

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

Xiang Lorraine Li

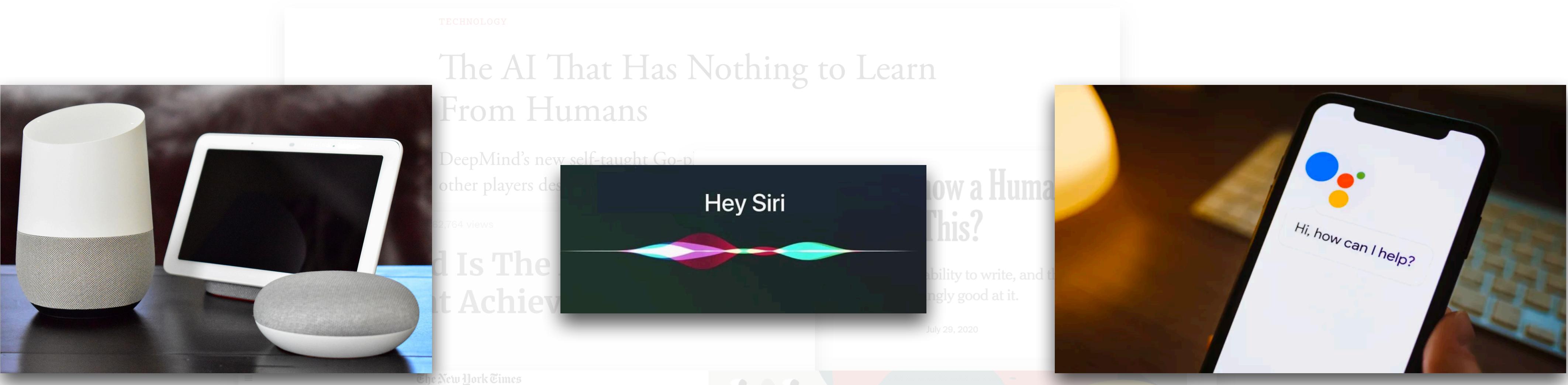
Assistant Professor at SCI Pitt

UMassAmherst

Robert and Donna Manning  
College of Information  
& Computer Sciences



# Impressive Progress in AI



# Impressive Progress in AI

For a traditional Chinese wedding, there are many specific customs, but here are some key ones:

## 1. Women:

- Traditional Option: It's common for women to wear a *cheongsam* with a wide red collar. It can be embellished with intricate embroidery.
- Modern Option: If the couple prefers something more modern, a well-tailored suit or a well-tailored dress can be chosen, reserved for the bride.
- Colors to Consider: Red is the color of good fortune and happiness, so wearing red is also favorable.



Diverse communities have different customs and traditions. Models need to understand and respect these differences to provide accurate and useful information that truly serves the needs of all users.

**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 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.

# 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.

# 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. Boiled water is too hot to drink.

Heat is needed to boil water.

Burner can provide heat.

Boiled water can cook food.

# What is Common Sense?

**Shared**

Water is liquid.

Water can be found in river.

Humans drink water.

**Implicit**

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. Boiled water is too hot to drink.

Heat is needed to boil water.

Burner can provide heat.

Boiled water can cook food.

# Why is Common Sense Challenging?

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. Boiled water is too hot to drink.

Heat is needed to boil water.

Burner can provide heat.

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.

Water is liquid.

Open the jelly jar.

can provide heat. Human needs water to live.

Humans drink water. Water needs to be held in a container.

Heat is needed to boil water.

Heat is needed to boil water.

Boiled water can cook food.

Boiled water is too hot to drink.

Water can be used for cleaning.

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.

Opening a jar needs tool

A knife with peanut butter could be the tool. Human can put peanut butter on the bread

Spread the peanut butter on the bread.

People who wants to lose weight usually avoid peanut butter.

Sweet water tastes good

Peanut butter is high calorie food.

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. Bread with peanut butter can be satisfying.

Water can be used for cleaning.

Sugar can melt in water

They boiled the water, then added sugar.

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.

Water can be used for

Water needs to be held in a container.

People who wants to lose weight usually avoid peanut butter.

Sweet water tastes good

Sugar water is wet. There is water in the river

River water is not directly drinkable.

Water needs to be h

Water needs to be held in a container.

Water can be used for cleaning.

Water needs to be held in a container.

Water can be used for

Water needs to be held in a container.

Water can be used to wash clothes.

Human feel satisfied after having sweet stuff.

Water needs to be held in a container.

Water needs to be held in a container.

Boiled water is too hot to drink.

Most bread is not sweet

Water needs to be held in a container.

Water needs to be held in a container.

Boiled water is too hot to drink.

Sugar water is also liquid.

Water needs to be held in a container.

Water needs to be held in a container.

Water can be used to wash clothes.

Water can be used for cleaning.

Water needs to be held in a container.

Water needs to be held in a container.

They boiled the water.

Water needs to be held in a container.

# Why is Common Sense Challenging?

## Massive

Food Chemistry  
Volume 303, 15 January 2020, 125385

Melatonin treatment maintains nutraceutical properties of pomegranate fruits during cold storage

Morteza Soleimani Aghdam <sup>a</sup>, Zisheng Luo <sup>b</sup>, Li Li <sup>b</sup>, Abbasali Jannatizadeh <sup>a</sup>, Javad Rezapour Fard <sup>c</sup>, Farhad Pirzad <sup>d</sup>

Show more ▾

+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.foodchem.2019.125385>

Get rights and content

Highlights

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

Donald T business, Jan. 3, 2022

The Ne subpoena business children, The inv Trump, Trump genera his chil

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.

PA MEDIA

IF NOT YOU, WHO? IF NOT NOW, WHEN? COP26 ACT NOW!

Article

Talk

## COVID-19 pandemic

From Wikipedia, the free encyclopedia

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The novel virus was first identified in December 2019 in Wuhan, China.

What do scientists think this week?

## Nutraceutical properties of lycopene

[Article in Spanish]  
Krzysztof N Waliszewski <sup>1</sup>, Gabriela Blasco

Affiliations + expand

PMID: 20485889 DOI: 10.1590/s0036-36342010000300010 Paper

They boiled the water, then added sugar.

### Abstract

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

DNA vaccine encoding Middle East respiratory syndrome coronavirus (MERS-CoV) spike protein induces protective immunity and virulence.

and widely distributed in various countries since December 2020. Other recommended preventive measures include social distancing, self-isolation for those who have been exposed or are symptomatic.

### Abstract

Principal component analysis is a widely used method for the analysis of a given data set in a high-dimensional Euclidean space. Here we present two analogues of principal component analysis in the setting of phylogenetic trees. In one approach, we study the Stiefel tropical linear space of the data points in the tropical projective torus; in the other a

Hang Chi <sup>a</sup>, Xueying Zheng <sup>a,b</sup>, Xiwen Wang <sup>a</sup>, Chong Wang <sup>a</sup>, Hualei Wang <sup>a,c</sup>, Weiwei Gai <sup>a</sup>, Stanley Perlman <sup>a</sup>, Songtao Yang <sup>a,d,e</sup>, Jinjun Zhao <sup>a,f</sup>, Xianzhu Xia <sup>a,g,h</sup>

<sup>a</sup>Key Laboratory of Jinan Province for Zoonosis Prevention and Control, Institute of Military Veterinary, Academy of Military Medical Science, Changchun, China

<sup>b</sup>School of Public Health, Shandong University, Jinan, China

<sup>c</sup>Jiangsu Co-Innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou, China

<sup>d</sup>State Key Laboratory of Respiratory Diseases, Guangzhou Institute of Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

<sup>e</sup>State Key Laboratory of Respiratory Diseases, Guangzhou Institute of Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

<sup>f</sup>State Key Laboratory of Respiratory Diseases, Guangzhou Institute of Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

<sup>g</sup>Department of Microbiology, College of Veterinary Medicine, Jilin University, Changchun, China

<sup>h</sup>Department of Clinical Laboratory, College of Medicine, Jilin University, Changchun, China

© 2017 Elsevier Ltd. All rights reserved.

The Middle East respiratory syndrome coronavirus (MERS-CoV) is an emerging pathogen that can cause outbreaks in the Arabian peninsula and in travellers from this region, raising the concern of a global pandemic could occur. Here, we show that a DNA vaccine encoding the first 25 amino acids of the MERS-CoV spike (S) protein, an SARS-specific functional protein, can induce a strong cellular immune response with a high titer of neutralizing antibodies ( $10^5$ – $10^6$ ) without adjuvant. DNA vaccination with the MERS-CoV S1 gene markedly increased the frequencies of CD4<sup>+</sup> and CD8<sup>+</sup> T cells secreting IFN- $\gamma$  and other cytokines, including IL-2, TNF- $\alpha$ , IL-10, IL-12, IL-17, IL-21, IL-22, IL-23, IL-27, IL-31, IL-35, IL-36, IL-37, IL-39, IL-40, IL-41, IL-42, IL-43, IL-44, IL-45, IL-46, IL-47, IL-48, IL-49, IL-50, IL-51, IL-52, IL-53, IL-54, IL-55, IL-56, IL-57, IL-58, IL-59, IL-60, IL-61, IL-62, IL-63, IL-64, IL-65, IL-66, IL-67, IL-68, IL-69, IL-70, IL-71, IL-72, IL-73, IL-74, IL-75, IL-76, IL-77, IL-78, IL-79, IL-80, IL-81, IL-82, IL-83, IL-84, IL-85, IL-86, IL-87, IL-88, IL-89, IL-90, IL-91, IL-92, IL-93, IL-94, IL-95, IL-96, IL-97, IL-98, IL-99, IL-100, IL-101, IL-102, IL-103, IL-104, IL-105, IL-106, IL-107, IL-108, IL-109, IL-110, IL-111, IL-112, IL-113, IL-114, IL-115, IL-116, IL-117, IL-118, IL-119, IL-120, IL-121, IL-122, IL-123, IL-124, IL-125, IL-126, IL-127, IL-128, IL-129, IL-130, IL-131, IL-132, IL-133, IL-134, IL-135, IL-136, IL-137, IL-138, IL-139, IL-140, IL-141, IL-142, IL-143, IL-144, IL-145, IL-146, IL-147, IL-148, IL-149, IL-150, IL-151, IL-152, IL-153, IL-154, IL-155, IL-156, IL-157, IL-158, IL-159, IL-160, IL-161, IL-162, IL-163, IL-164, IL-165, IL-166, IL-167, IL-168, IL-169, IL-170, IL-171, IL-172, IL-173, IL-174, IL-175, IL-176, IL-177, IL-178, IL-179, IL-180, IL-181, IL-182, IL-183, IL-184, IL-185, IL-186, IL-187, IL-188, IL-189, IL-190, IL-191, IL-192, IL-193, IL-194, IL-195, IL-196, IL-197, IL-198, IL-199, IL-200, IL-201, IL-202, IL-203, IL-204, IL-205, IL-206, IL-207, IL-208, IL-209, IL-210, IL-211, IL-212, IL-213, IL-214, IL-215, IL-216, IL-217, IL-218, IL-219, IL-220, IL-221, IL-222, IL-223, IL-224, IL-225, IL-226, IL-227, IL-228, IL-229, IL-230, IL-231, IL-232, IL-233, IL-234, IL-235, IL-236, IL-237, IL-238, IL-239, IL-240, IL-241, IL-242, IL-243, IL-244, IL-245, IL-246, IL-247, IL-248, IL-249, IL-250, IL-251, IL-252, IL-253, IL-254, IL-255, IL-256, IL-257, IL-258, IL-259, IL-260, IL-261, IL-262, IL-263, IL-264, IL-265, IL-266, IL-267, IL-268, IL-269, IL-270, IL-271, IL-272, IL-273, IL-274, IL-275, IL-276, IL-277, IL-278, IL-279, IL-280, IL-281, IL-282, IL-283, IL-284, IL-285, IL-286, IL-287, IL-288, IL-289, IL-290, IL-291, IL-292, IL-293, IL-294, IL-295, IL-296, IL-297, IL-298, IL-299, IL-300, IL-301, IL-302, IL-303, IL-304, IL-305, IL-306, IL-307, IL-308, IL-309, IL-310, IL-311, IL-312, IL-313, IL-314, IL-315, IL-316, IL-317, IL-318, IL-319, IL-320, IL-321, IL-322, IL-323, IL-324, IL-325, IL-326, IL-327, IL-328, IL-329, IL-330, IL-331, IL-332, IL-333, IL-334, IL-335, IL-336, IL-337, IL-338, IL-339, IL-340, IL-341, IL-342, IL-343, IL-344, IL-345, IL-346, IL-347, IL-348, IL-349, IL-350, IL-351, IL-352, IL-353, IL-354, IL-355, IL-356, IL-357, IL-358, IL-359, IL-360, IL-361, IL-362, IL-363, IL-364, IL-365, IL-366, IL-367, IL-368, IL-369, IL-370, IL-371, IL-372, IL-373, IL-374, IL-375, IL-376, IL-377, IL-378, IL-379, IL-380, IL-381, IL-382, IL-383, IL-384, IL-385, IL-386, IL-387, IL-388, IL-389, IL-390, IL-391, IL-392, IL-393, IL-394, IL-395, IL-396, IL-397, IL-398, IL-399, IL-400, IL-401, IL-402, IL-403, IL-404, IL-405, IL-406, IL-407, IL-408, IL-409, IL-410, IL-411, IL-412, IL-413, IL-414, IL-415, IL-416, IL-417, IL-418, IL-419, IL-420, IL-421, IL-422, IL-423, IL-424, IL-425, IL-426, IL-427, IL-428, IL-429, IL-430, IL-431, IL-432, IL-433, IL-434, IL-435, IL-436, IL-437, IL-438, IL-439, IL-440, IL-441, IL-442, IL-443, IL-444, IL-445, IL-446, IL-447, IL-448, IL-449, IL-450, IL-451, IL-452, IL-453, IL-454, IL-455, IL-456, IL-457, IL-458, IL-459, IL-460, IL-461, IL-462, IL-463, IL-464, IL-465, IL-466, IL-467, IL-468, IL-469, IL-470, IL-471, IL-472, IL-473, IL-474, IL-475, IL-476, IL-477, IL-478, IL-479, IL-480, IL-481, IL-482, IL-483, IL-484, IL-485, IL-486, IL-487, IL-488, IL-489, IL-490, IL-491, IL-492, IL-493, IL-494, IL-495, IL-496, IL-497, IL-498, IL-499, IL-500, IL-501, IL-502, IL-503, IL-504, IL-505, IL-506, IL-507, IL-508, IL-509, IL-510, IL-511, IL-512, IL-513, IL-514, IL-515, IL-516, IL-517, IL-518, IL-519, IL-520, IL-521, IL-522, IL-523, IL-524, IL-525, IL-526, IL-527, IL-528, IL-529, IL-530, IL-531, IL-532, IL-533, IL-534, IL-535, IL-536, IL-537, IL-538, IL-539, IL-540, IL-541, IL-542, IL-543, IL-544, IL-545, IL-546, IL-547, IL-548, IL-549, IL-550, IL-551, IL-552, IL-553, IL-554, IL-555, IL-556, IL-557, IL-558, IL-559, IL-560, IL-561, IL-562, IL-563, IL-564, IL-565, IL-566, IL-567, IL-568, IL-569, IL-570, IL-571, IL-572, IL-573, IL-574, IL-575, IL-576, IL-577, IL-578, IL-579, IL-580, IL-581, IL-582, IL-583, IL-584, IL-585, IL-586, IL-587, IL-588, IL-589, IL-590, IL-591, IL-592, IL-593, IL-594, IL-595, IL-596, IL-597, IL-598, IL-599, IL-600, IL-601, IL-602, IL-603, IL-604, IL-605, IL-606, IL-607, IL-608, IL-609, IL-610, IL-611, IL-612, IL-613, IL-614, IL-615, IL-616, IL-617, IL-618, IL-619, IL-620, IL-621, IL-622, IL-623, IL-624, IL-625, IL-626, IL-627, IL-628, IL-629, IL-630, IL-631, IL-632, IL-633, IL-634, IL-635, IL-636, IL-637, IL-638, IL-639, IL-640, IL-641, IL-642, IL-643, IL-644, IL-645, IL-646, IL-647, IL-648, IL-649, IL-650, IL-651, IL-652, IL-653, IL-654, IL-655, IL-656, IL-657, IL-658, IL-659, IL-660, IL-661, IL-662, IL-663, IL-664, IL-665, IL-666, IL-667, IL-668, IL-669, IL-670, IL-671, IL-672, IL-673, IL-674, IL-675, IL-676, IL-677, IL-678, IL-679, IL-680, IL-681, IL-682, IL-683, IL-684, IL-685, IL-686, IL-687, IL-688, IL-689, IL-690, IL-691, IL-692, IL-693, IL-694, IL-695, IL-696, IL-697, IL-698, IL-699, IL-700, IL-701, IL-702, IL-703, IL-704, IL-705, IL-706, IL-707, IL-708, IL-709, IL-710, IL-711, IL-712, IL-713, IL-714, IL-715, IL-716, IL-717, IL-718, IL-719, IL-720, IL-721, IL-722, IL-723, IL-724, IL-725, IL-726, IL-727, IL-728, IL-729, IL-730, IL-731, IL-732, IL-733, IL-734, IL-735, IL-736, IL-737, IL-738, IL-739, IL-740, IL-741, IL-742, IL-743, IL-744, IL-745, IL-746, IL-747, IL-748, IL-749, IL-750, IL-751, IL-752, IL-753, IL-754, IL-755, IL-756, IL-757, IL-758, IL-759, IL-7510, IL-7511, IL-7512, IL-7513, IL-7514, IL-7515, IL-7516, IL-7517, IL-7518, IL-7519, IL-7520, IL-7521, IL-7522, IL-7523, IL-7524, IL-7525, IL-7526, IL-7527, IL-7528, IL-7529, IL-7530, IL-7531, IL-7532, IL-7533, IL-7534, IL-7535, IL-7536, IL-7537, IL-7538, IL-7539, IL-7540, IL-7541, IL-7542, IL-7543, IL-7544, IL-7545, IL-7546, IL-7547, IL-7548, IL-7549, IL-7550, IL-7551, IL-7552, IL-7553, IL-7554, IL-7555, IL-7556, IL-7557, IL-7558, IL-7559, IL-75510, IL-75511, IL-75512, IL-75513, IL-75514, IL-75515, IL-75516, IL-75517, IL-75518, IL-75519, IL-75520, IL-75521, IL-75522, IL-75523, IL-75524, IL-75525, IL-75526, IL-75527, IL-75528, IL-75529, IL-75530, IL-75531, IL-75532, IL-75533, IL-75534, IL-75535, IL-75536, IL-75537, IL-75538, IL-75539, IL-75540, IL-75541, IL-75542, IL-75543, IL-75544, IL-75545, IL-75546, IL-75547, IL-75548, IL-75549, IL-75550, IL-75551, IL-75552, IL-75553, IL-75554, IL-75555, IL-75556, IL-75557, IL-75558, IL-75559, IL-75560, IL-75561, IL-75562, IL-75563, IL-75564, IL-75565, IL-75566, IL-75567, IL-75568, IL-75569, IL-75570, IL-75571, IL-75572, IL-75573, IL-75574, IL-75575, IL-75576, IL-75577, IL-75578, IL-75579, IL-75580, IL-75581, IL-75582, IL-75583, IL-75584, IL-75585, IL-75586, IL-75587, IL-75588, IL-75589, IL-7

# Why is Common Sense Challenging?

**Massive**

Water is liquid.

Water can be found in ocean.

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.

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?

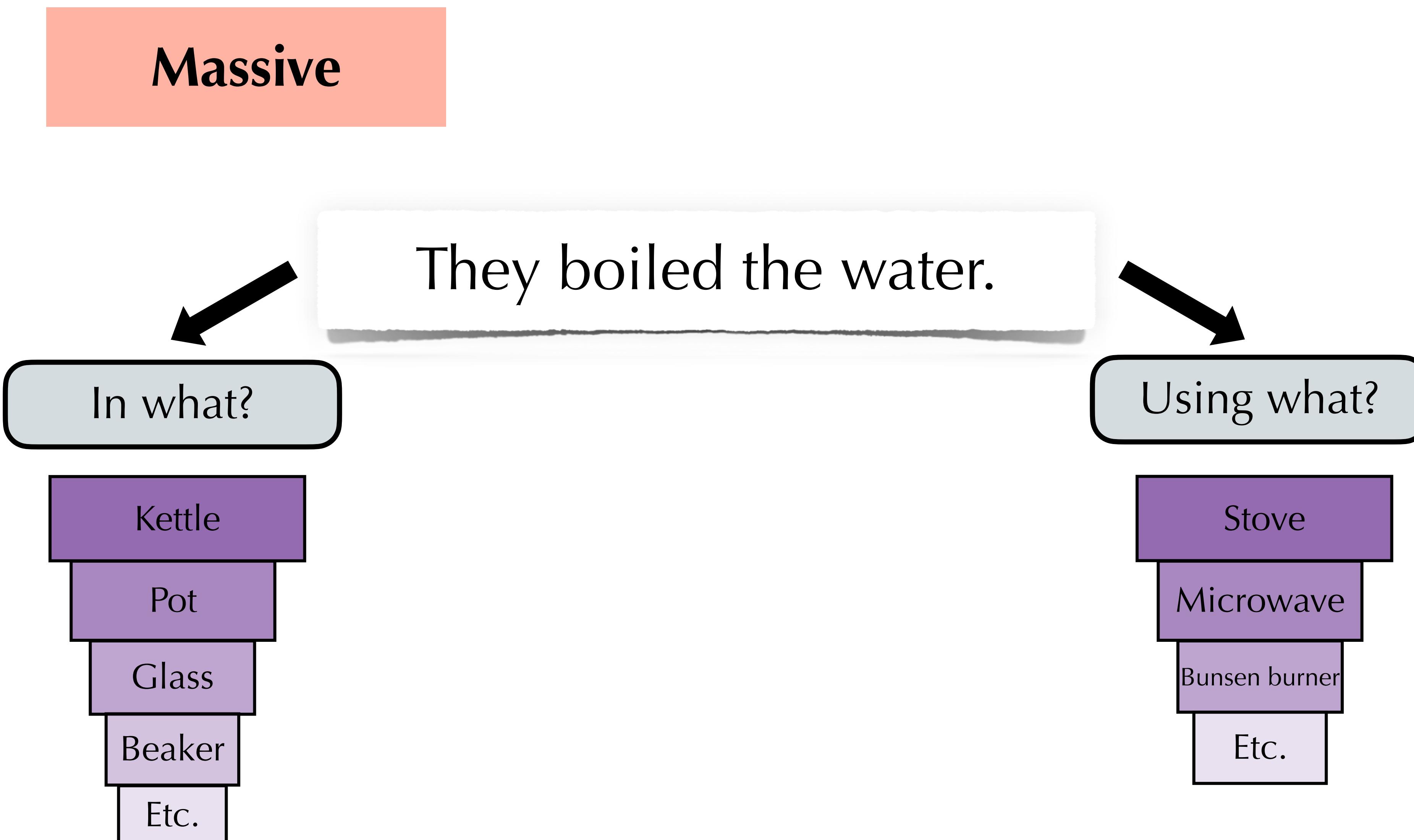
Stove

Microwave

Bunsen burner

Etc.

# Why is Common Sense Challenging?



# Why is Common Sense Challenging?

Massive

Probabilistic

They boiled the water.

In what?

Using what?

Kettle

Pot

Glass

Beaker

Etc.

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?

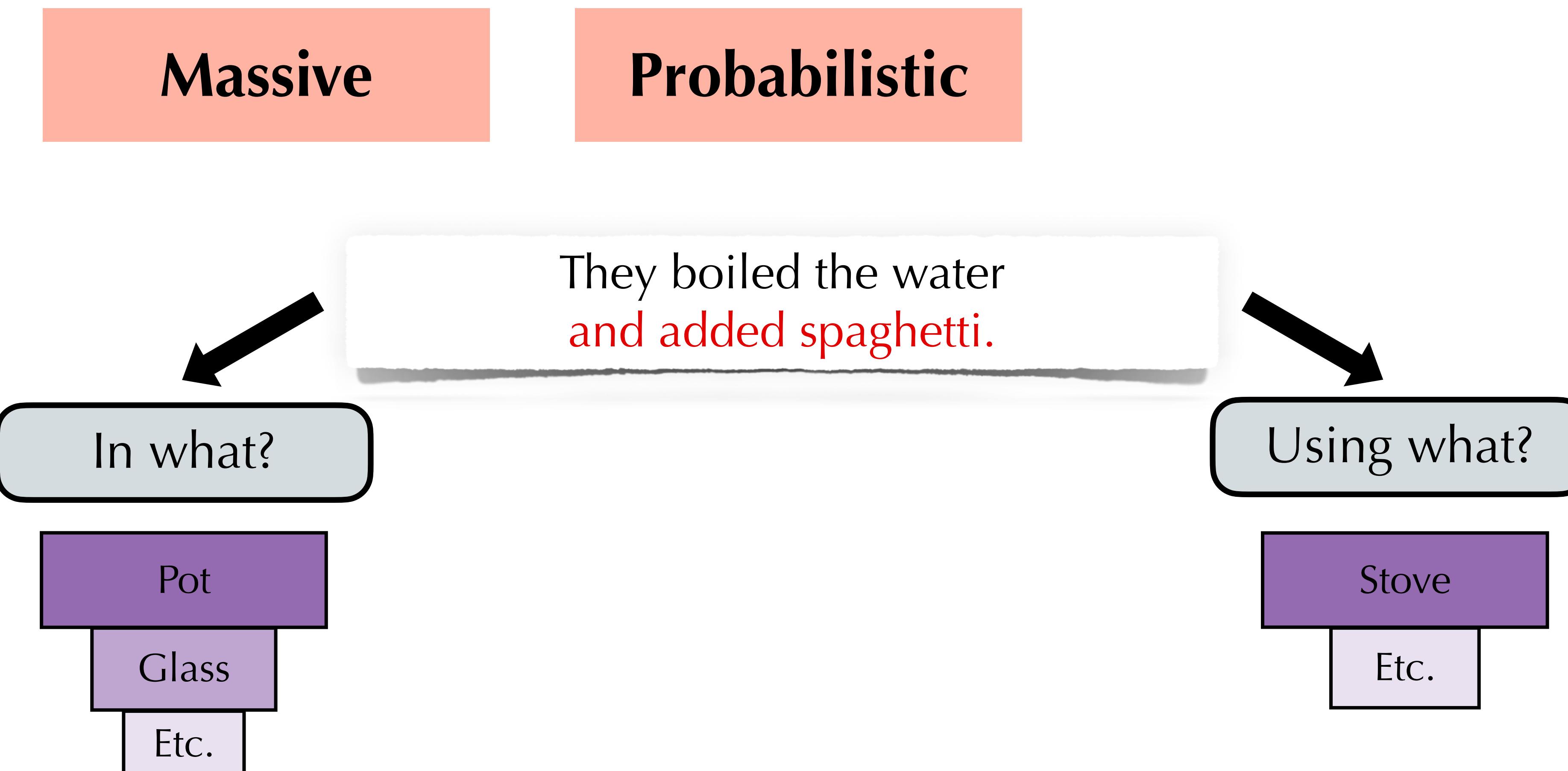
Stove

Microwave

Bunsen burner

Etc.

