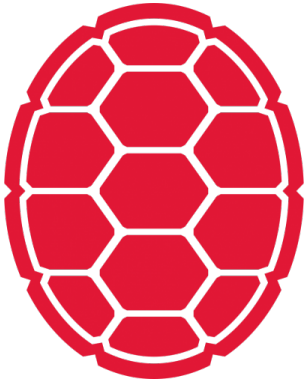
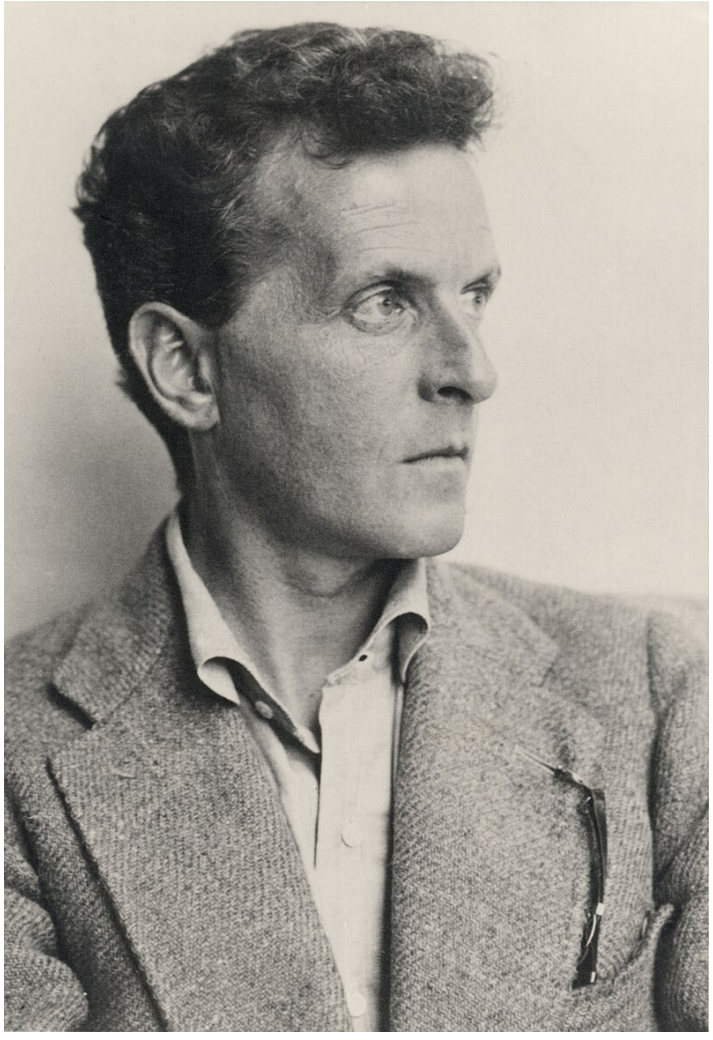


Language Games

Jordan Boyd-Graber
University of Maryland
2023

Why Language Games?



- Wittgenstein: Language is a product of the context in which it is used
- Interaction allows us to figure out how to use language
- Score lets us know how well we're doing

Image: DeepMind

Peter Morgan/Reuters

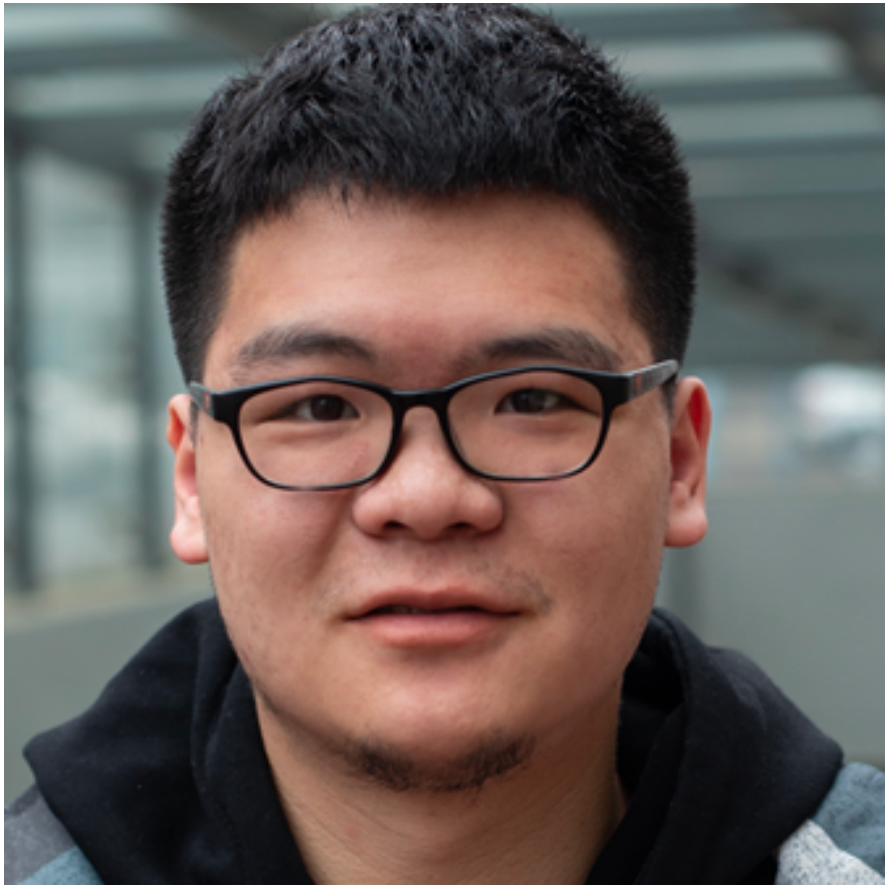
Today's Topics

- Fact Checking
  - Getting harder examples
  - Playing well with humans
- Question Answering
  - Getting harder examples (accuracy and calibration)
  - Measuring human vs. computer accuracy and calibration
- Negotiation
  - Measuring human vs. computer ability
  - Detecting lies

