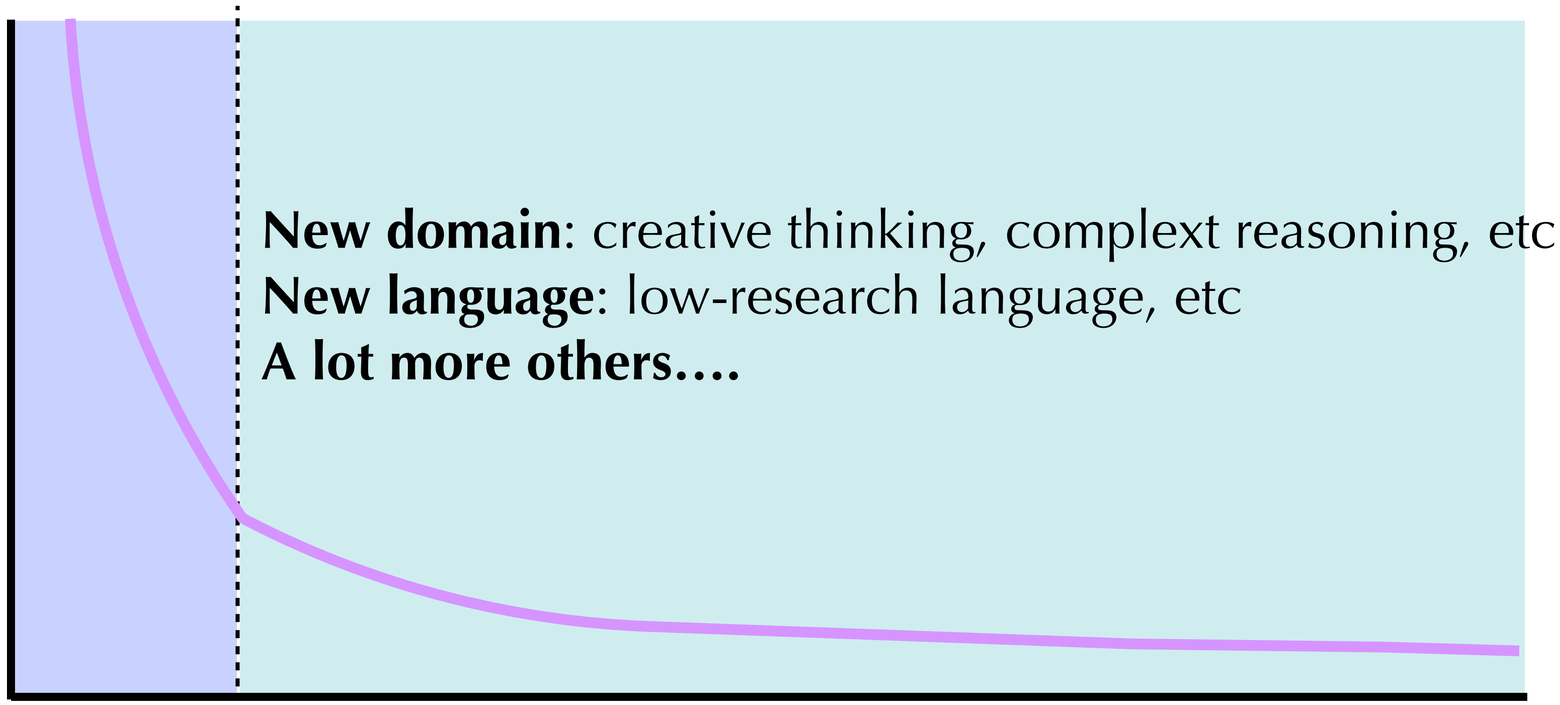
Takeways

- **Abstract reasoning performance is language-dependent**, with English modalities leading to the best performance

- We show that this performance is associated with **game parameters and group content**, allowing controlled comparison

- We demonstrate the value of a multilingual-focused training paradigm, letting **small models perform on-par with far larger models**

Takeaways from the talk



Summary

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

- Every Answer Matters: Evaluating Commonsense with Probabilistic Measures. [ACL 2024]
- Leveraging Large Models for Evaluating Novel Content: A Case Study on Advertisement Creativity. [EMNLP 2025]

Benchmark: Long-tail Question: Commonsense Reasoning Evaluation

- UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations. [NAACL 2024]
- In search of the long-tail: systematic generation of long-tail knowledge via logical rule guided search [EMNLP 2024]
- Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games. [EMNLP 2025]

Thanks to the amazing students!

Joey Hou

Qi Cheng

Cesar Geurra-Solano

Zhuochun Li