# Why is Common Sense Challenging?



# Why is Common Sense Challenging?

Massive

Probabilistic

Contextual

They boiled the water  
and added spaghetti.

In what?

Pot

Glass

Etc.

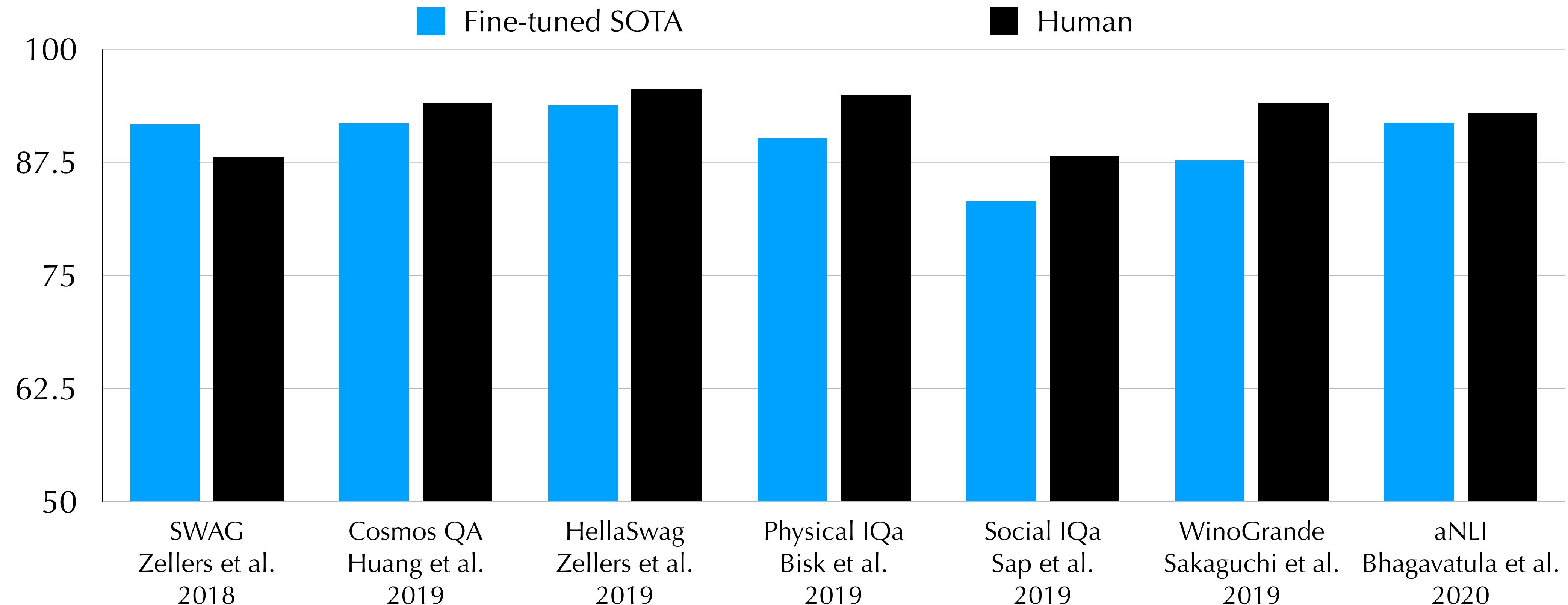
Using what?

Stove

Etc.

# Common Sense in Language Model

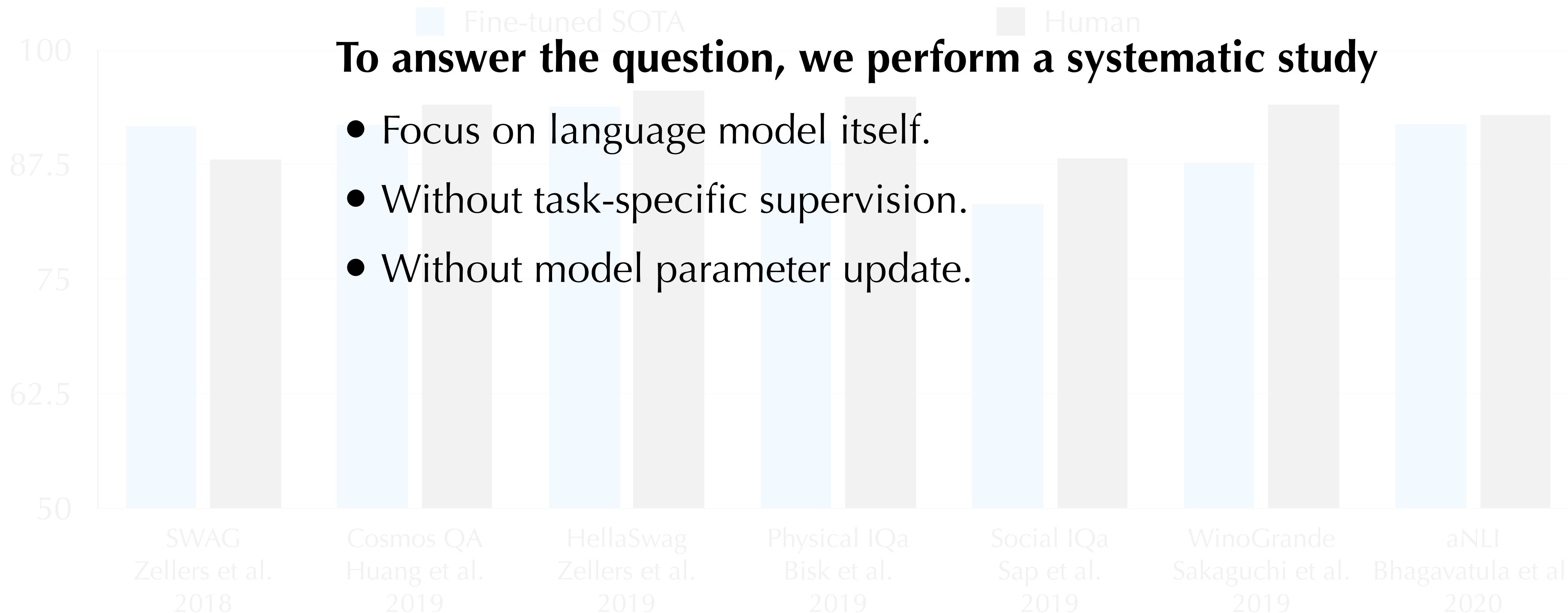
Models based on large language models show impressive performance on many **commonsense question answering** tasks.



# Do language models learn common sense?

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

Zero-shot evaluation on language models



To answer the question, we perform a systematic study

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

# Do language models learn common sense?

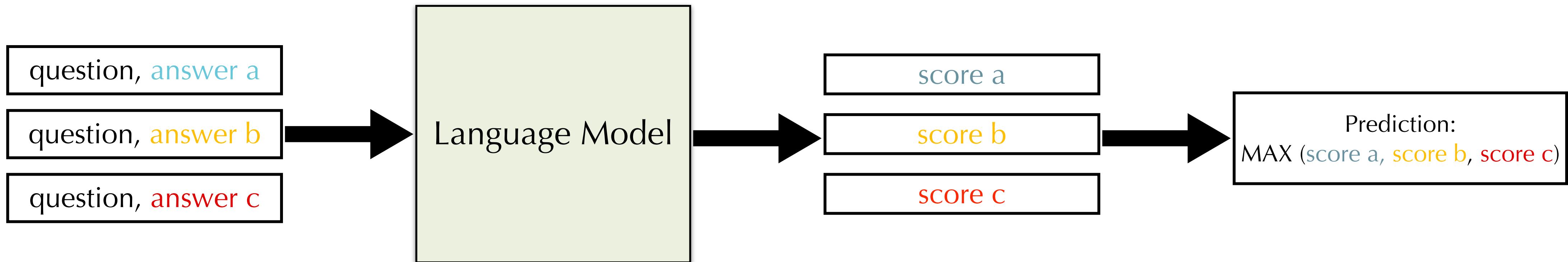
Dataset	Example	Number of Choices	Reasoning Type
<b>Physical IQa</b> (Bisk et al. 2019)	<b>Question:</b> To apply eyeshadow without a brush, should I use a cotton swab or a toothpick? <b>Answer:</b> Cotton swab.	2	Physical
<b>Social IQa</b> (Sap et al. 2019)	<b>Question:</b> Tracy had accidentally pressed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

Four multiple choice selection QA datasets.

# Do language models learn common sense?

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

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



# Zero-shot Performance: random baseline

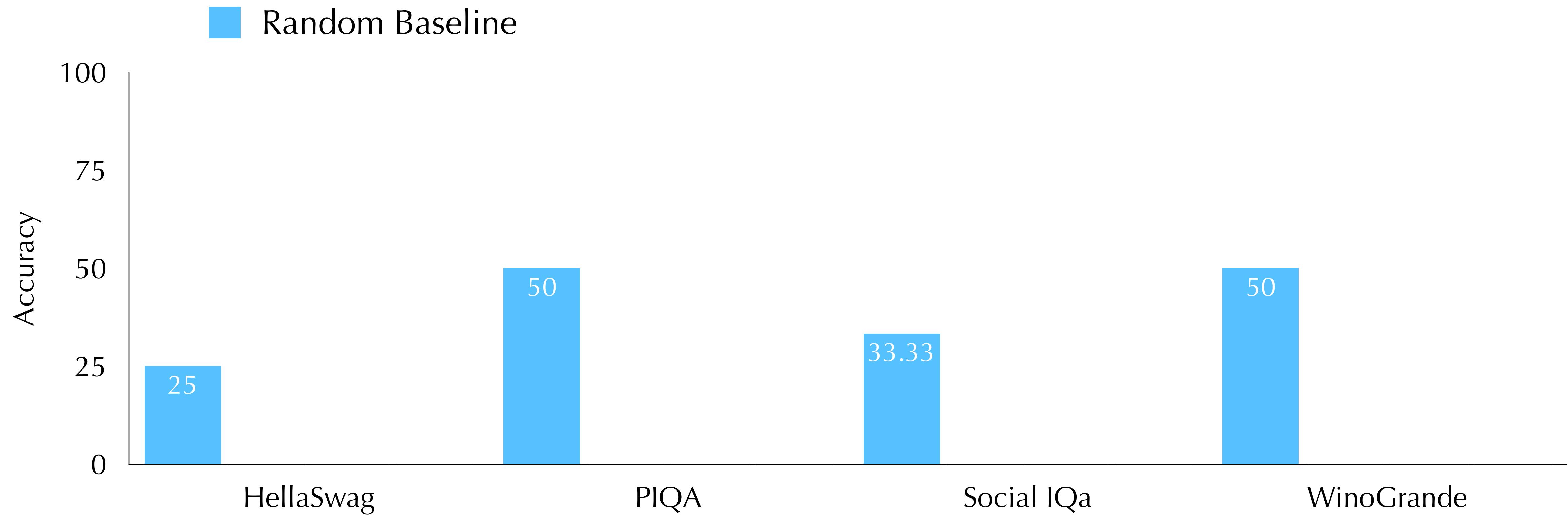


Figure: the dev accuracy for each dataset evaluated on Gopher.

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

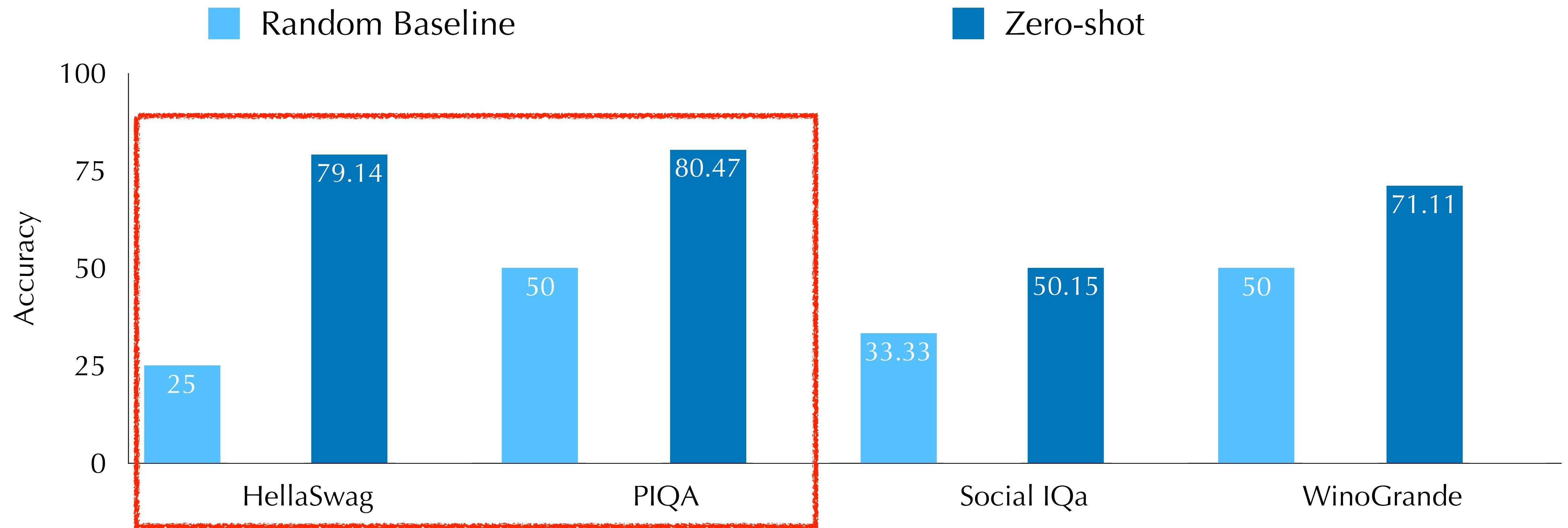


Figure: the dev accuracy for each dataset evaluated on Gopher.

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

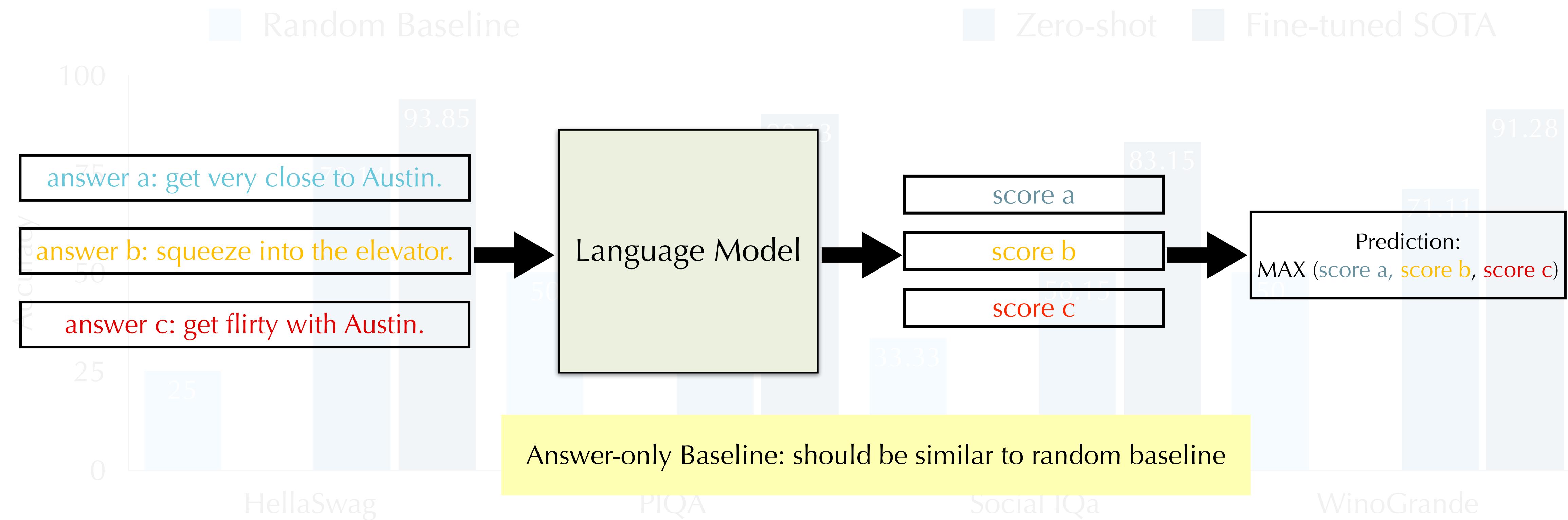


Figure: the dev accuracy for each dataset evaluated on Gopher.

# Models pick the correct answer without seeing the question

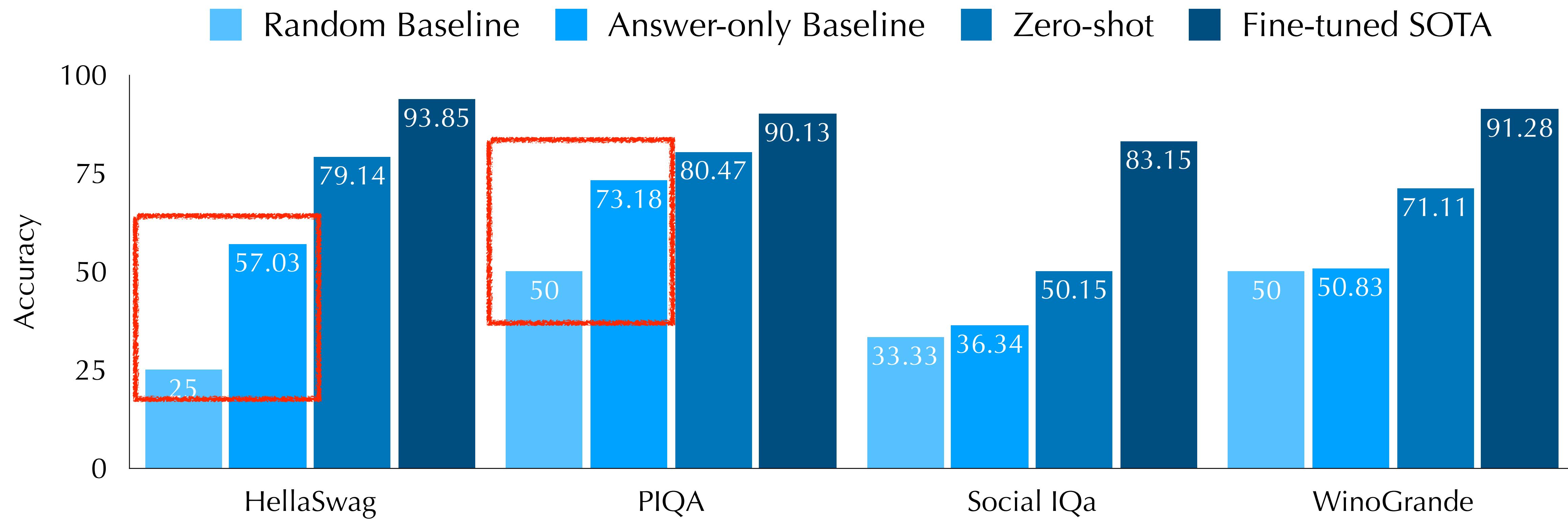


Figure: the dev accuracy for each dataset evaluated on Gopher.

# We need better commonsense evaluation!

Dataset Bias!

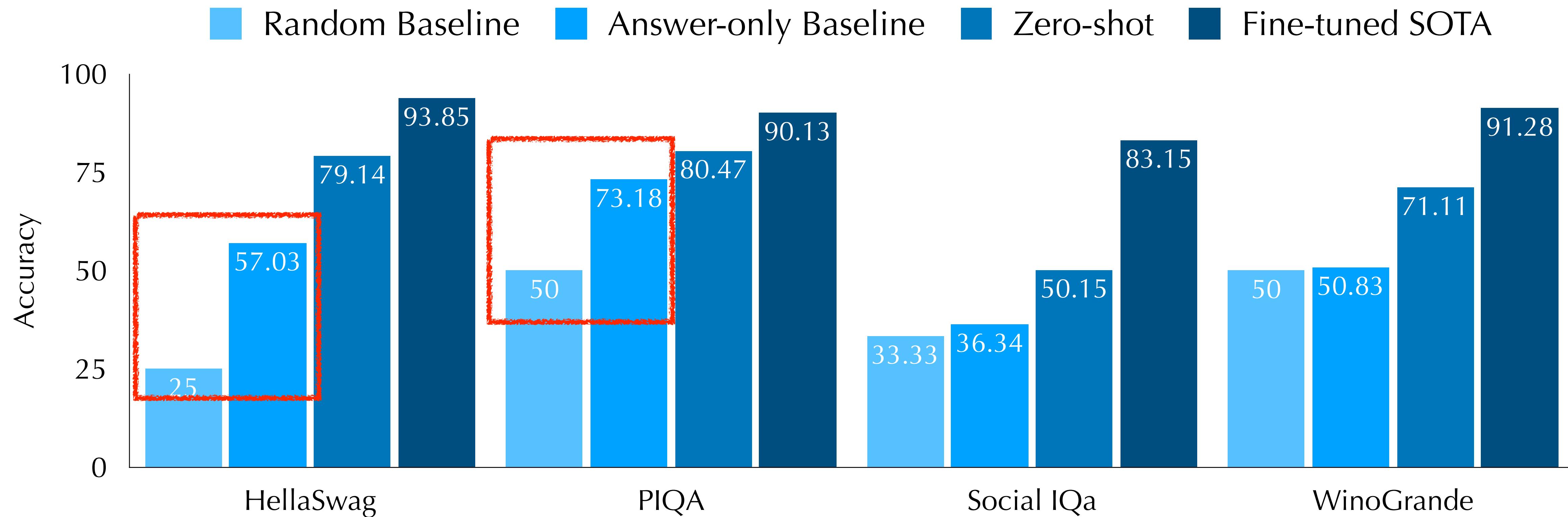


Figure: the dev accuracy for each dataset evaluated on Gopher.

# Outline

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

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

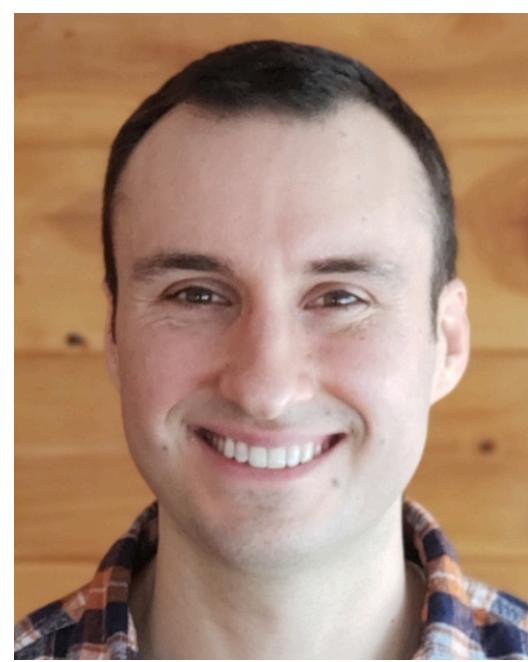
- UNcommonsense Reasoning: Abductive Reasoning about Uncommon Situations. [NAACL 2024]
- In search of the long-tail: systematic generation of long-tail knowledge via logical rule guided search [EMNLP 2024]
- Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games. [EMNLP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

# Every Answer Matters: Evaluating Commonsense with Probabilistic Measures

Qi Cheng



Michael Boratko



Pranay Yelugam



Tim O' Gorman



Nalini Singh



Andrew  
McCallum

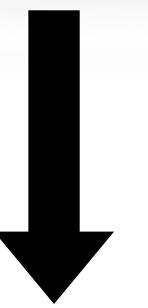


Xiang Lorraine Li

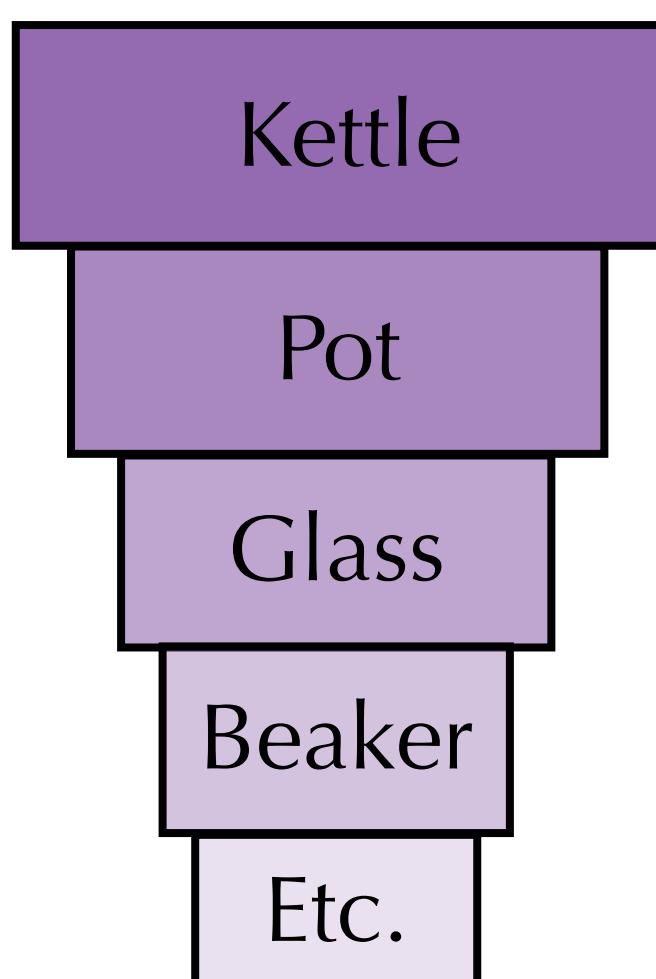


# Probabilistic Evaluation of Commonsense

They boiled the water.

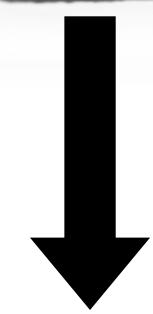


In what?

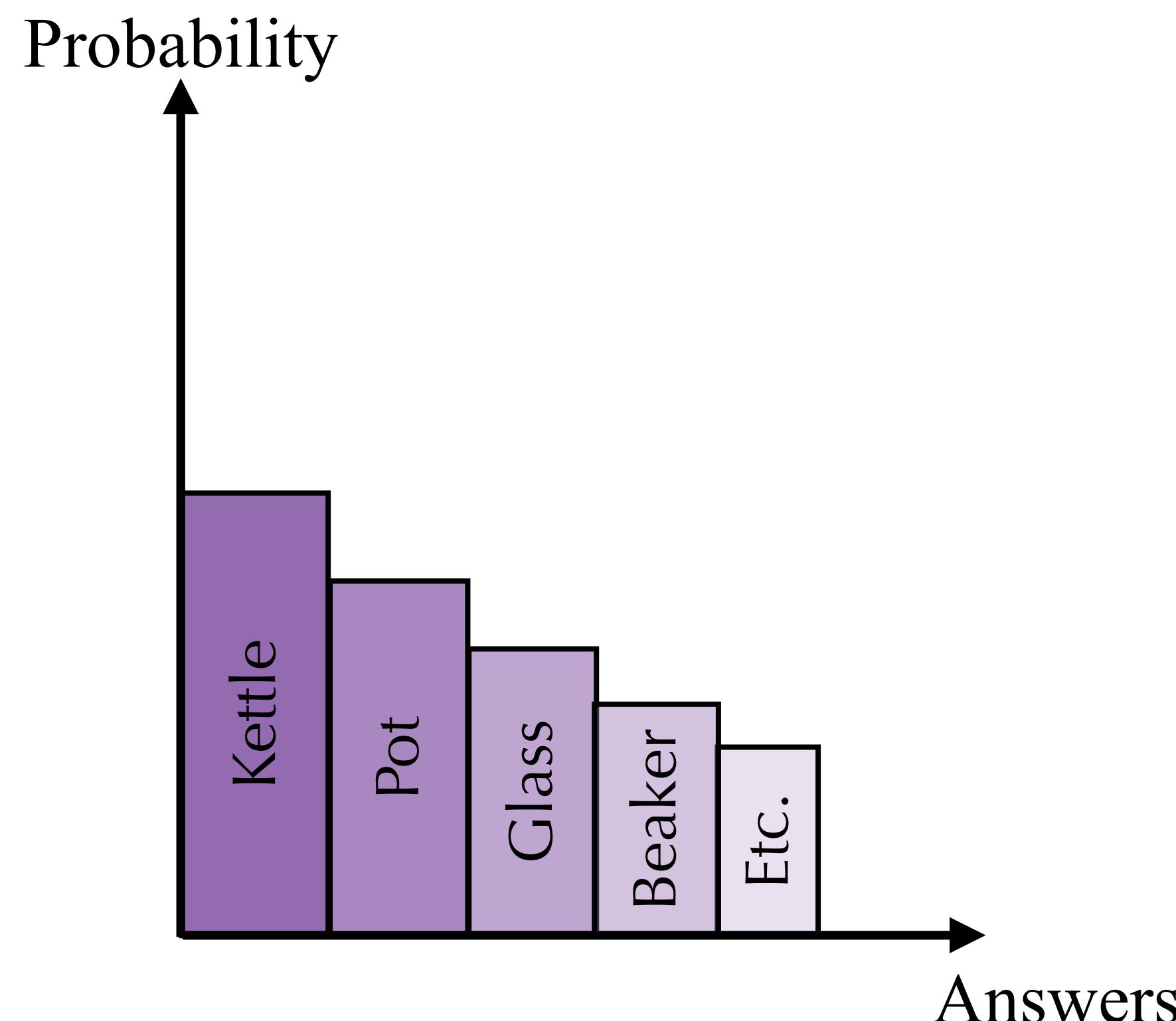


# Probabilistic Evaluation of Commonsense

They boiled the water.



In what?



Question Answering

Dialogue

Any language tasks!

# CFC Data Collection

We crowd-source high-quality evaluation data

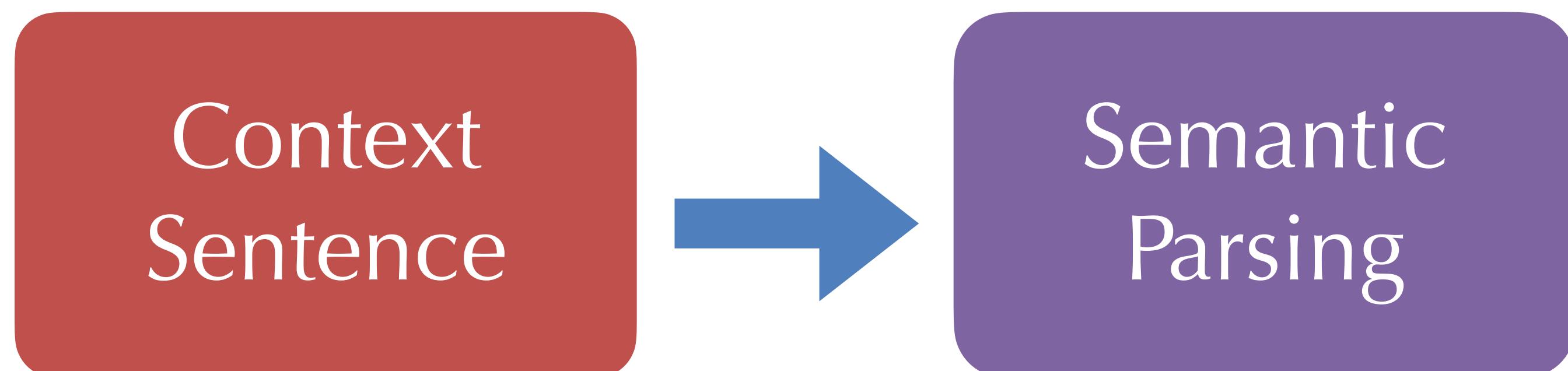
Context  
Sentence

“Dog catches the  
thrown frisbee.”

CommonGen (Image Captions)

# CFC Data Collection

We crowd-source high-quality evaluation data



“Dog catches the thrown frisbee.”



CommonGen (Image Captions)

AMR Parsing

# CFC Data Collection

We crowd-source high-quality evaluation data



“Dog catches the thrown frisbee.”



“Who throws the frisbee?”

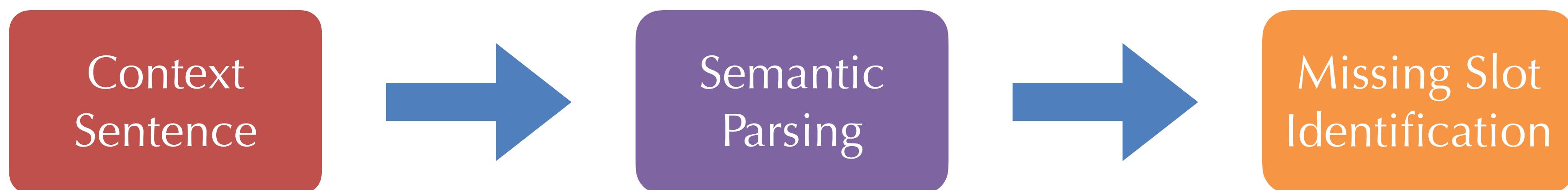
CommonGen (Image Captions)

AMR Parsing

AMR-unknown

# CFC Data Collection

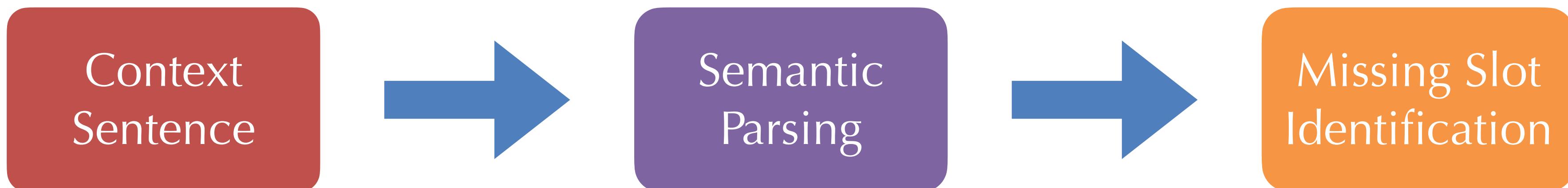
We crowd-sourced high-quality **101 questions (manual filtering)**



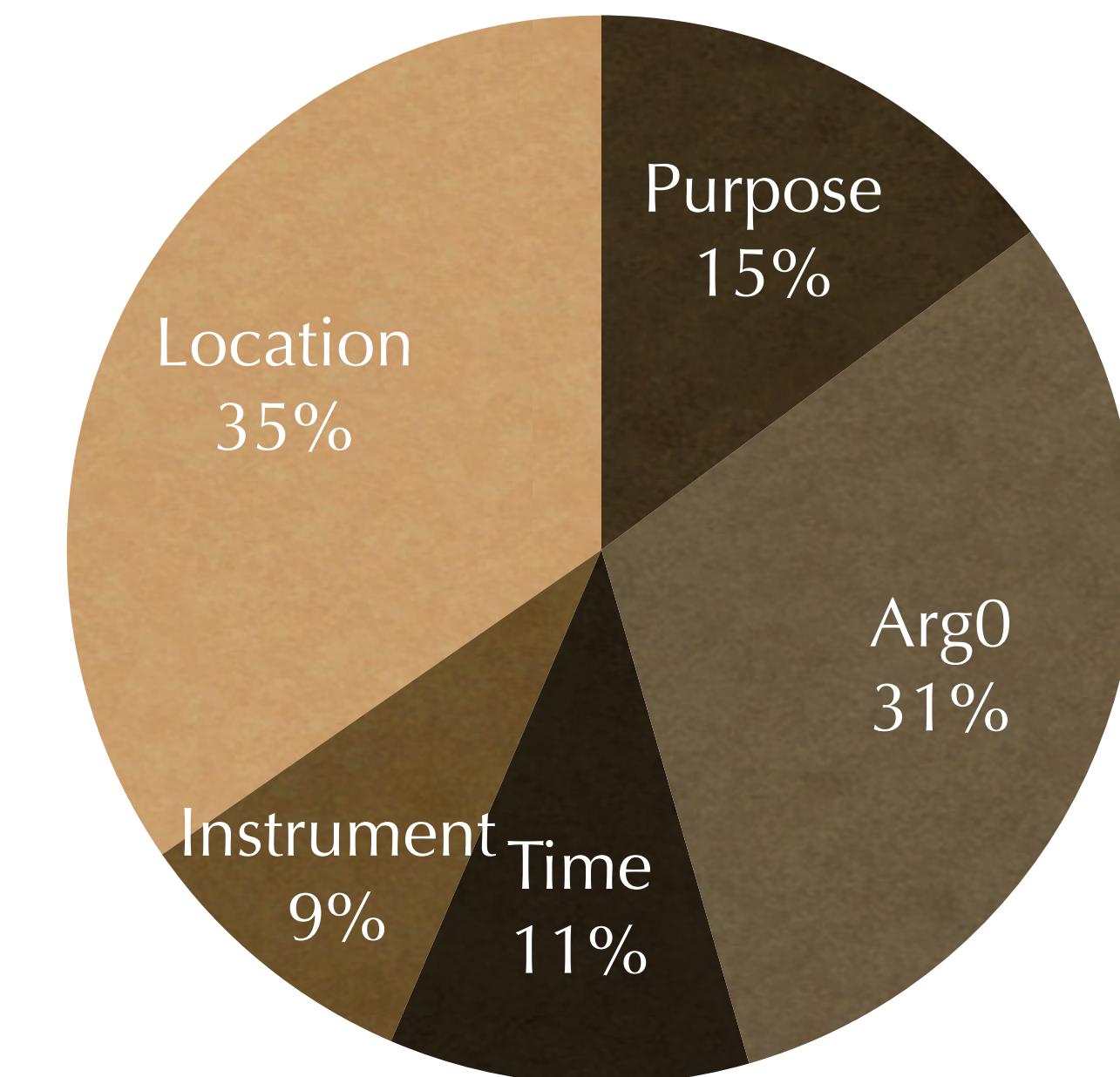
Missing Slot	Definition	Examples
Arg0	Who/what does the event?	Sentence: putting cheese on the pizza. Arg0? Answers: person, cook
Purpose	What is the goal for doing the event?	Sentence: putting cheese on the pizza. Purpose? Answers: get nutrition, stop being hungry
Instrument	What kind of tools are used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

# CFC Data Collection

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



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



# CFC Data Collection

“They boiled the water” Purpose?

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

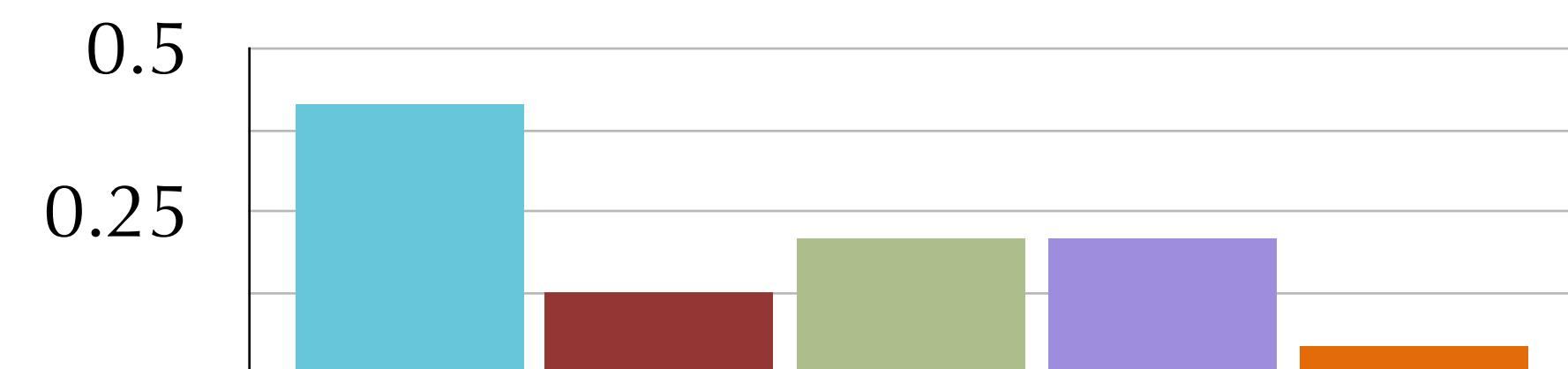


# CFC Data Collection

“They boiled the water” Purpose?

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

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

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

# CFC Data Collection

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

- Classic problem in statistics.
  - The approximated error rate is bounded by [1]
$$\rightarrow \mathbb{P}(D_{KL}(g_{n,k} || f) \geq \epsilon) \leq e^{-n\epsilon} \left[ \frac{3c_1}{c_2} \sum_{i=0}^{k-2} K_{i-1} \left( \frac{e\sqrt{n}}{2\pi} \right)^i \right]$$
    - $n$ : number of samples
    - $k$ : number of category in the categorical distribution
    - $\epsilon$ : KL error rate

# CFC Data Collection

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

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

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

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

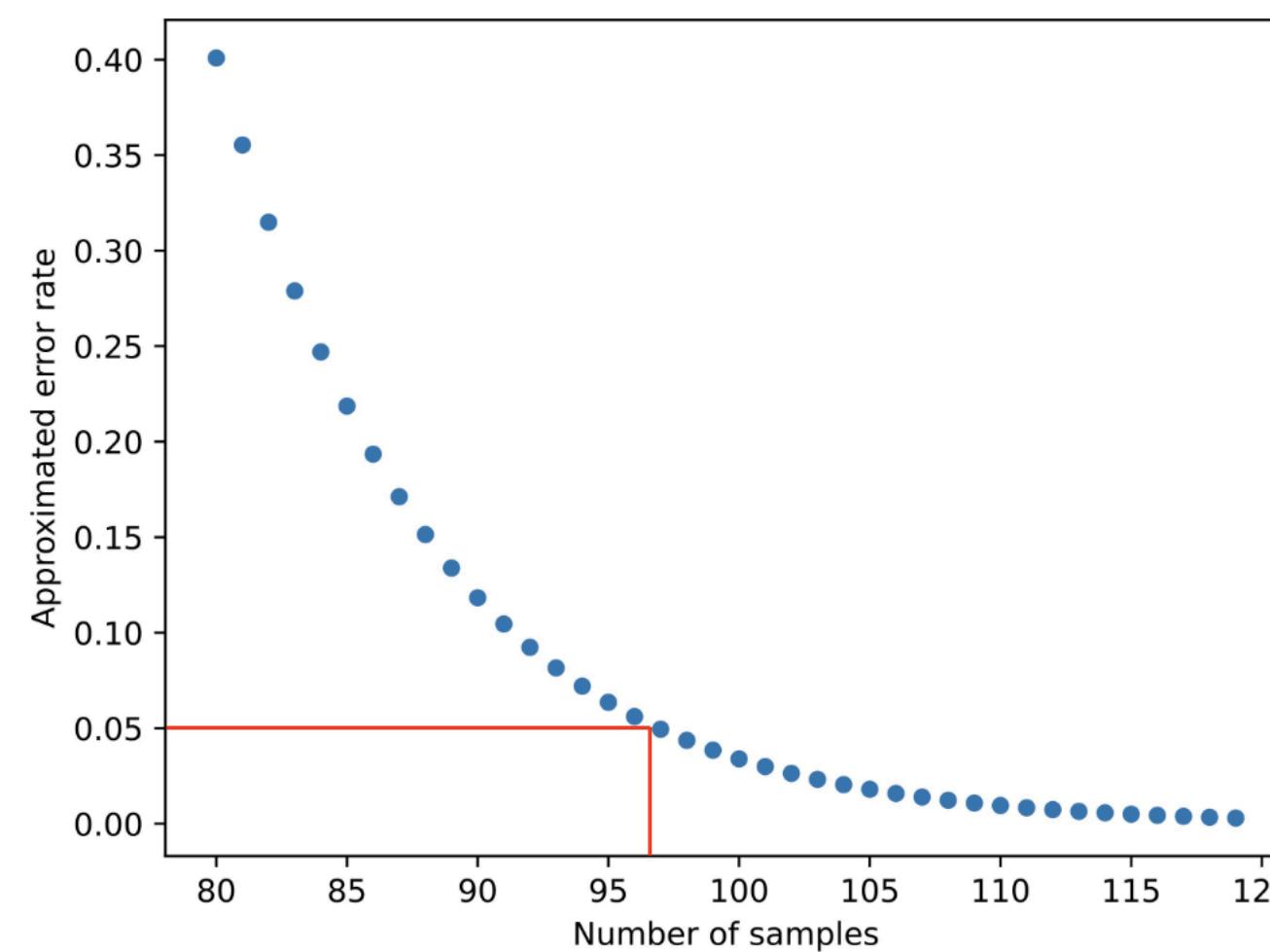
# CFC Data Collection

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

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

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

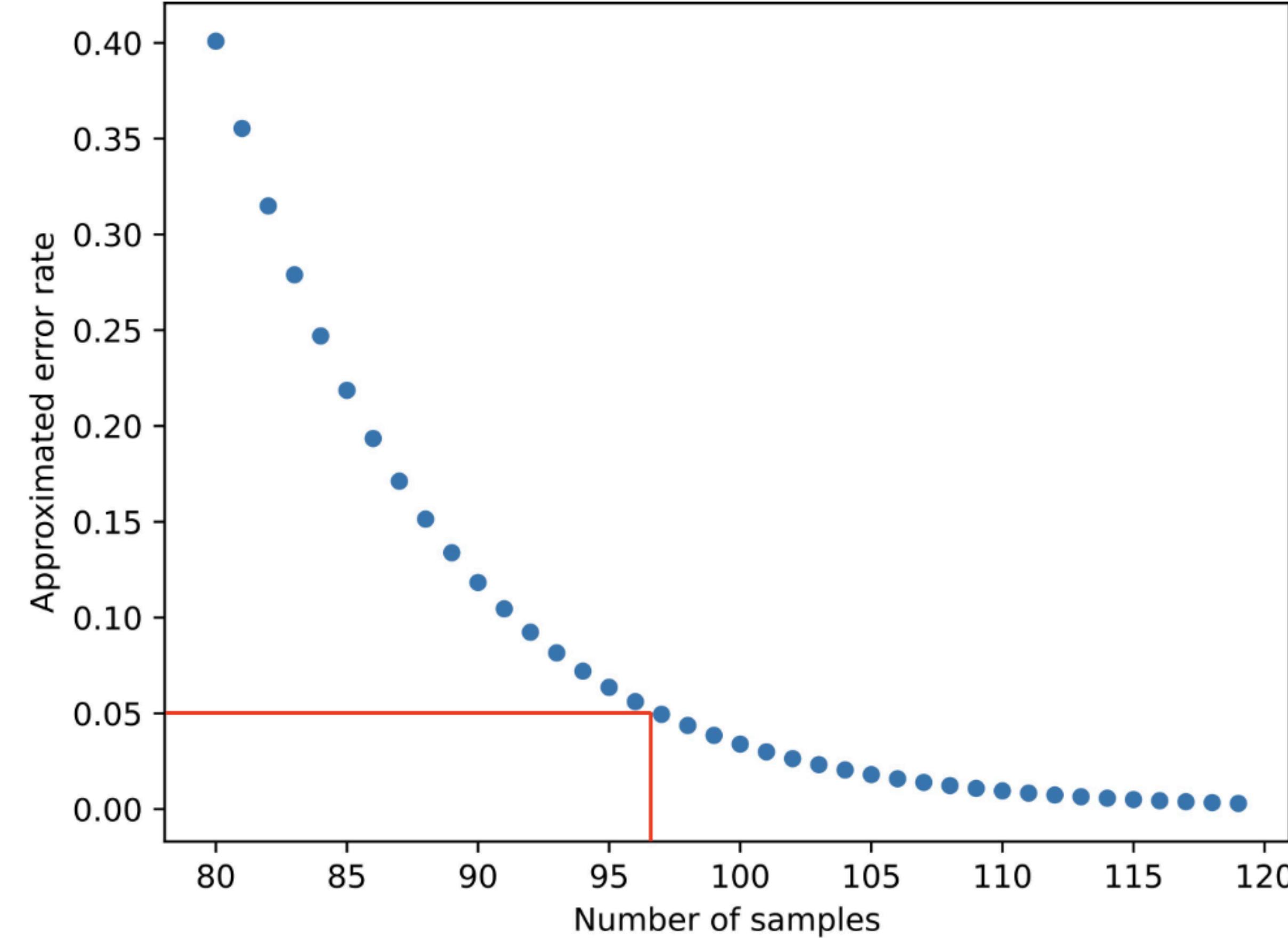
- $n$ : number of samples
- $k$ : number of category in the categorical distribution = 8
- $\epsilon$ : KL error rate = 0.2



# CFC Data Collection

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

~97. we collect 100 answers for each question.



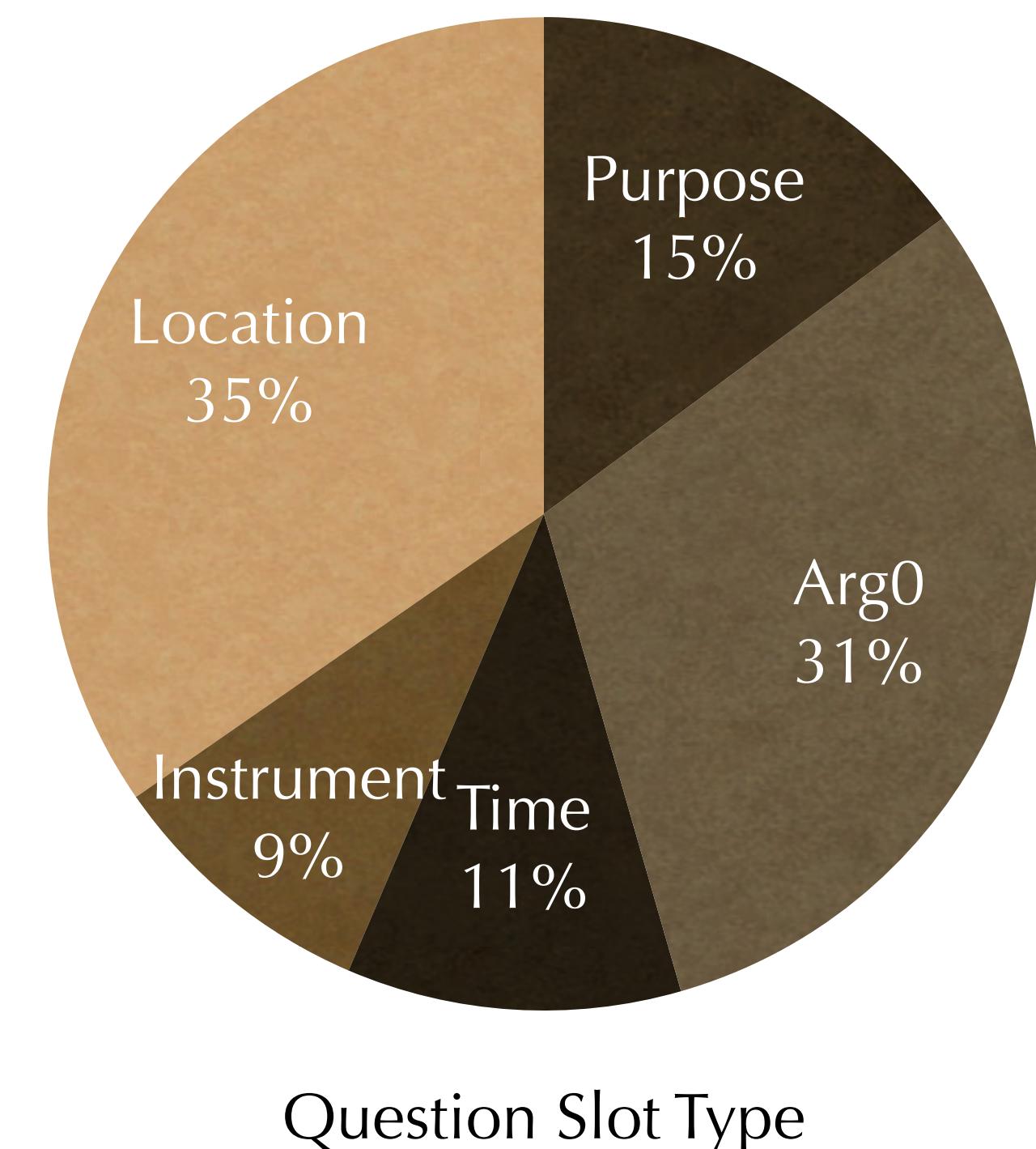
# CFC Data Statistics

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

- 55 Dev Questions
- 46 Test Questions

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

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



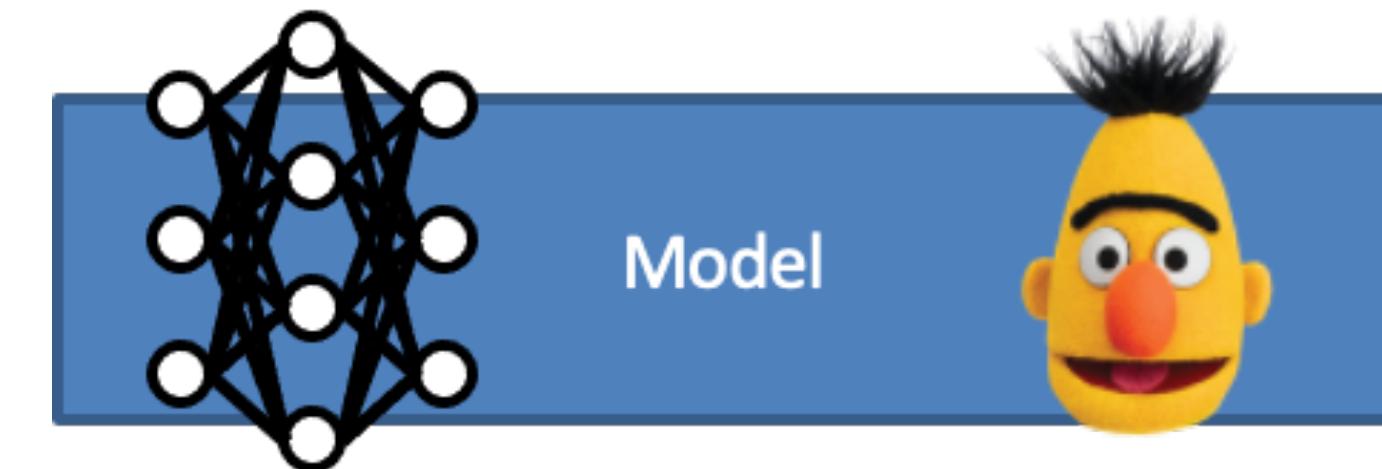
cook, cook noodles, cook pasta, bake cake, boil eggs, pasta, make pasta, cook meal,  
to make tea, coffee, make coffee, to make it safe to drink, to sterilize it, to remove  
germs and make it safe to drink ...

# CFC Probabilistic Evaluation

“They boiled the water” Purpose?



Crowd Workers

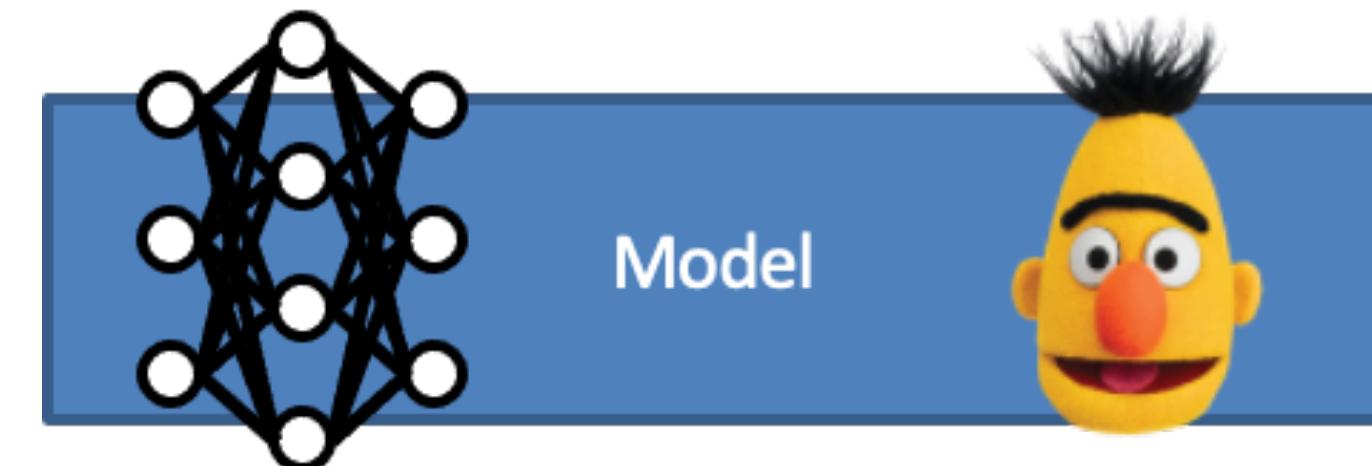


Model

# CFC Probabilistic Evaluation

“They boiled the water” Purpose?

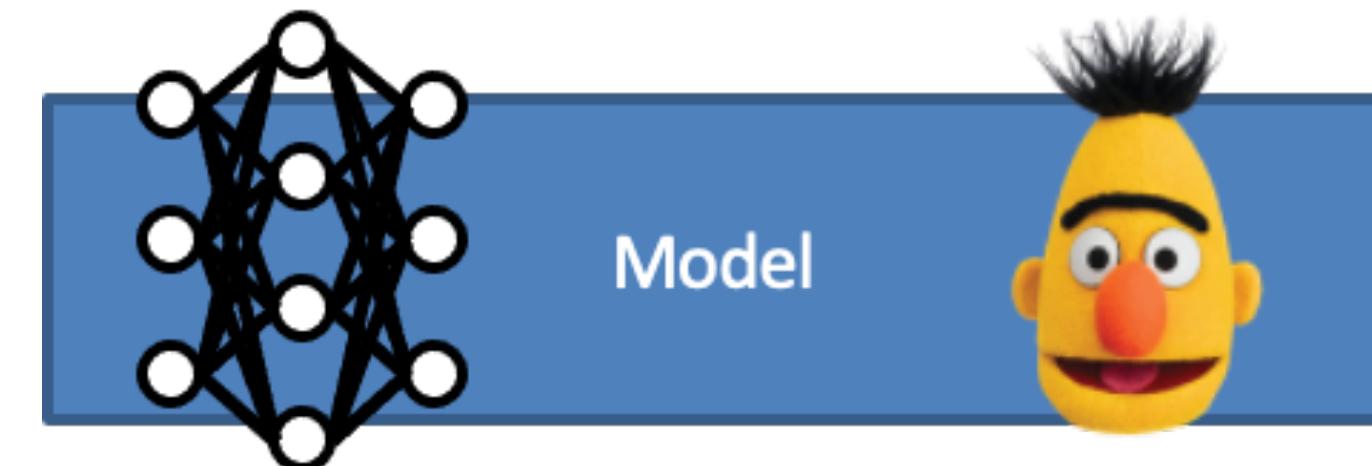
cooking		clean
cook	make tea	disinfect
for making tea		disinfecting
to cook	cook food	making dinner
cooking spaghetti		cleaning
steaming vegetables		cleaning tools
boiling potatoes	for a hot drink	kill bacteria
boiling chicken	making pasta	purify
		purification
	make safe to drink	
	sterilization	
	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



# CFC Probabilistic Evaluation

“They boiled the water” Purpose?

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



# CFC Probabilistic Evaluation

“They boiled the water” Cause?

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

cooking

cook dinner

to make hard boiled eggs

making food

sterilize 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

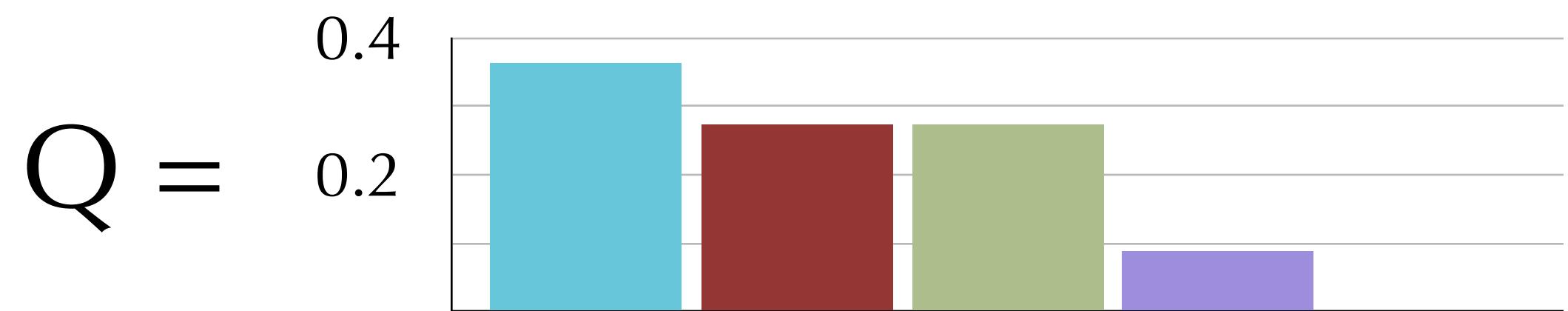
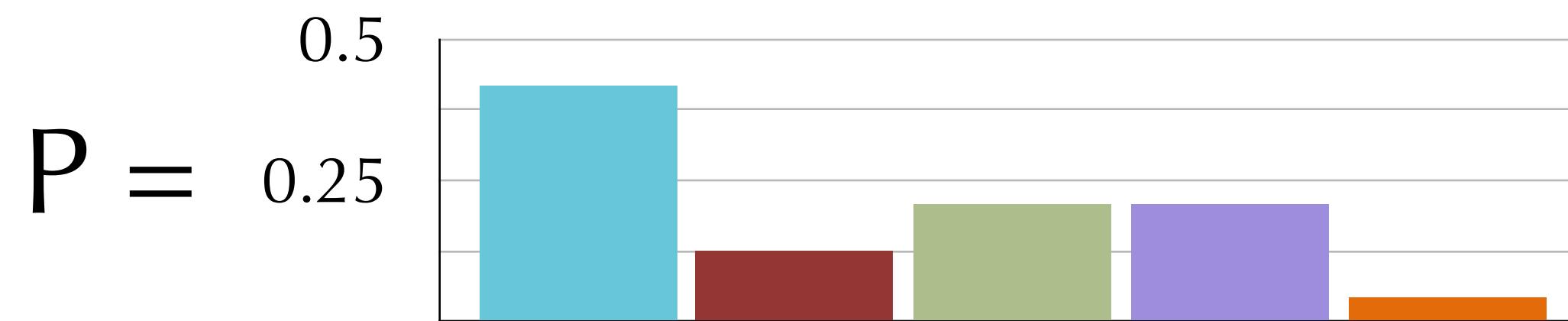
for an experiment

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  
sterilize instruments  
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

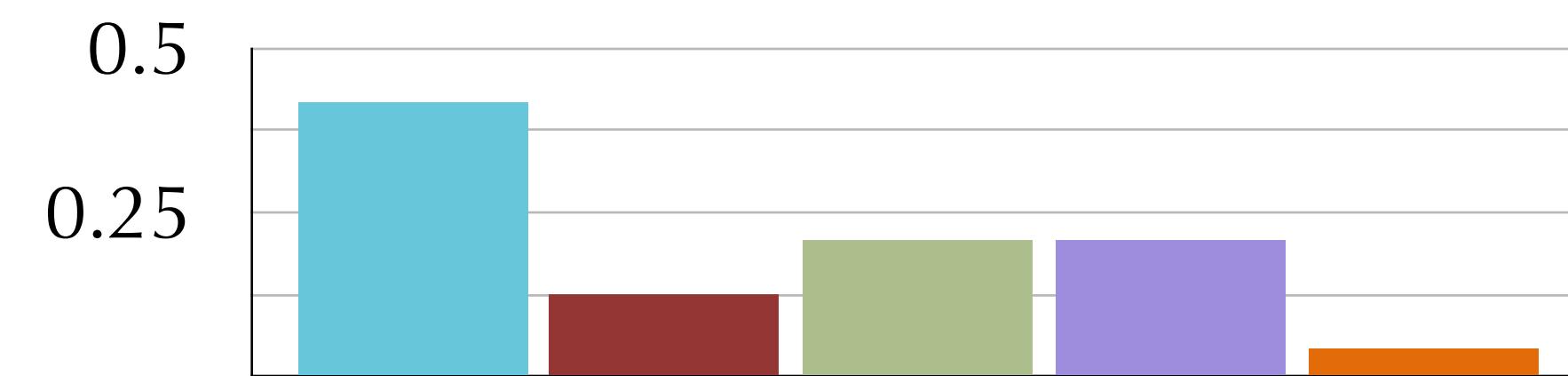
steriliza instruments

cooking

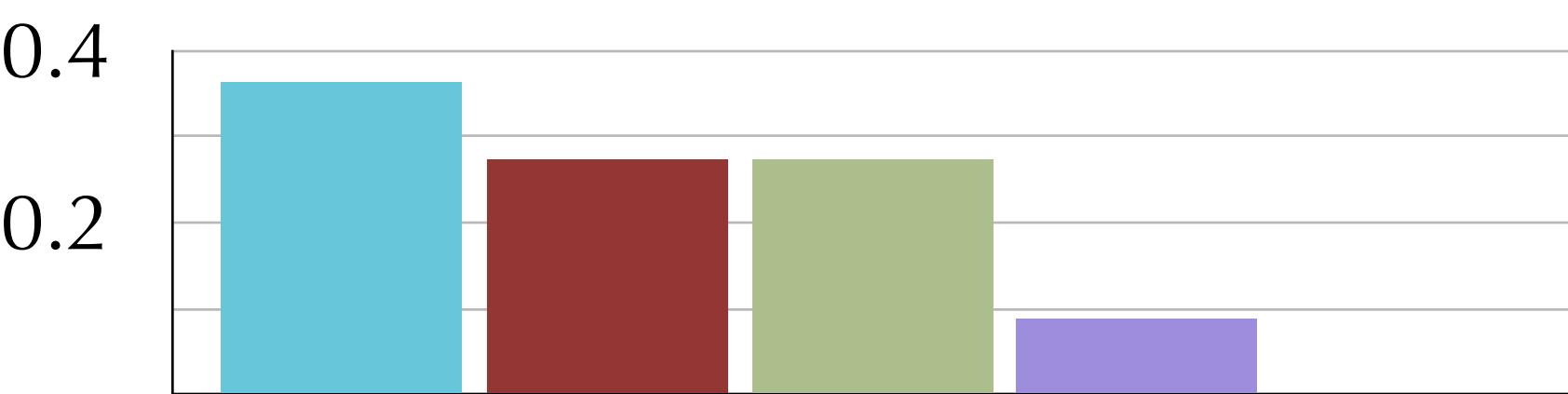
to make hard boiled eggs

making food cook dinner

kill parasites



$KL(P \parallel Q)$



# CFC Automatic Evaluation

*For each question:*

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

$H \leftarrow$  evaluation answers (model)

For each human scorer:

Cluster G

Match H to clusters of G

Calculate score

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

# CFC Automatic Evaluation

*For each question:*

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

$H \leftarrow$  evaluation answers (model)

For each human scorer:

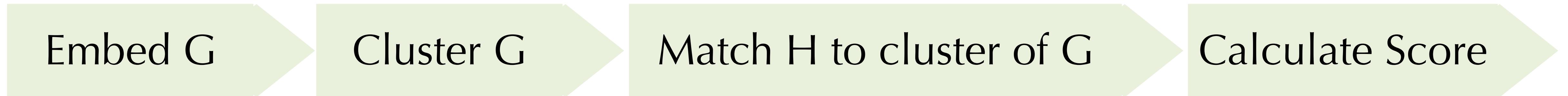
Cluster G

Match H to clusters of G

Calculate score

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

# CFC Automatic Evaluation



# 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**

for making tea

make tea

for a hot drink

cooking

to cook

making dinner

cooking spaghetti

cook food

boiling chicken

boiling potatoes

making pasta

steaming vegetables

disinfect

disinfecting

clean

cleaning tools

cleaning

purify

kill bacteria

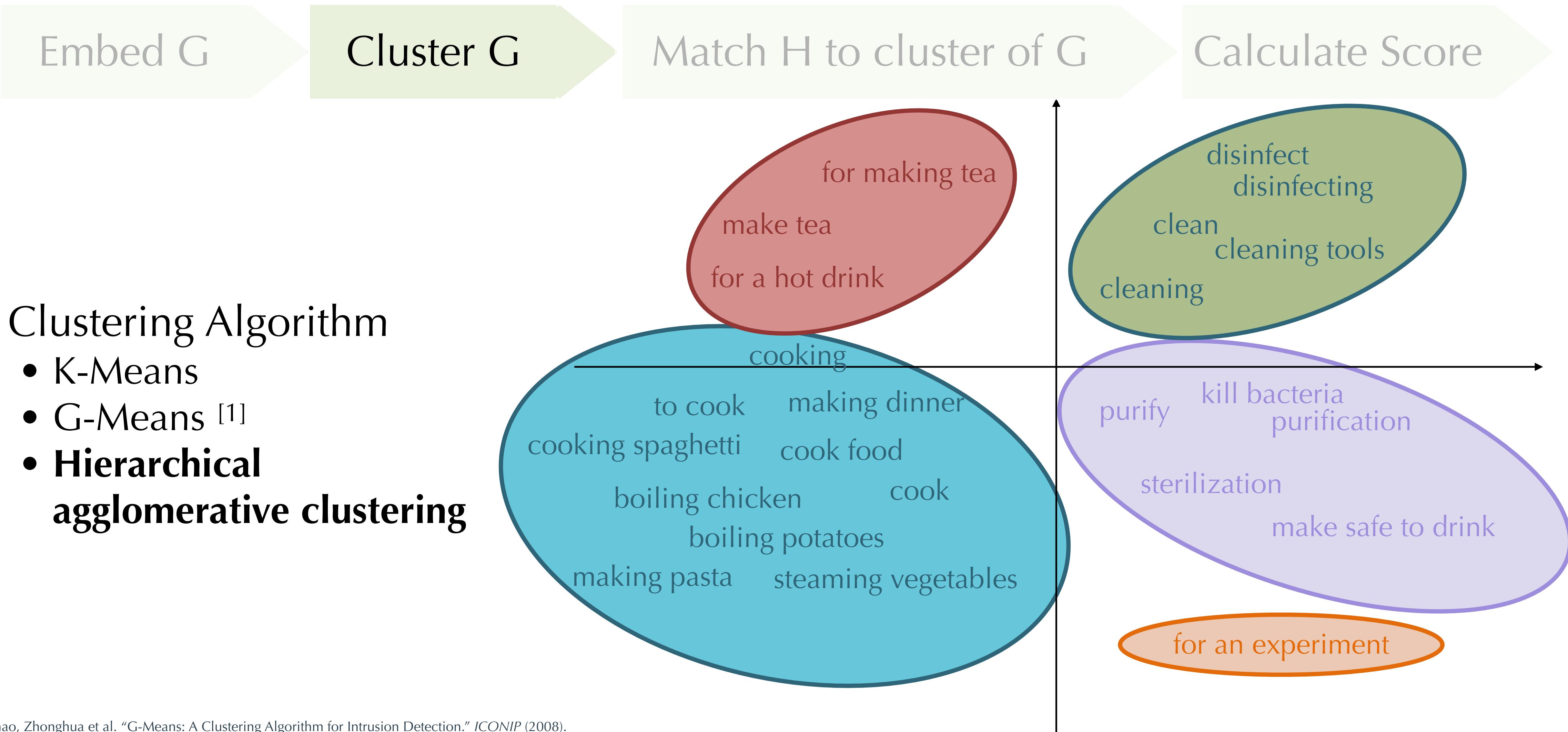
purification

sterilization

make safe to drink

for an experiment

# CFC Automatic Evaluation



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

# CFC Automatic Evaluation

Embed G

Cluster G

Match H to cluster of G

Calculate Score

make tea for a hot drink  
for making tea

clean cleaning disinfect  
disinfecting cleaning tools

purify kill bacteria  
make safe to drink  
purification sterilization

for an experiment

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

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

# CFC Automatic Evaluation

Embed G

Cluster G

Match H to cluster of G

Calculate Score

## Embeddings Based

- FastText

## Lexical Token Based

- WordNet



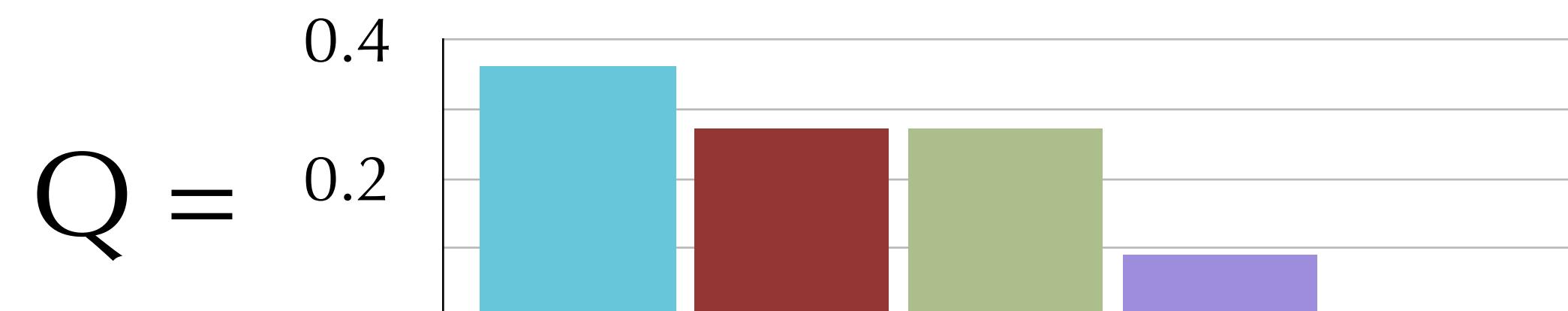
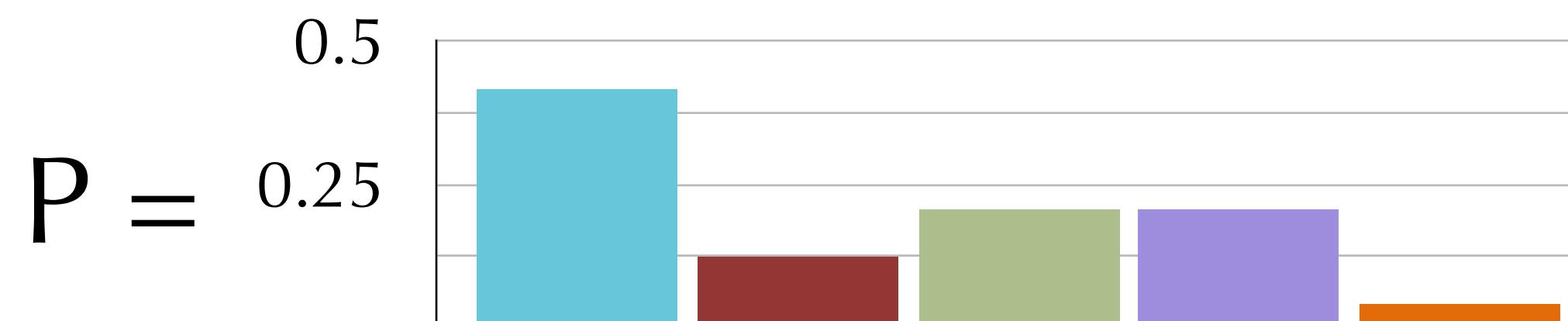
# CFC Automatic Evaluation

Embed G

Cluster G

Match H to cluster of G

Calculate Score



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

# Evaluating Automatic Metric

Given a question, and a large prediction set

- Sample **n** predicted answer sets.  
 $s_1, s_2, s_3, s_4, s_5\dots$
- Using **human** annotations, score answer sets:  
**H**:  $[s_2, s_5, s_4, s_3, s_1\dots]$
- Using **automatic** evaluation, score answer sets:  
**A**:  $[s_2, s_4, s_3, s_1, s_5]$
- 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

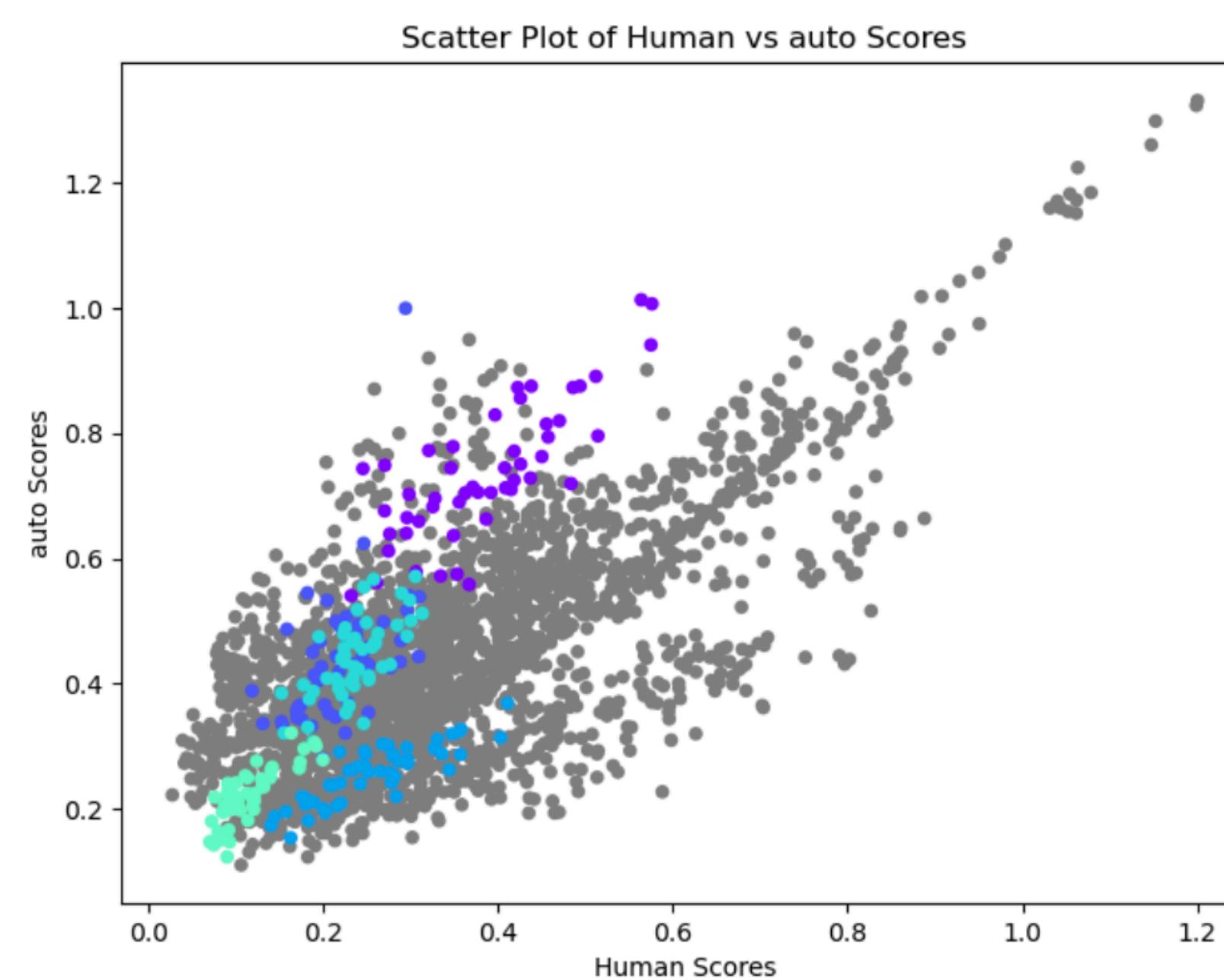
Table: Spearman correlation between human KL score and automatic KL score

# Evaluating Automatic Metric

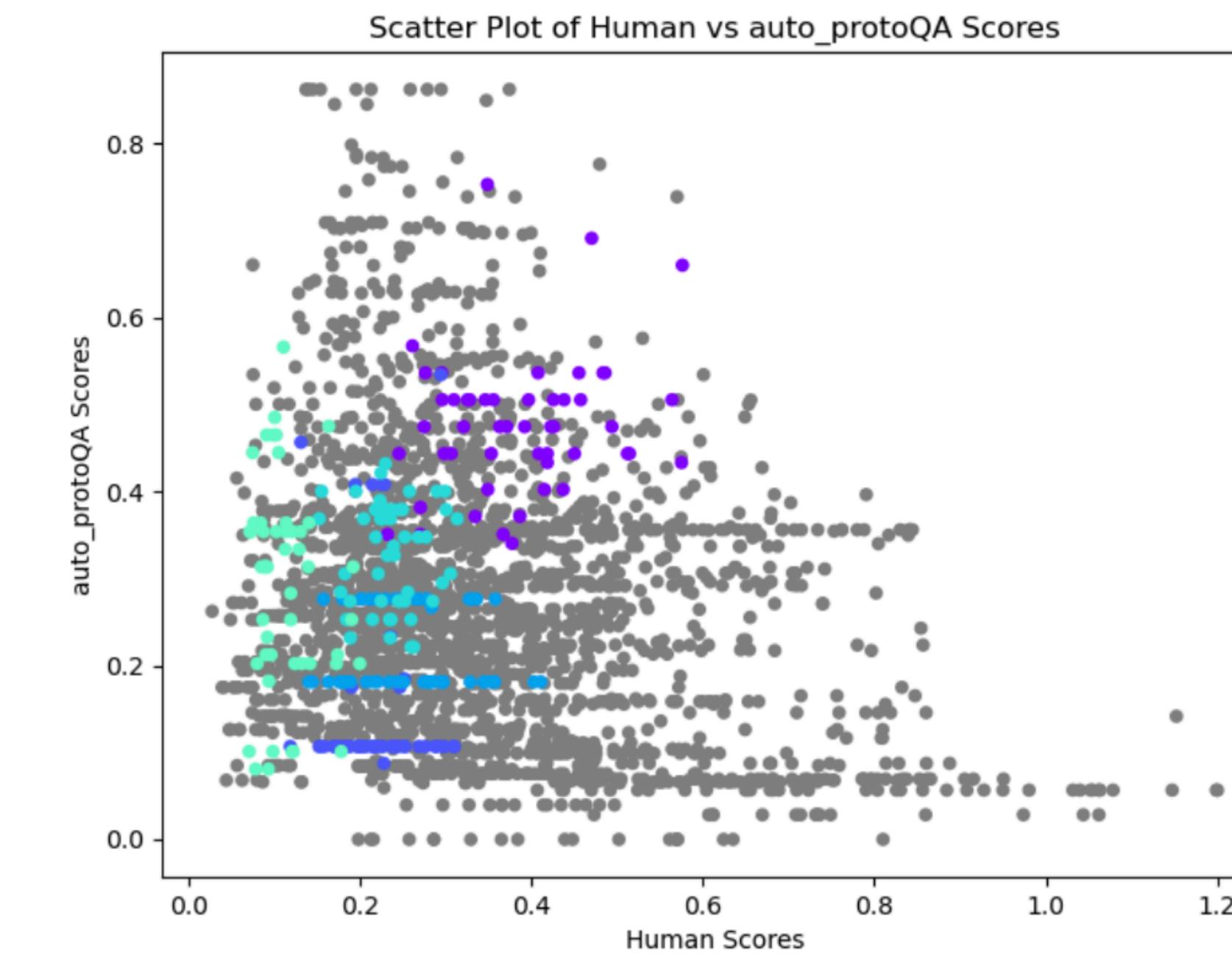
Clustering	Gmeans		Xmeans		Hierarchical agglomerative clustering (HAC)	
Matching	FastText	WordNet	FastText	WordNet	FastText	WordNet
ProtoQA Correlation	0.528	0.681	0.525	0.668	0.593	<b>0.698</b>
CFC Correlation	0.561	0.721	0.503	0.728	0.564	<b>0.728</b>

# 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



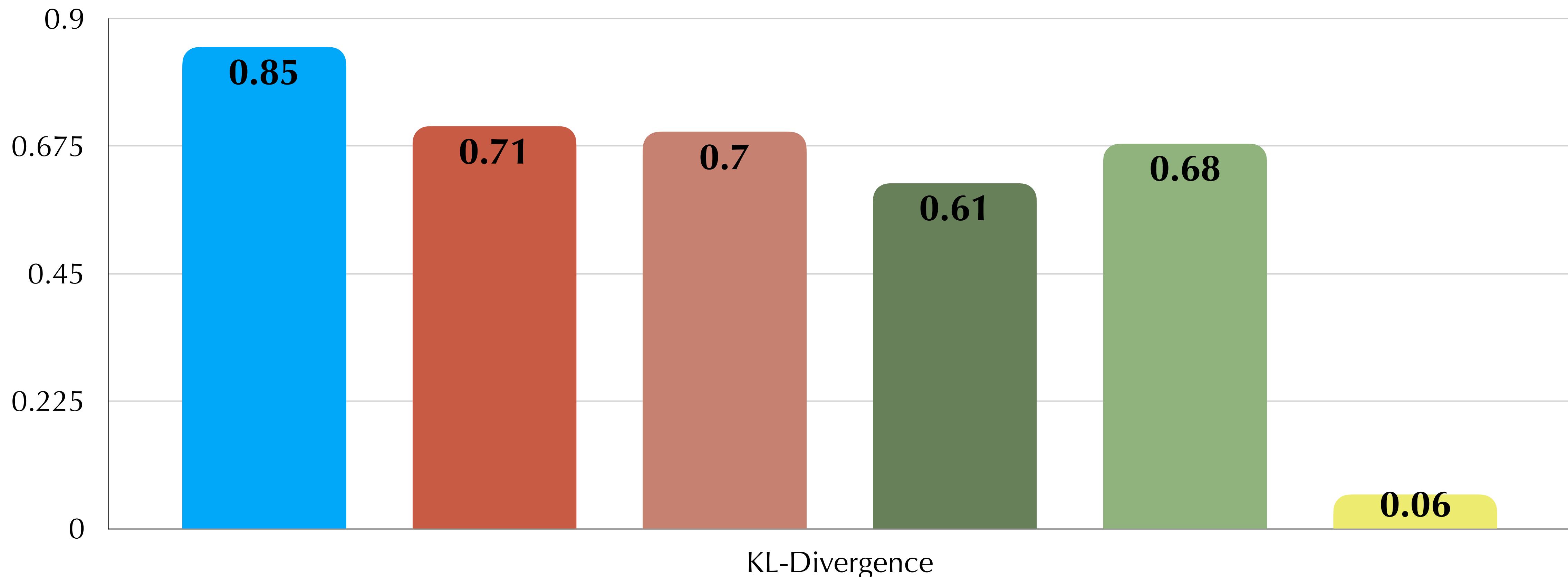
Ours



ProtoQA Evaluator

# Model Performance

■ Llama2 Few-Shot ■ GPT2 Large FT ■ GPT2 Large FT with ProtoQA ■ GPT-3.5 Few-Shot  
■ GPT4 Few-Shot ■ Human



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

**Joey Hou**



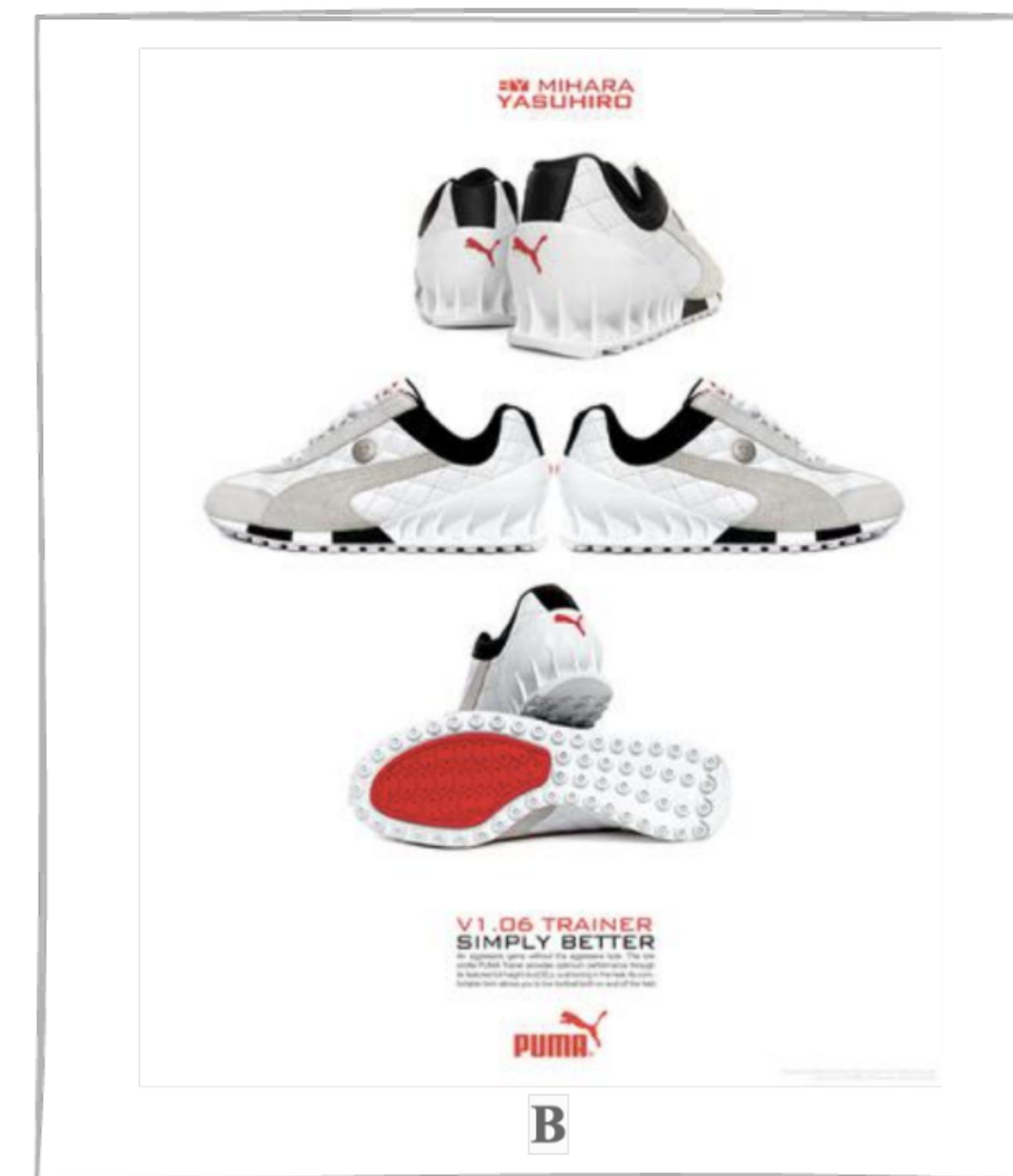
**Adriana Kovashka**



**Xiang Lorraine Li**



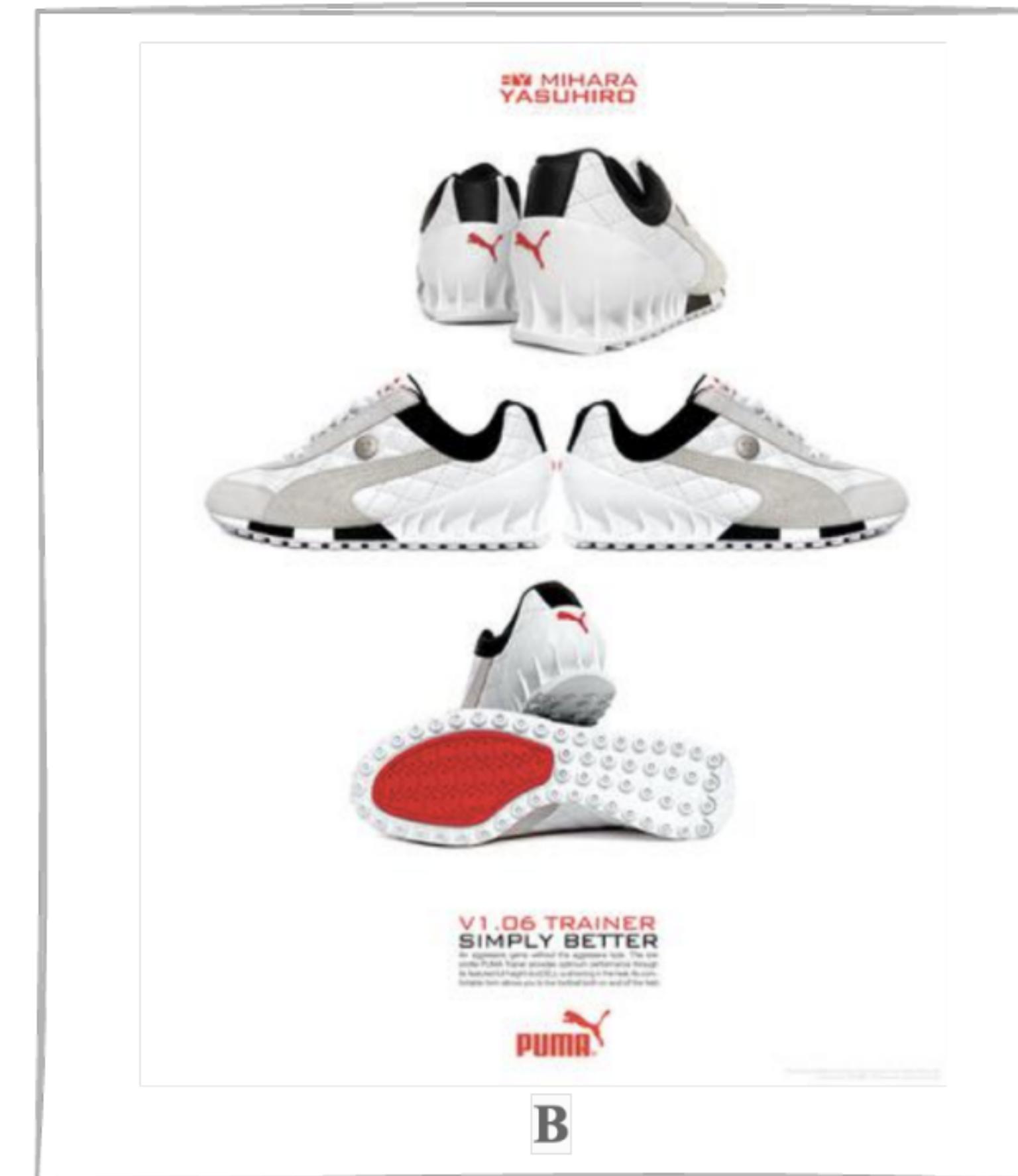
# Which image ads is more creative



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?

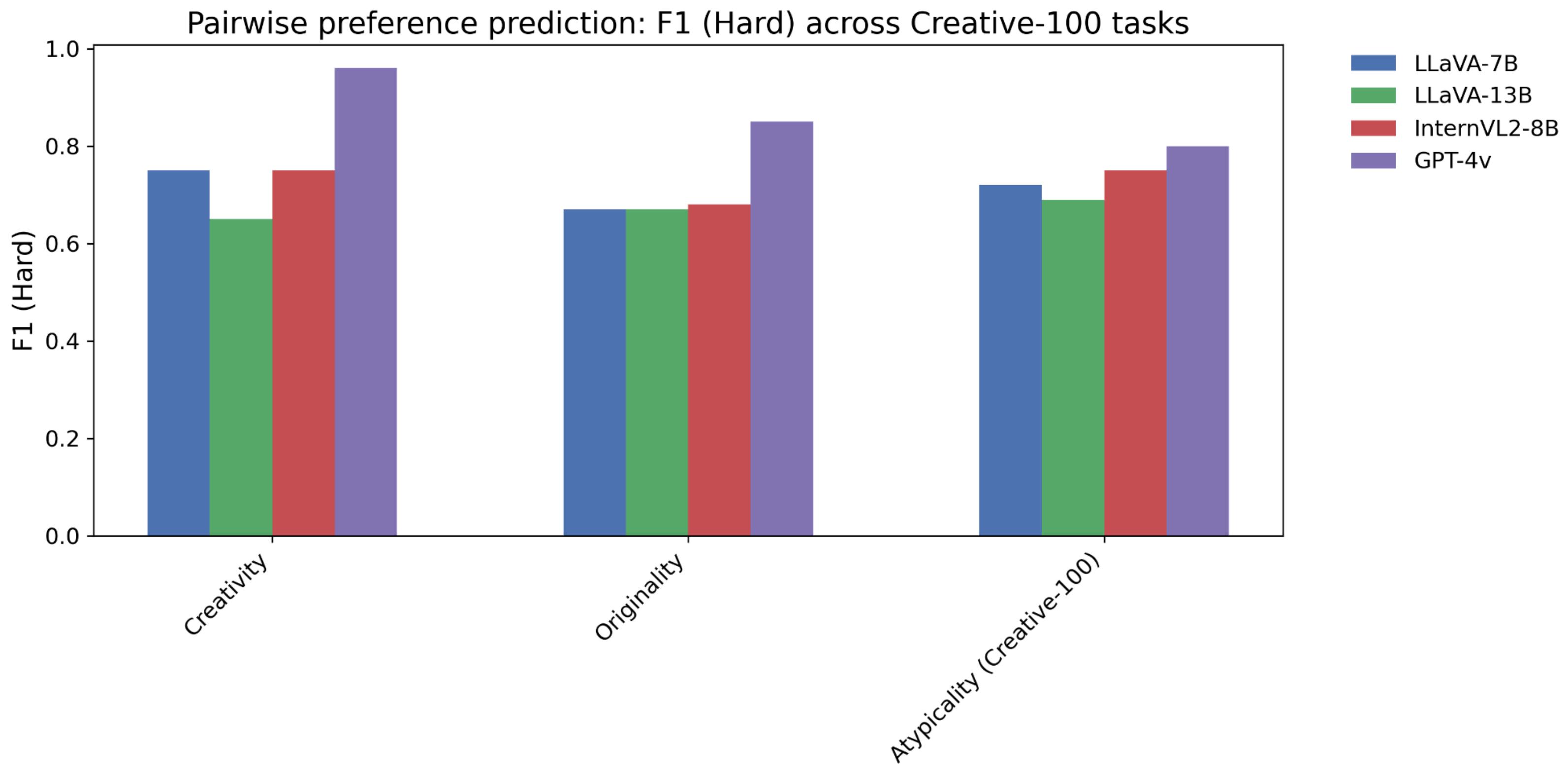
# Ads Creativity Dataset Construction

- **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

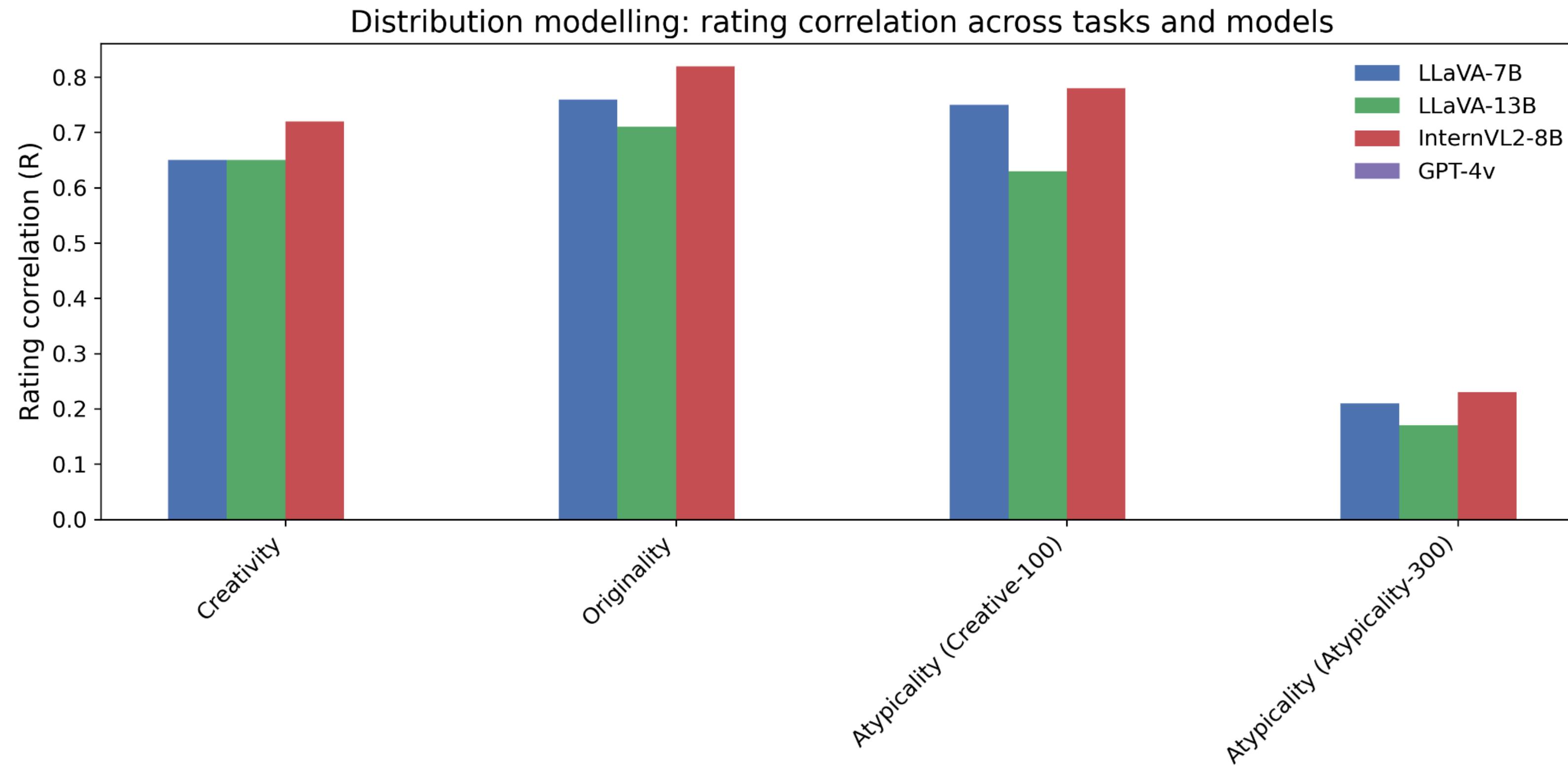
- **Distribution Modeling:** Simulate human group behavior
  - Task Setup: VLM simulates scores distribution by repeated prompting (for creativity, atypicality, originality)
  - Metrics
    - KL Divergence (human scores V.S. VLM scores)
    - Spearman's Correlation (human scores V.S. VLM scores)
- **Pairwise 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



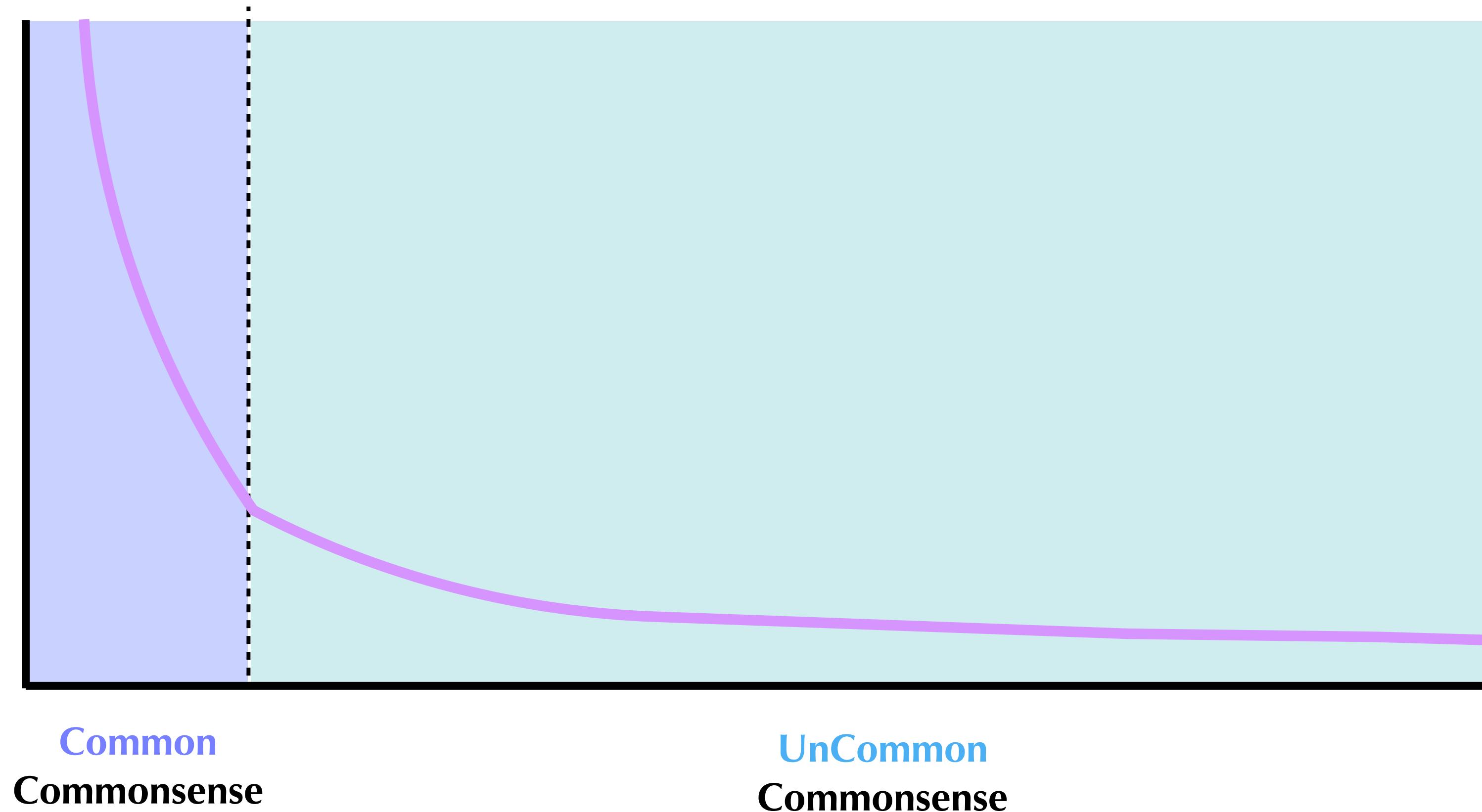
- Okay results for distributional modeling tasks, similar to CFC.

# Ads Creativity: Findings

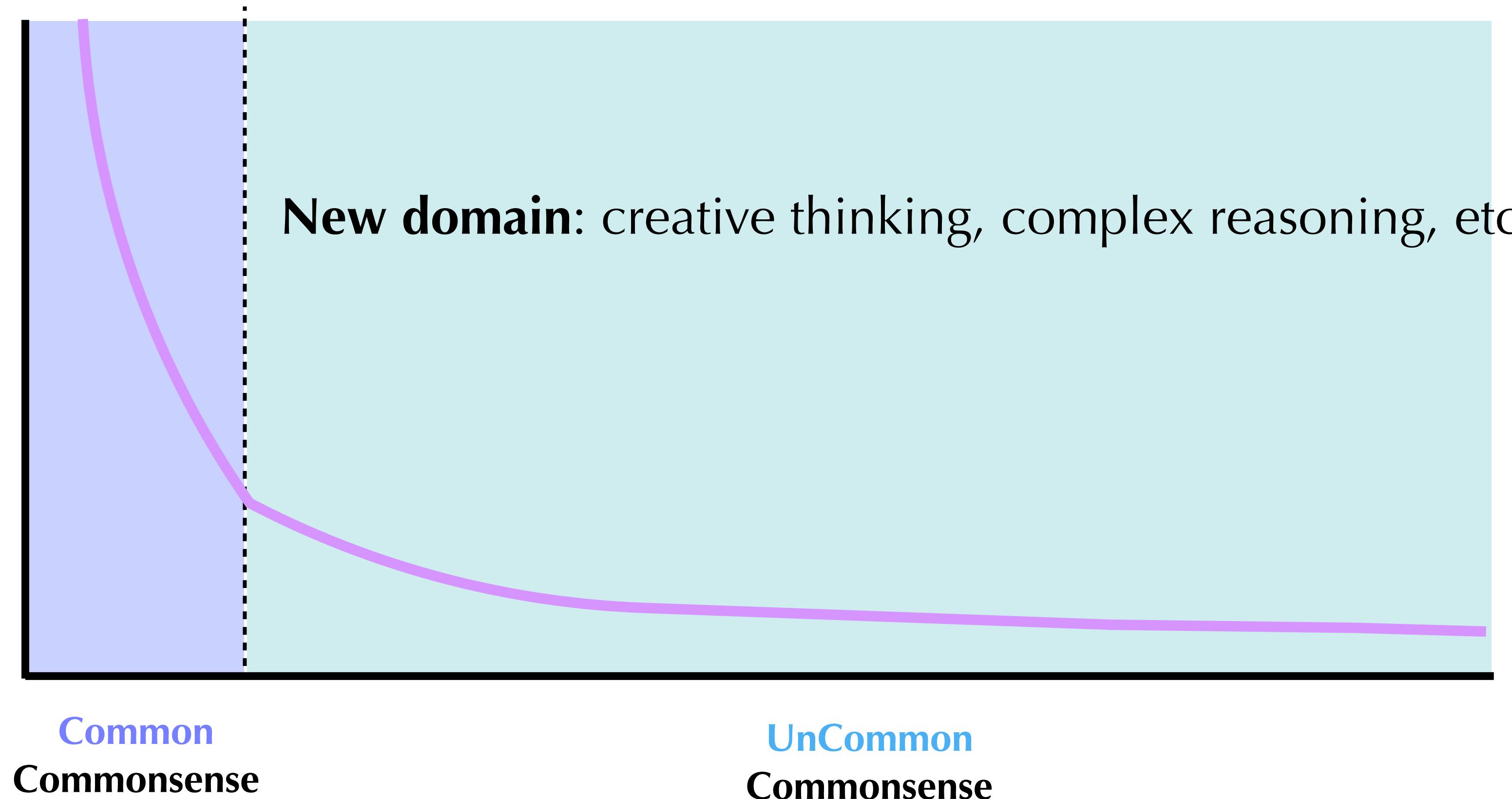
- The disagreement task is the hardest.
- The model can't predict when there is greater disagreement/subjectivity for specific instances.

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

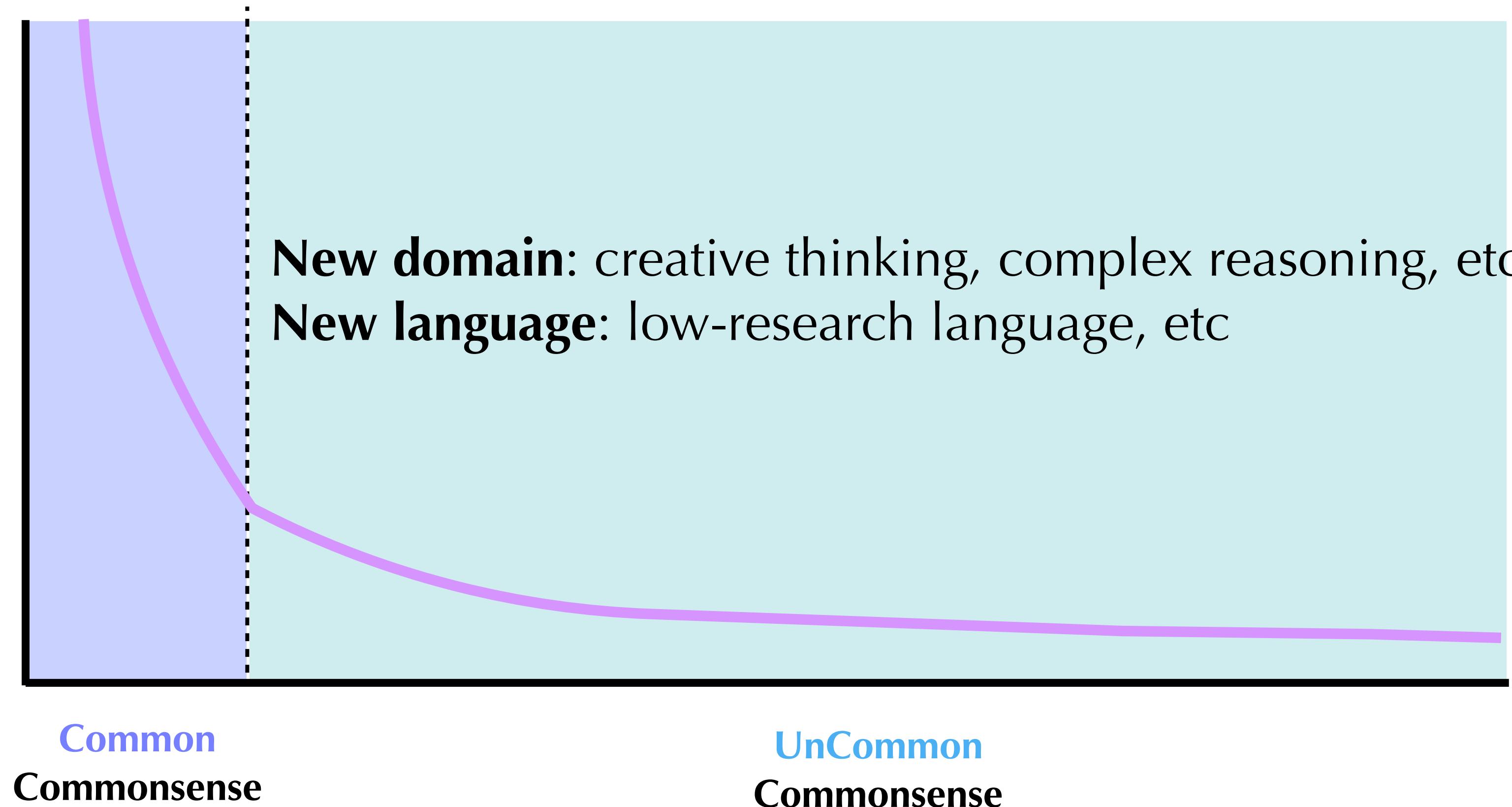
# Why is performance so bad?



# Why is performance so bad?

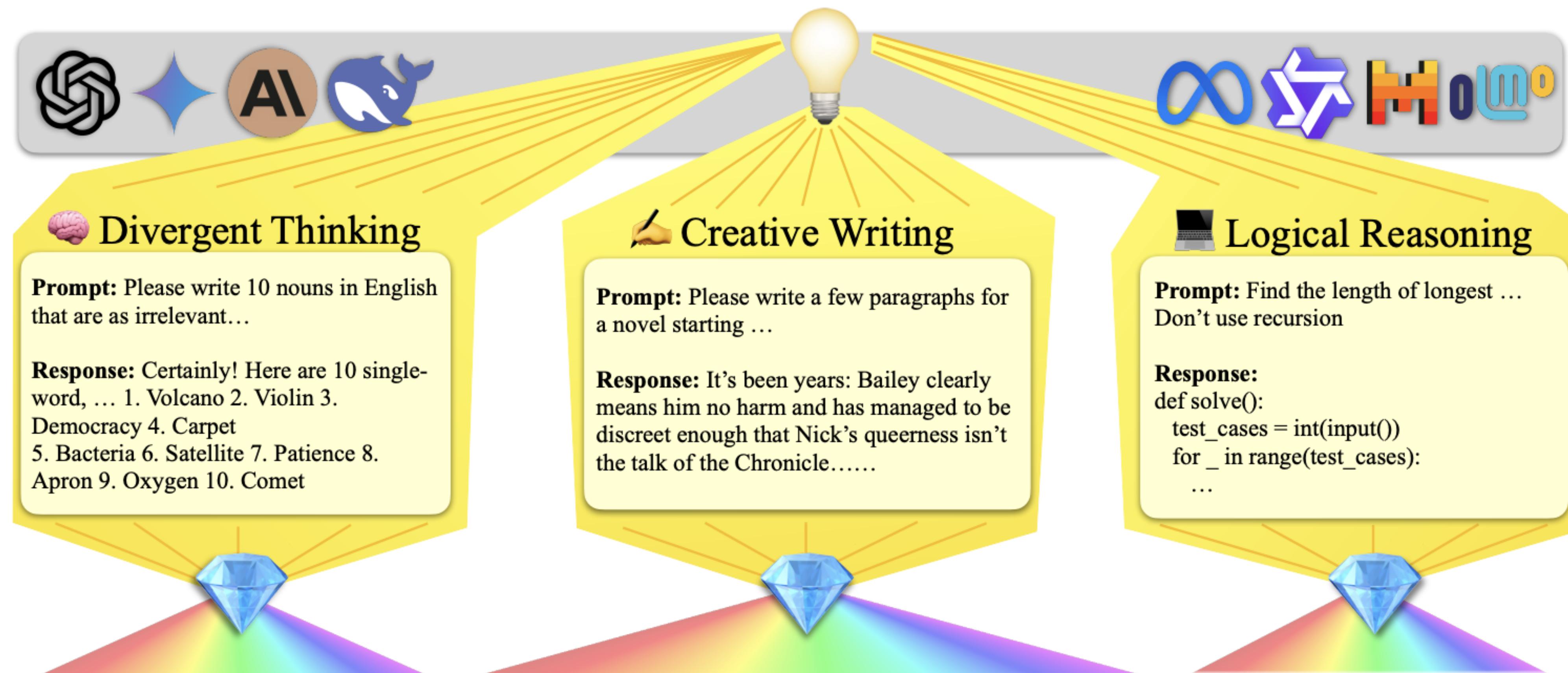


# Why is performance so bad?



# 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

- 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]

Complex reasoning

Multi-lingual

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

César Guerra-Solano<sup>1</sup>, Zhuochun Li<sup>2</sup>, Xiang Lorraine Li<sup>1</sup>

CS  
Undergrad!



<sup>1</sup>Department of Computer Science,

<sup>2</sup>Department of Informatics and Networked Systems  
University of Pittsburgh, PA, USA

# 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

<b>PARTS OF A TABLE SETTING</b> FORK, GLASS, NAPKIN, PLATE
<b>KINDS OF DIGITAL STORAGE</b> CARD, CLOUD, DISK, DRIVE
<b>UNITS OF VOLUME PLUS LETTER</b> BOUNCE, GALLEON, PINOT, QUARTZ
<b>INCREASED, WITH “UP”</b> FLEW, ROSE, SHOT, THRUST

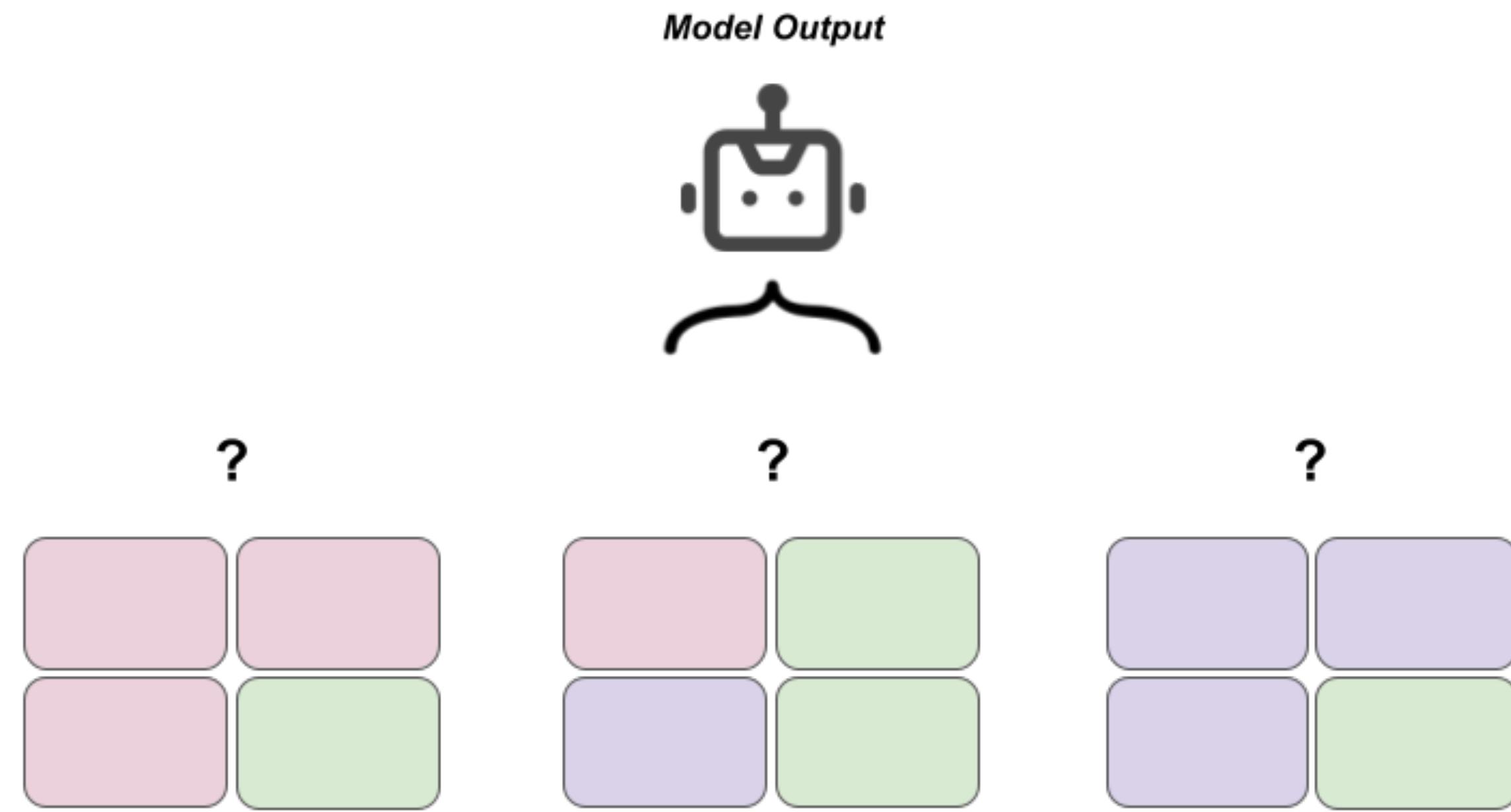
Mistakes Remaining: ● ● ● ●

[Shuffle](#)   [Deselect All](#)   [Submit](#)

[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

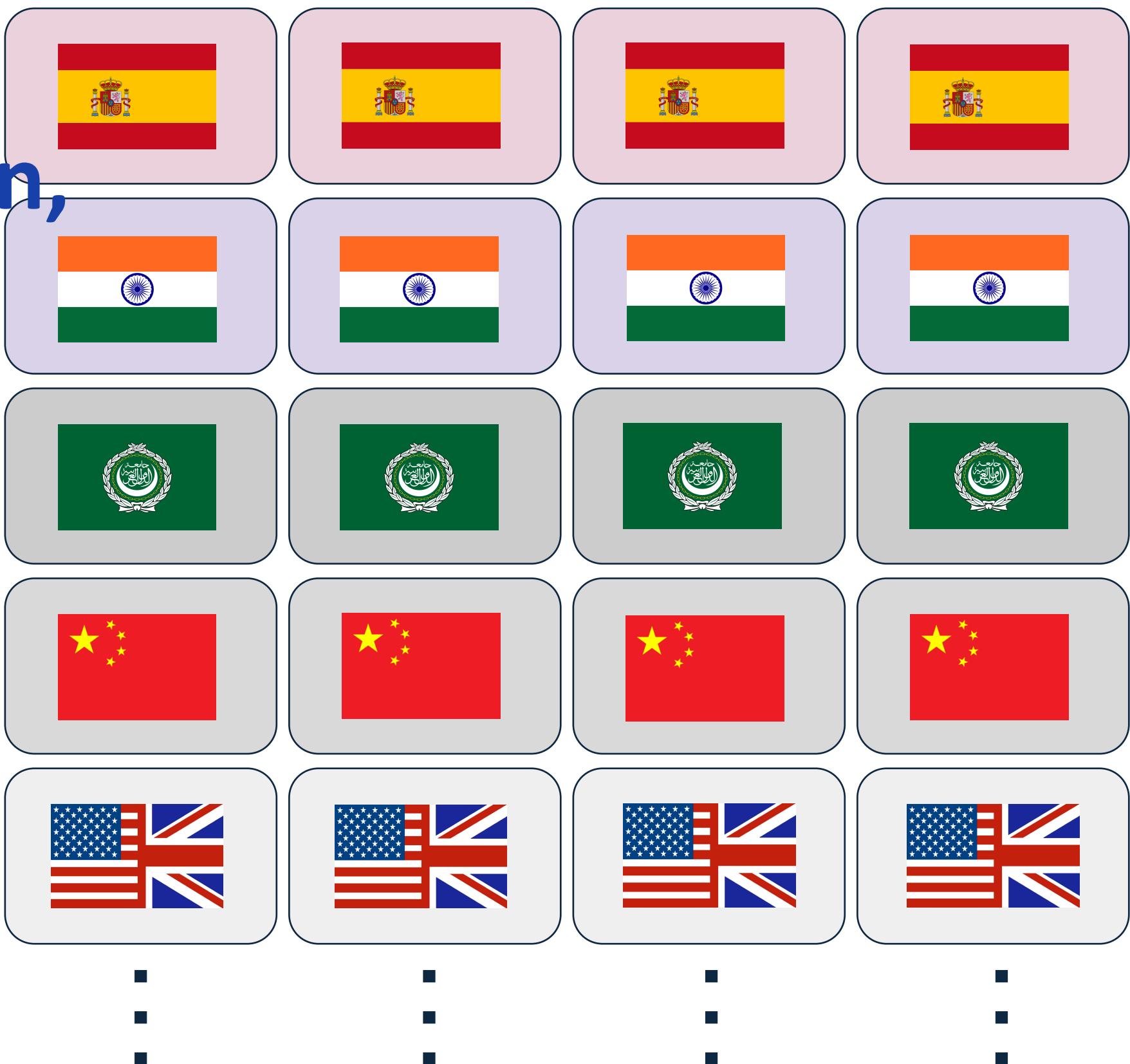


# Global Group

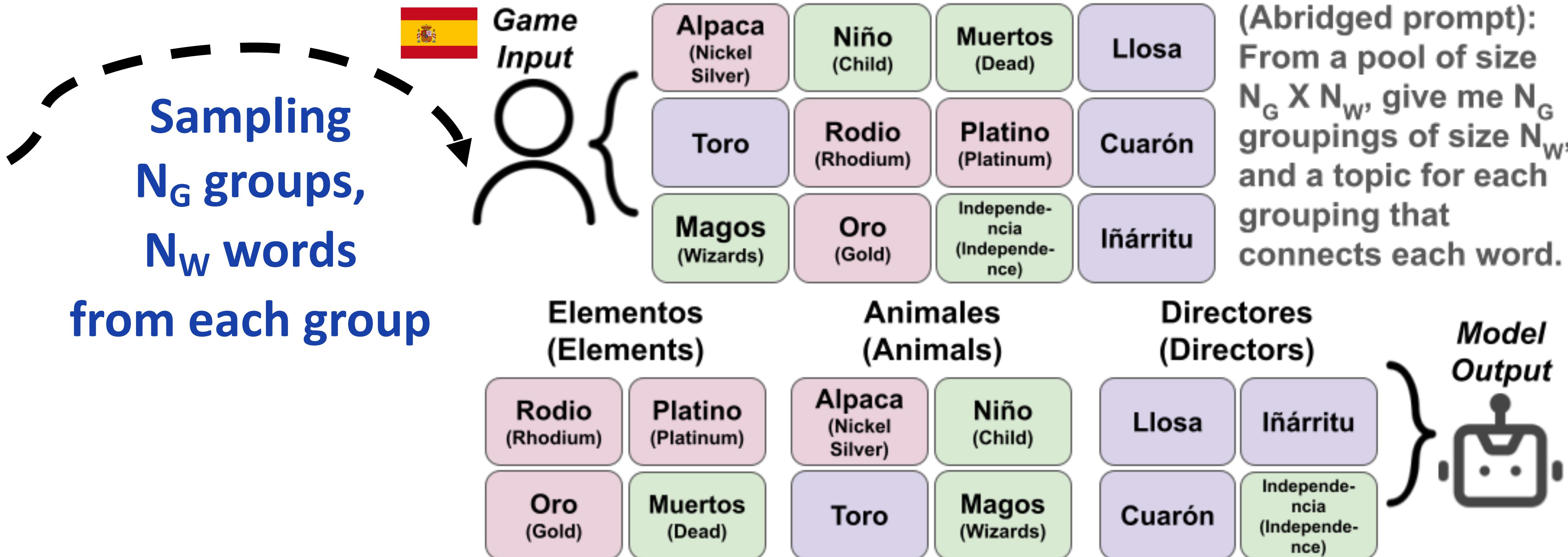


Native speaking  
annotators

Grouping creation, translation,  
tagging



# Global Group



# Global Group

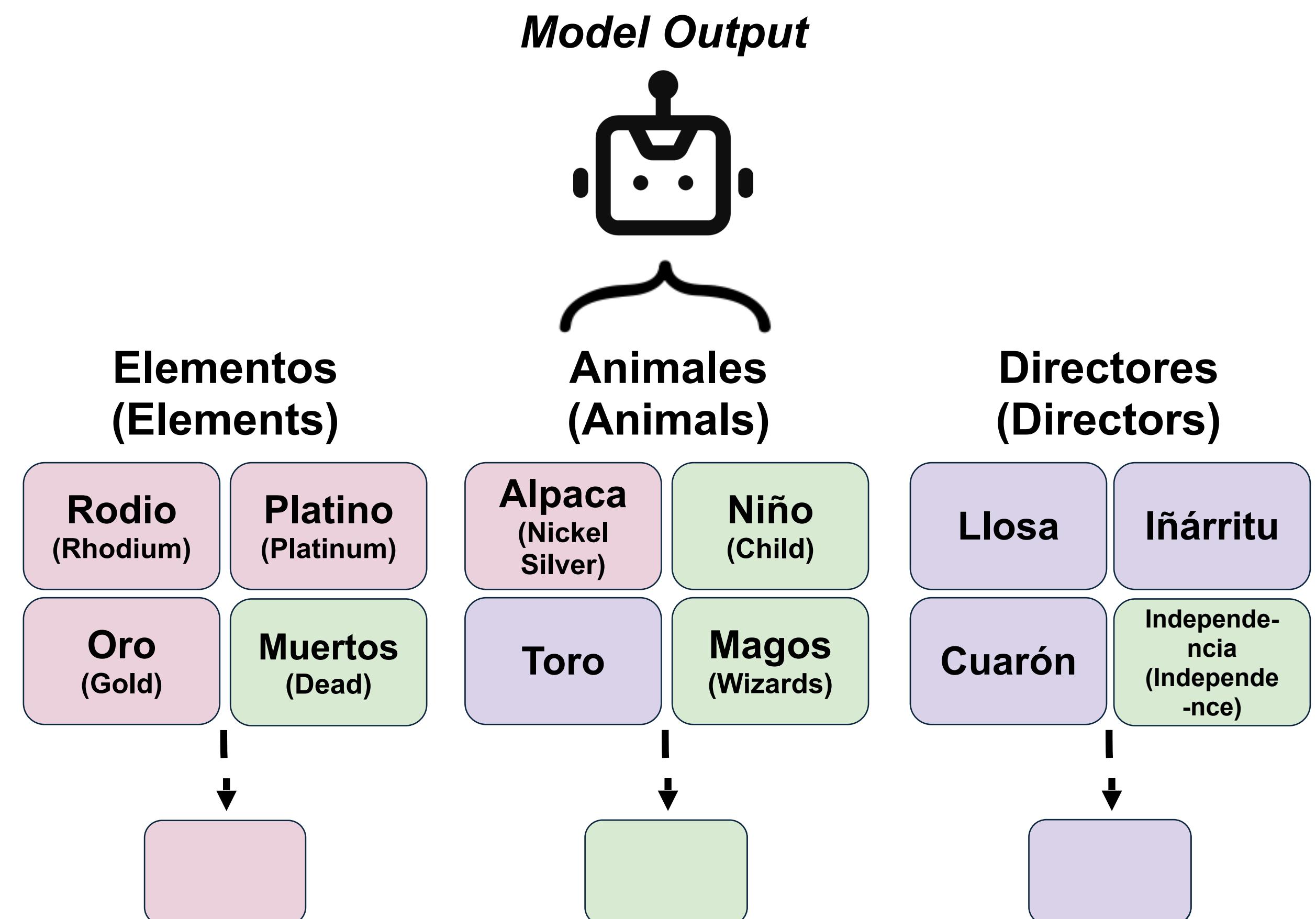
<b>Dataset Background Subsets</b>	<b>Available Languages</b>	<b>Non-Culturally-Related Group Topic (in en)</b>	<b>Culturally-Related (C-R) Group Topic (in en)</b>	<b>C-R Group Amount</b>	<b>Total Group Amount</b>	<b>Game Settings (<math>N_{words} \times N_{groups}</math>)</b>
en	en	Professions	American 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

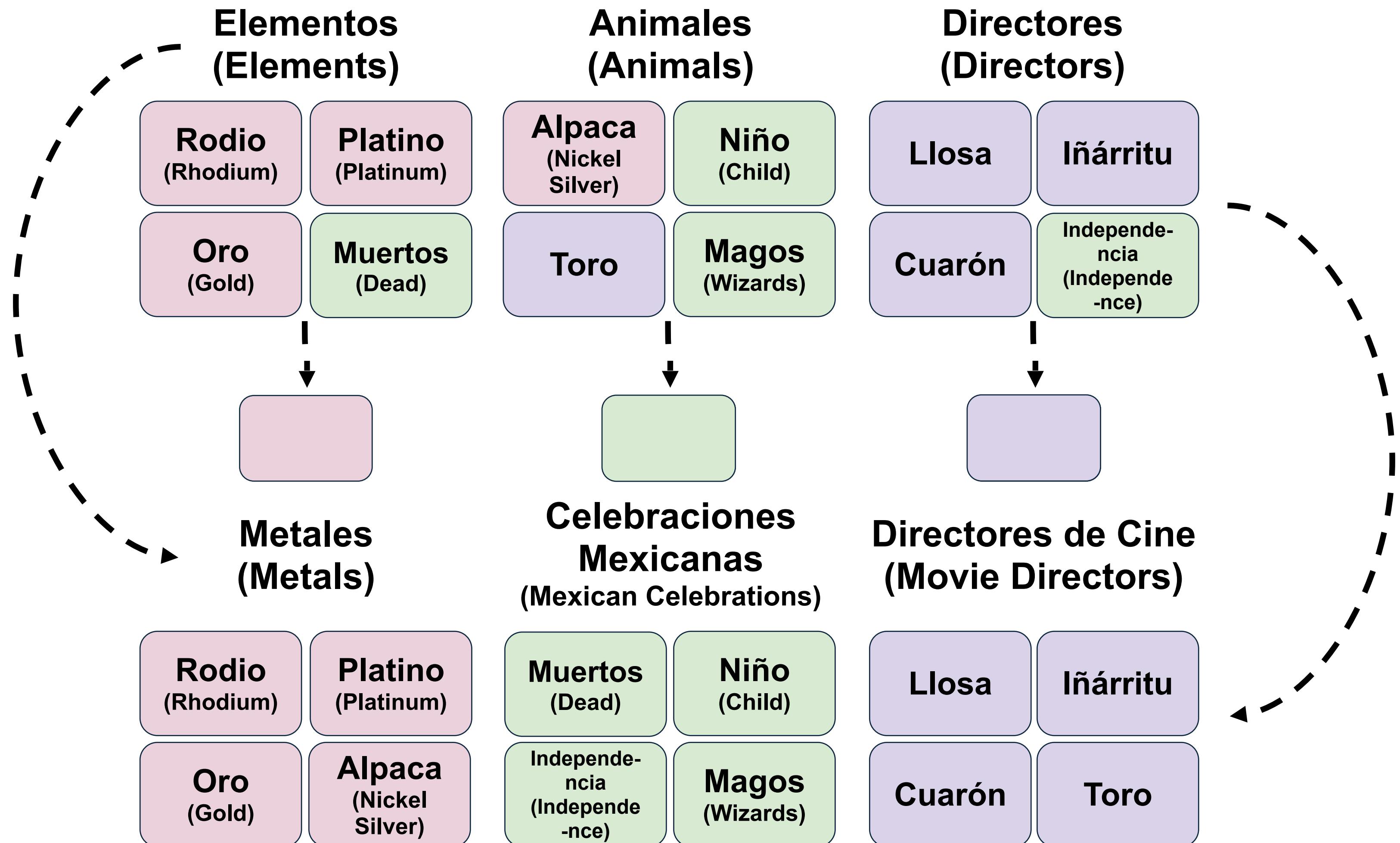
# Model Evaluation

- Match attempts to true groups based on majority



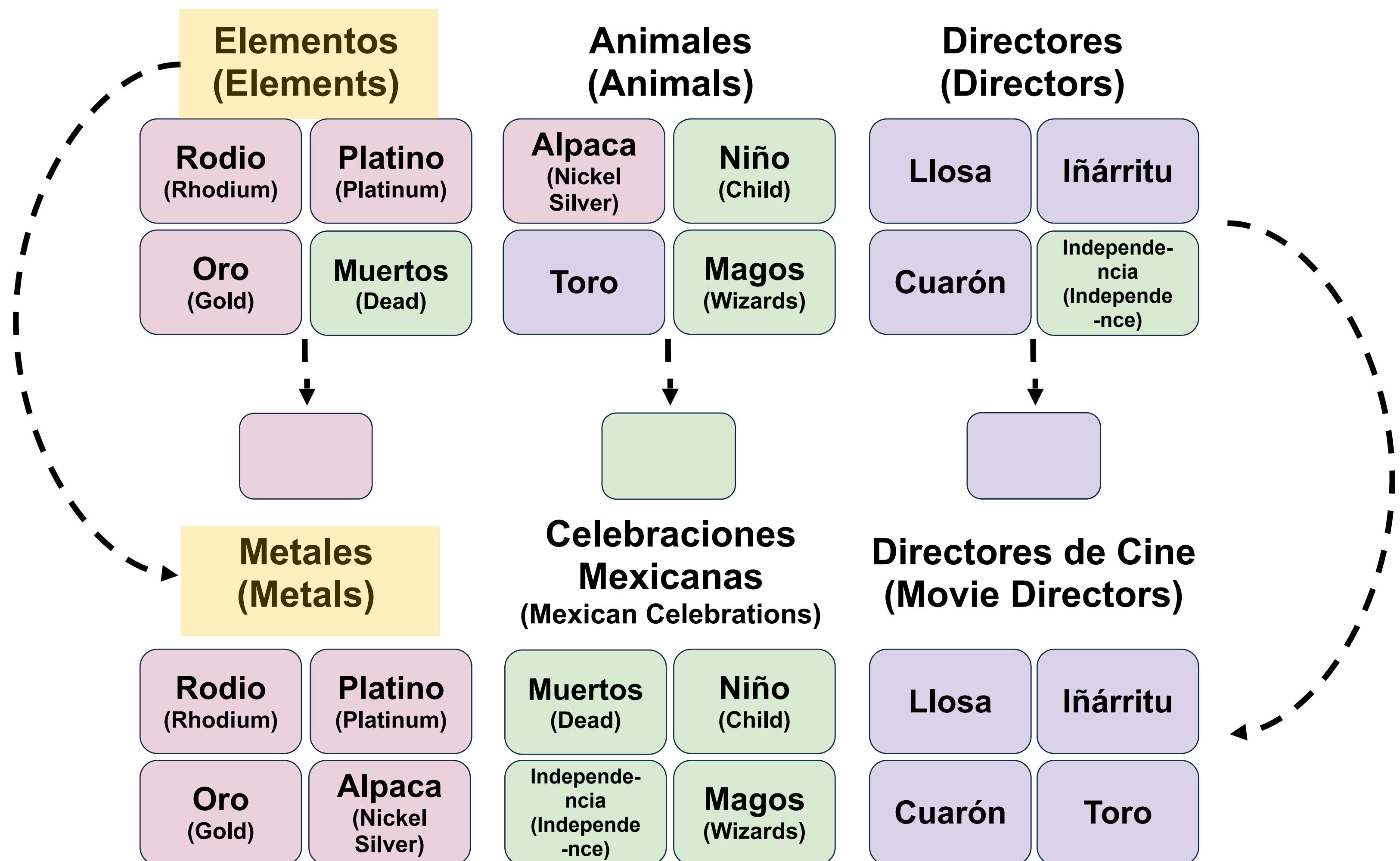
# Model Evaluation

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



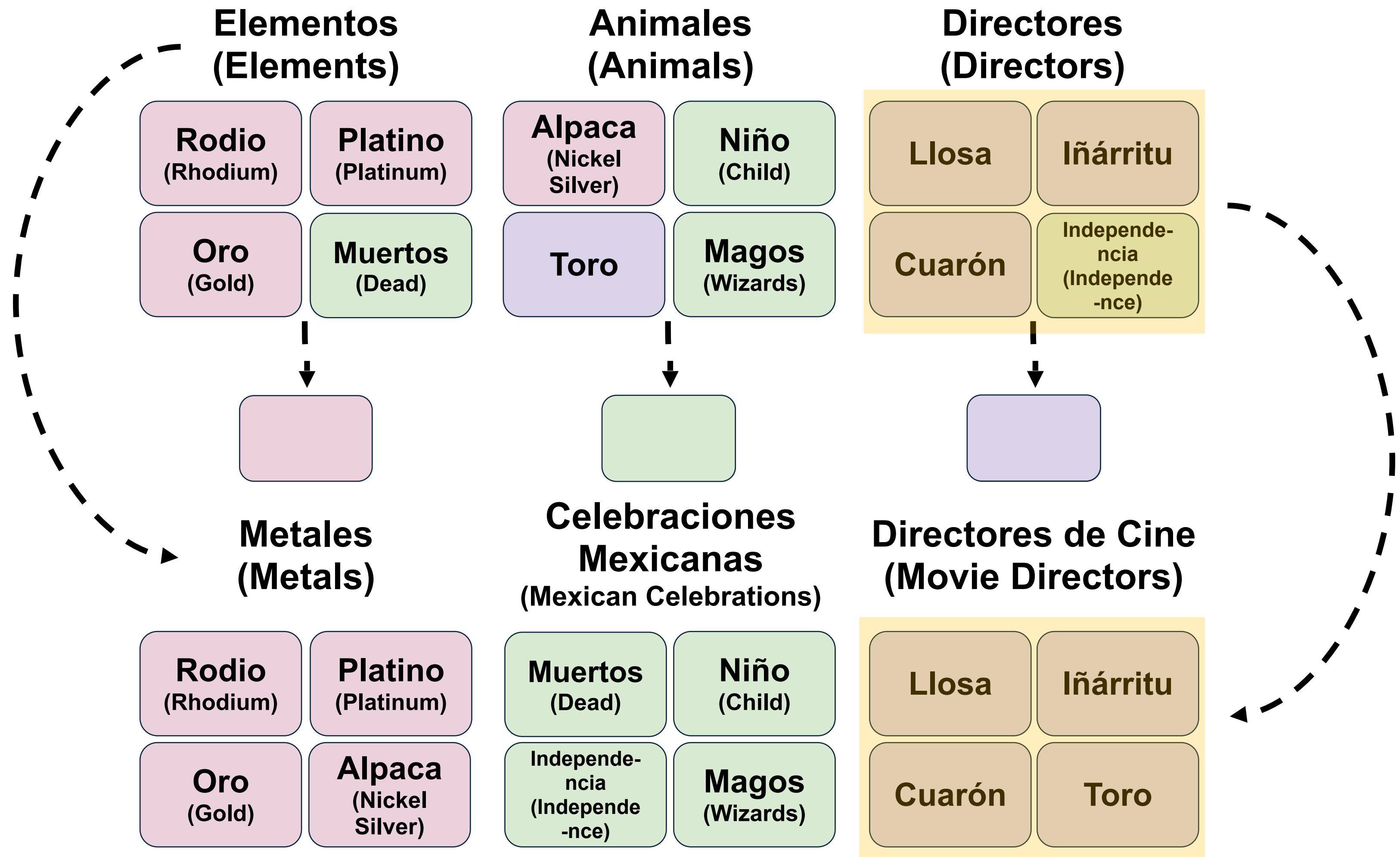
# Model Evaluation

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

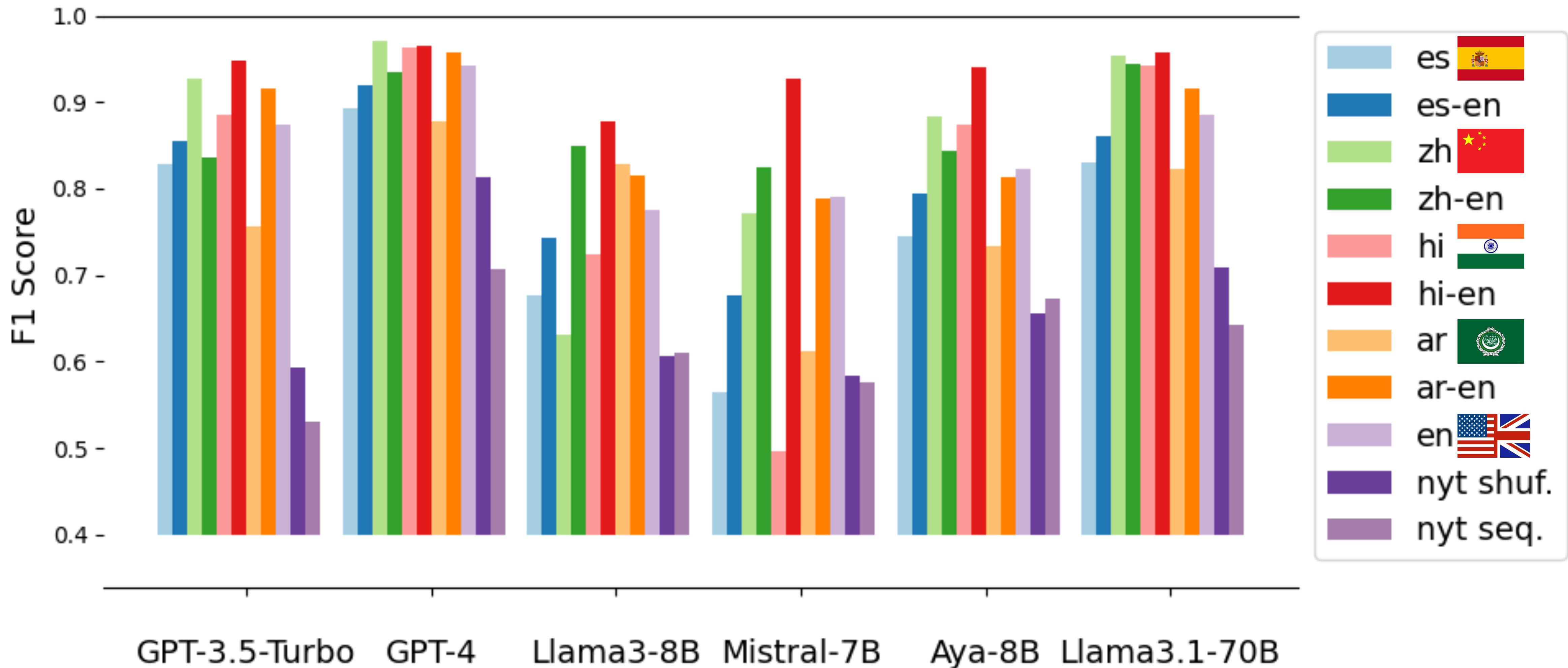


# Model Evaluation

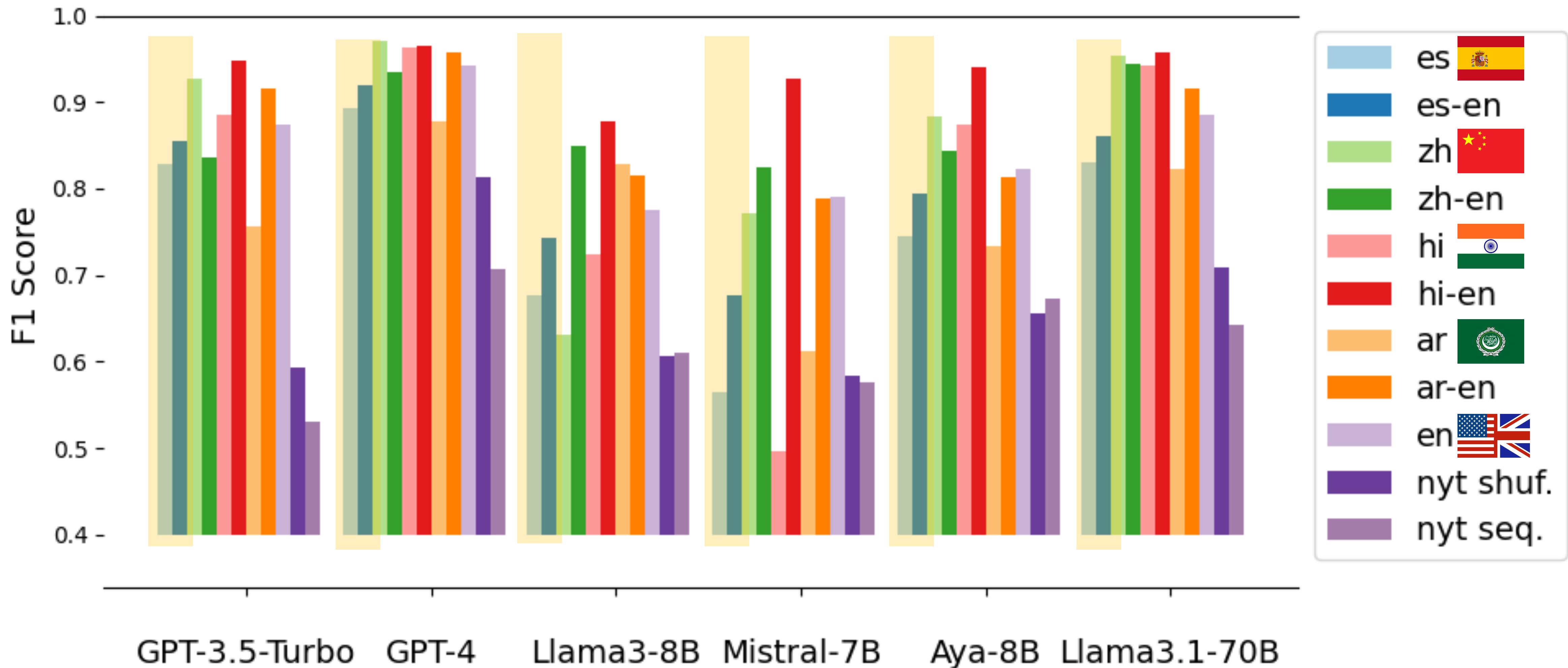
- Match attempts to true groups based on majority
- Scoring topics and groups:
  - **Topic Achieved Score**
  - **F1 Score**



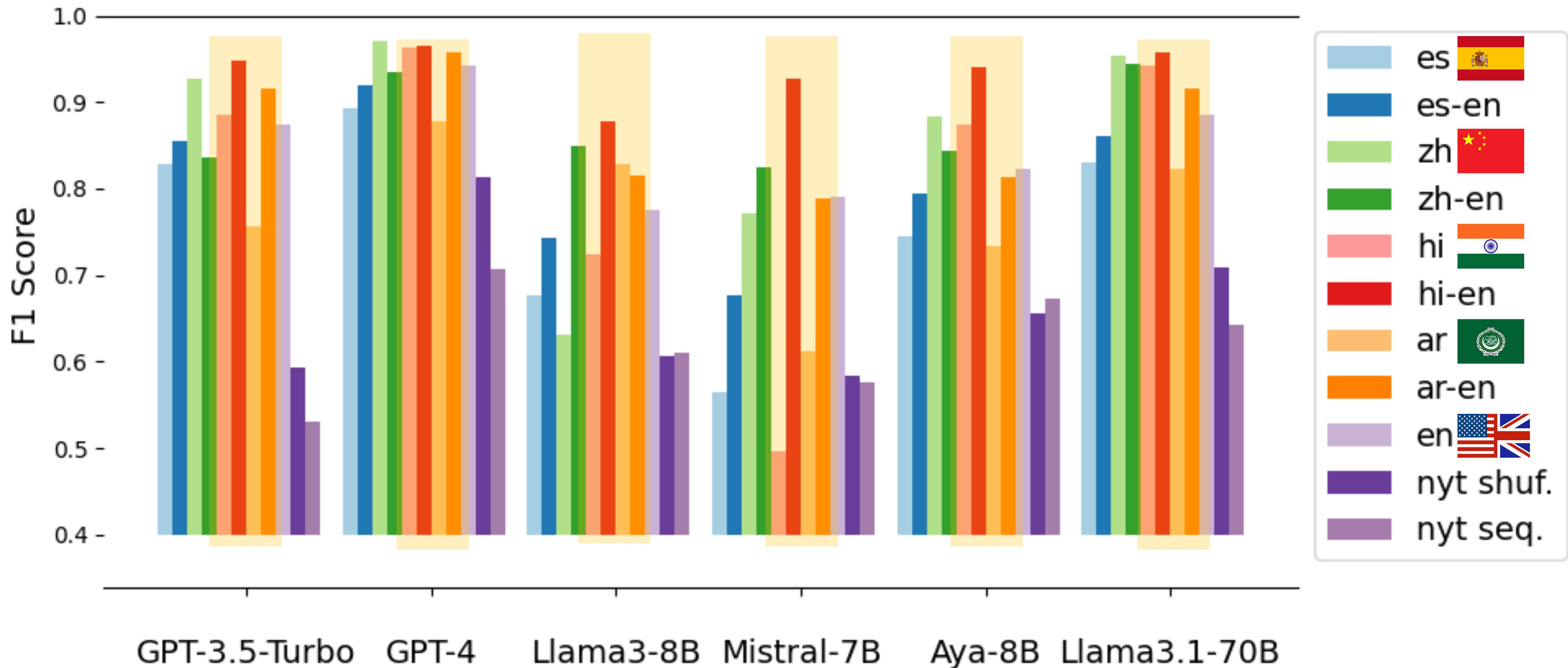
# English Representations Improve Performance



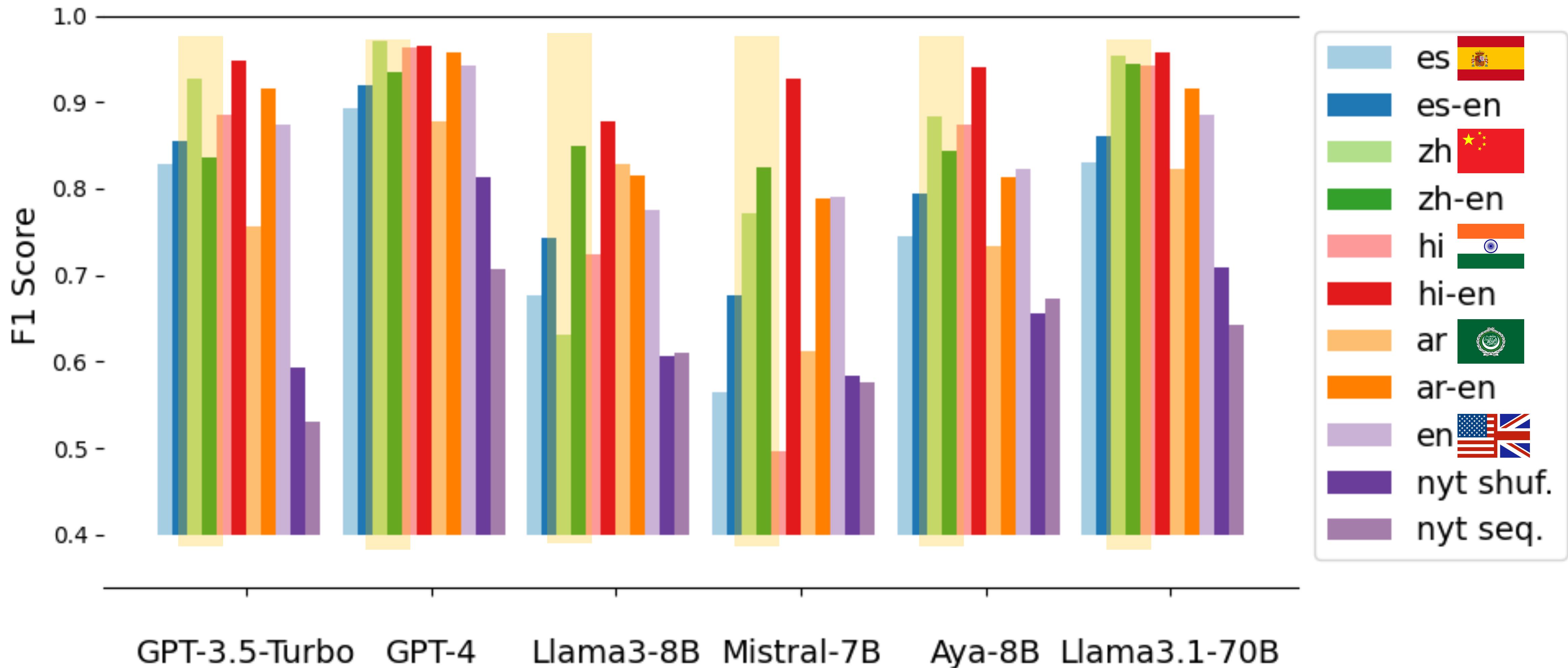
# English Representations Improve Performance



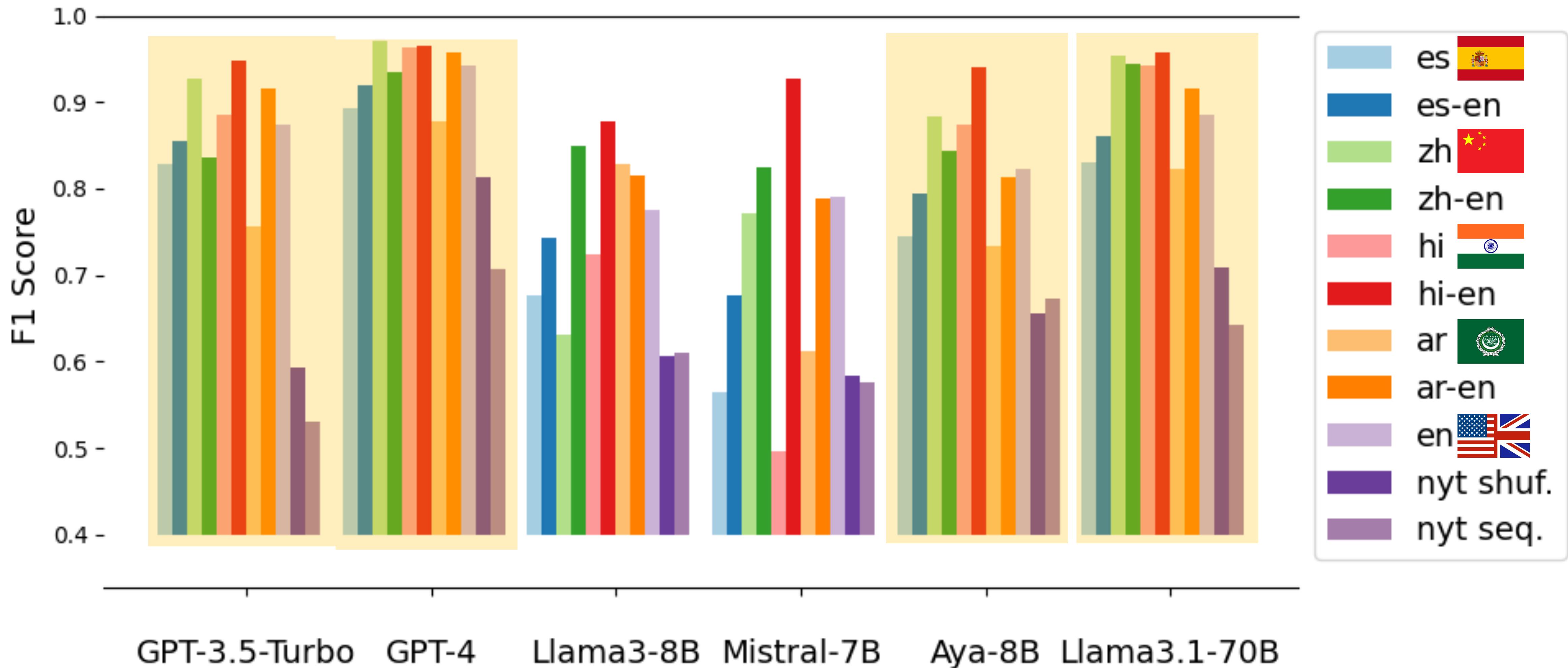
# English Representations Improve Performance



# English Representations Improve Performance

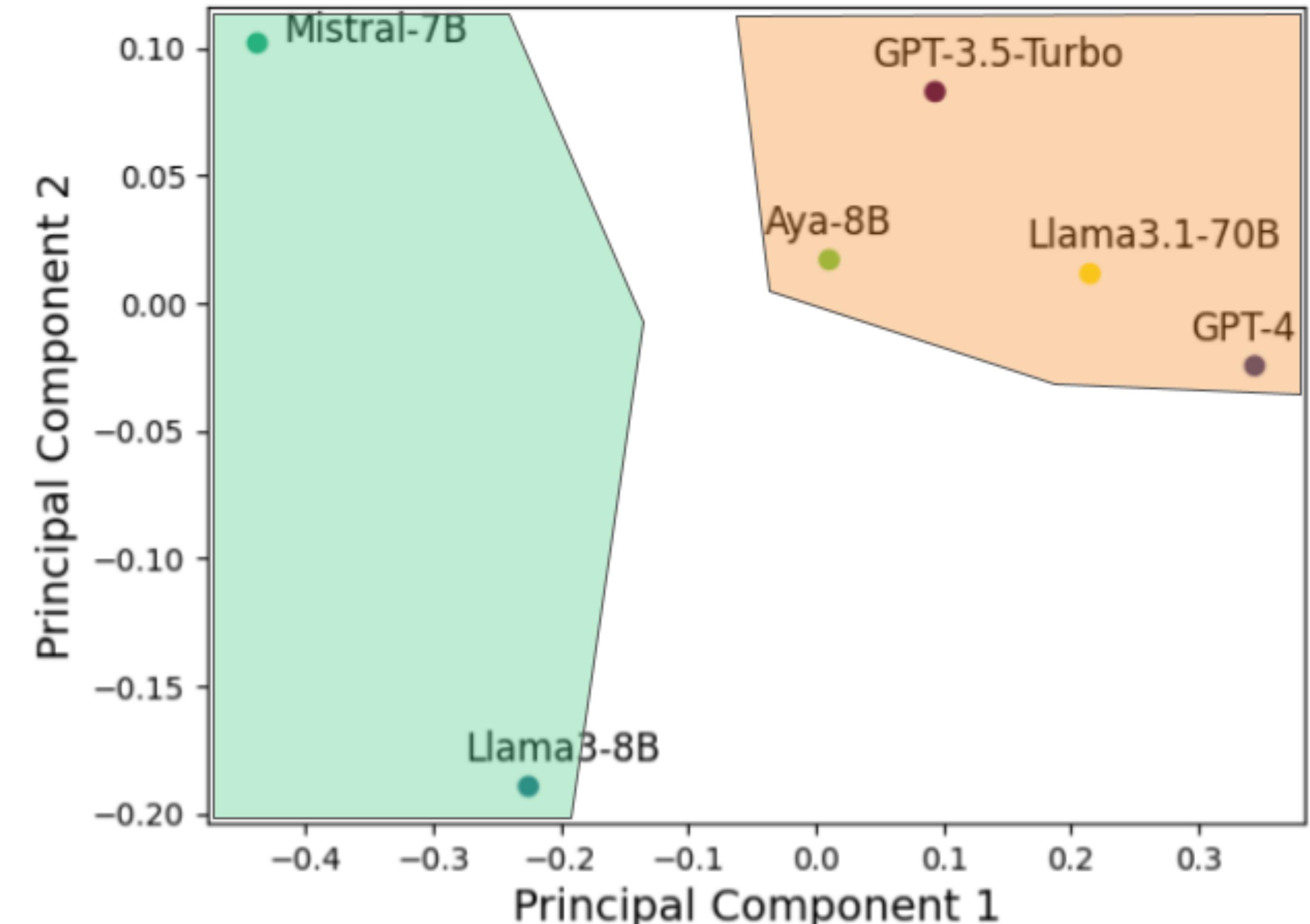


# English Representations Improve Performance



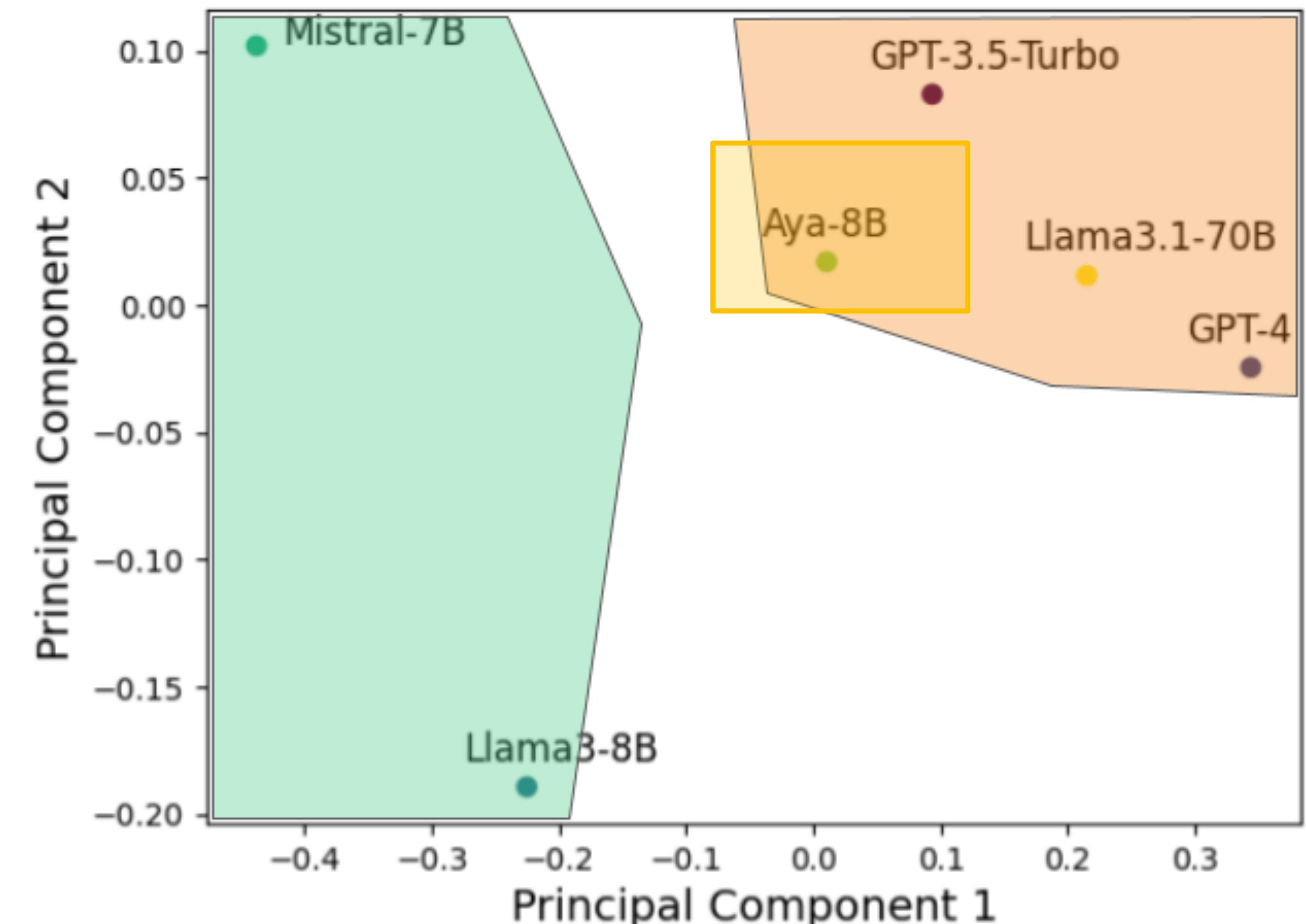
# Multilingual Training Greatly Improves Small Models

- 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

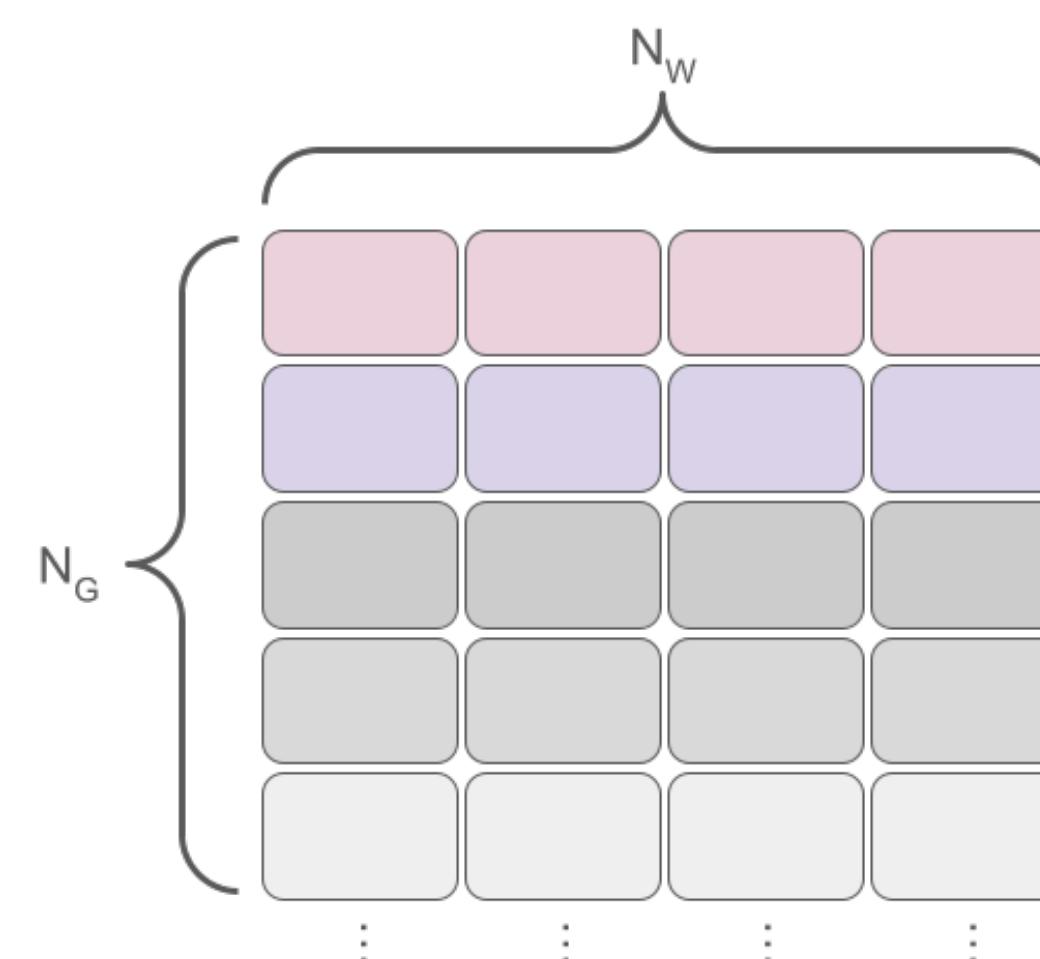
- We hypothesize, outside of language, **key game parameters are associated with difficulty**
  - Need for a controlled setting

# 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

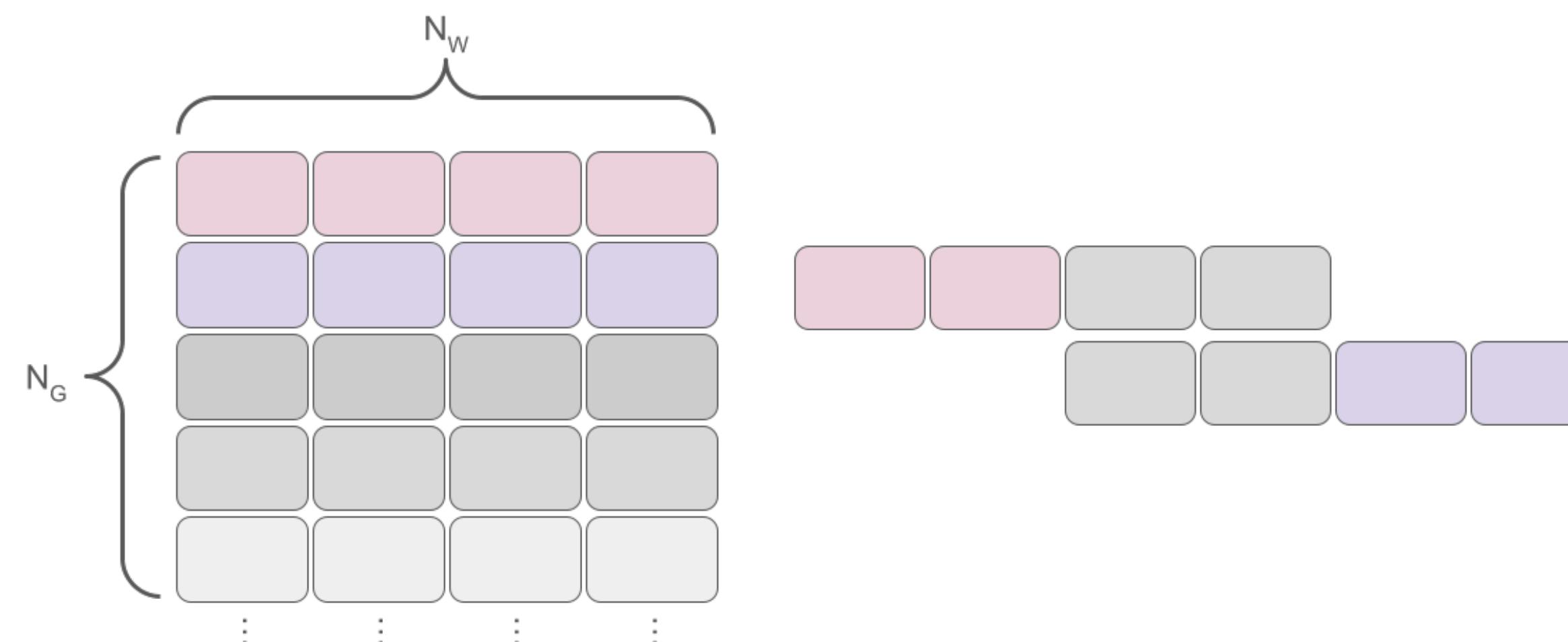
# Key Game Parameters Correlate with Difficulty

- 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



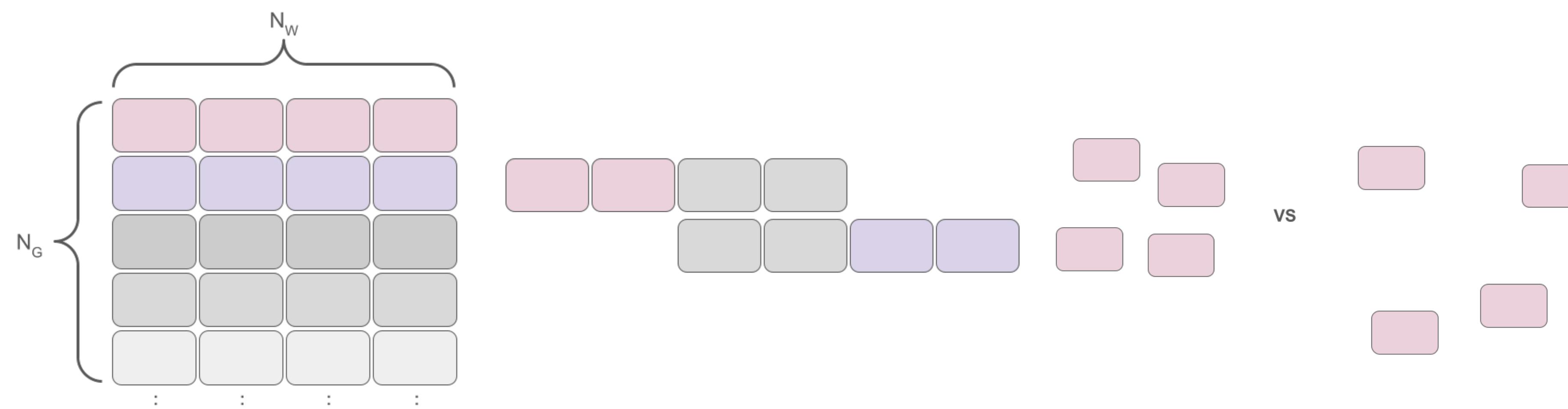
# Key Game Parameters Correlate with Difficulty

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

# Key Game Parameters Correlate with Difficulty

- Inconsistent trend with group size

Models	Average F1 Score		
	2	3	4
<i>Group Size</i>			
GPT-3.5-Turbo	0.881	<b>0.890</b>	0.884
GPT-4	0.925	0.940	<b>0.949</b>
<i>Group Count</i>			
GPT-3.5-Turbo	<b>0.930</b>	0.883	0.842
GPT-4	<b>0.962</b>	0.940	0.912
<i>Adjusted Rand Index Range</i>			
GPT-3.5-Turbo	0.924	0.927	<b>0.951</b>
GPT-4	0.969	0.974	<b>0.990</b>
<i>Candidate Proposal Overlap Range</i>			
GPT-3.5-Turbo	<b>0.932</b>	0.873	0.813
GPT-4	<b>0.977</b>	0.923	0.863
			0.831
			0.859

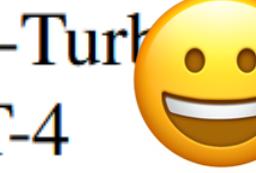
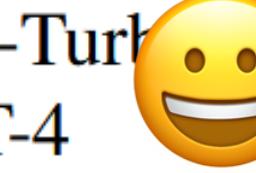
# Key Game Parameters Correlate with Difficulty

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

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

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

# Performance is Related to Group Content

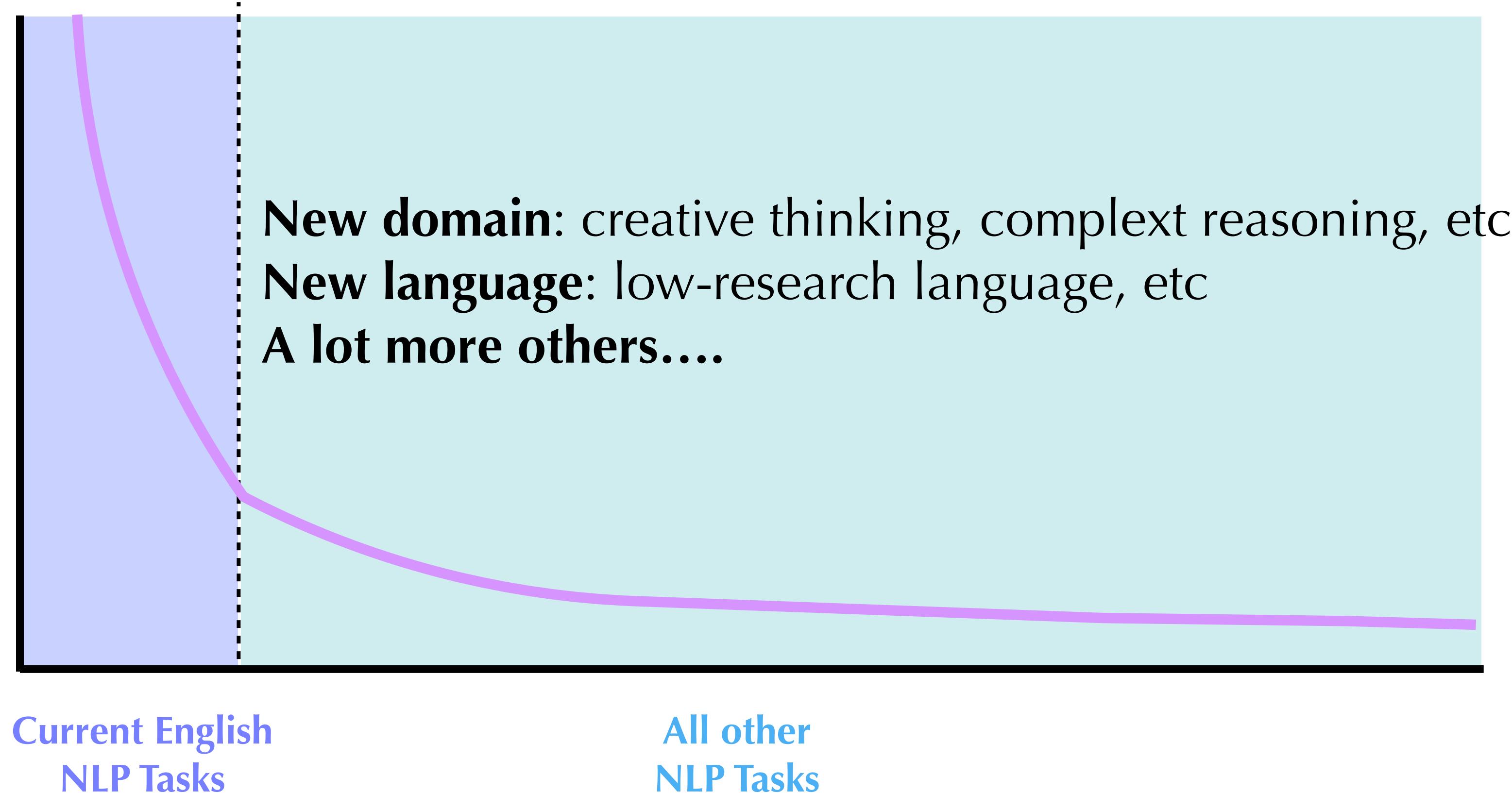
Models	Mean F1 Score in Groups of Tag		
	General Knowledge	Cultural and Pop Culture Knowledge	Linguistic
Group Tag			
GPT-3.5-Turbo	<b>0.902</b>	0.871	0.826
GPT-4	<b>0.954</b>	0.937	0.916
Llama3-8B	<b>0.804</b>	0.767	0.724
Llama3.1-70B	<b>0.935</b>	0.900	0.859
Mistral-7B	<b>0.779</b>	0.694	0.669
Aya-8B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



Current English  
NLP Tasks

All other  
NLP Tasks

# 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]

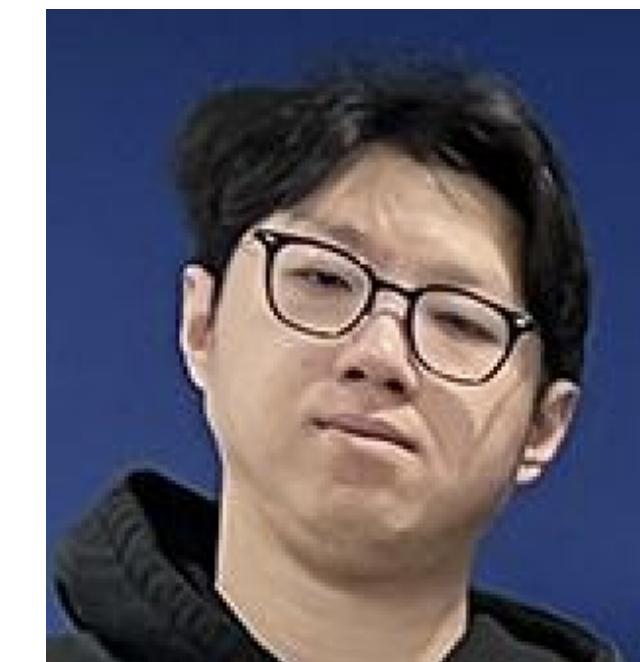
## Benchmark: Long-tail Question: Commonsense Reasoning Evaluation

- 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