Fact Checking with Humans in the Loop

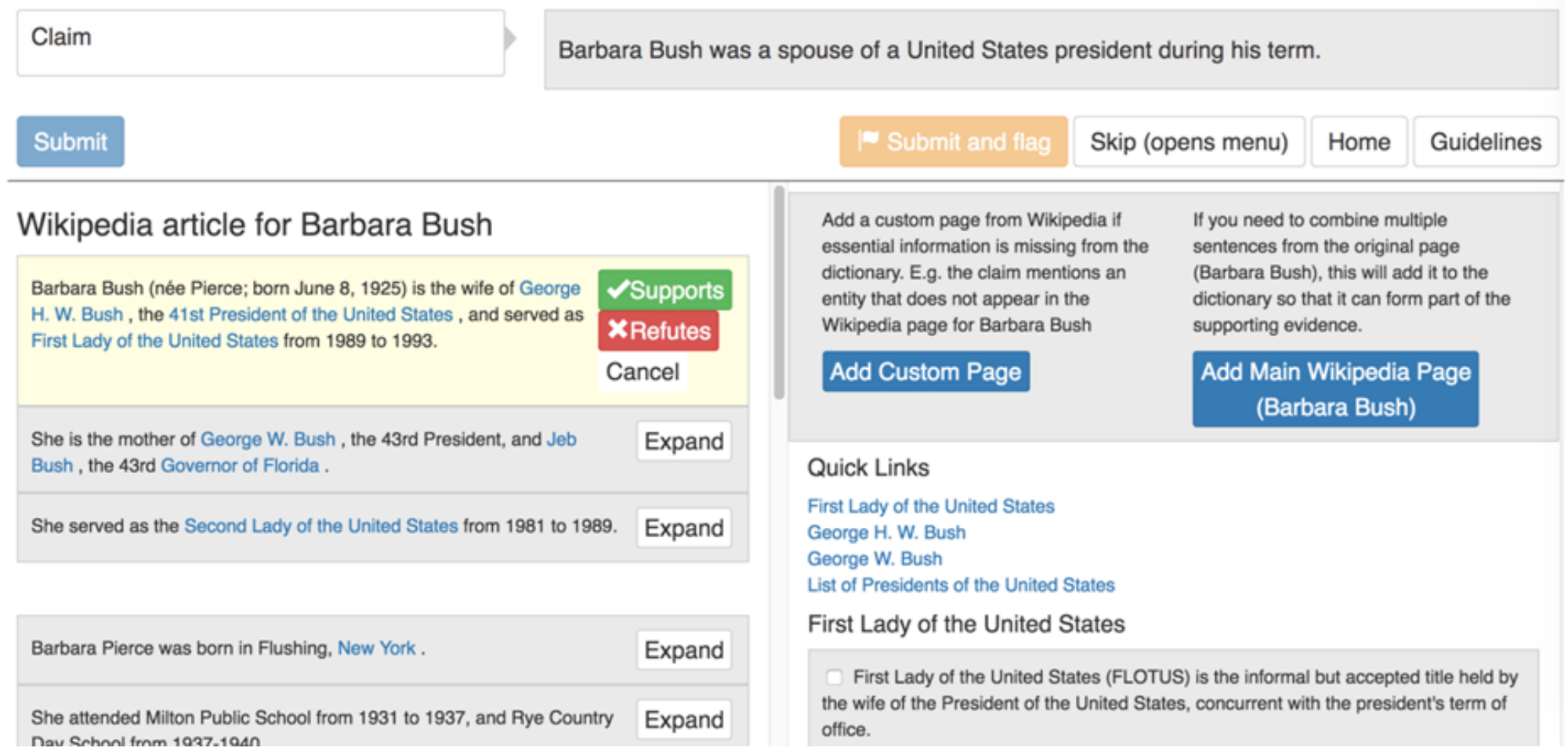
Large Language Models Help Humans Verify Truthfulness—Except When They Are Convincingly Wrong

Chenglei Si[1] **Navita Goyal**[2] **Sherry Tongshuang Wu**[3]
Chen Zhao[4] **Shi Feng**[5] **Hal Daumé III**[2,6] **Jordan Boyd-Graber**[2]
[1]Stanford University [2]University of Maryland [3]Carnegie Mellon University
[4]NYU Shanghai [5]New York University [6] Microsoft Research
clsi@stanford.edu



Research Questions

- Are natural language explanations more effective than retrieved passages for human fact-checking?
- Can contrastive explanations—arguing for or against a fact being true—mitigate over-reliance and be more effective than non-contrastive explanations?
- Are there complementary benefits in presenting both natural language explanations and retrieved passages?



Fact **E**xtraction and **VER**ification

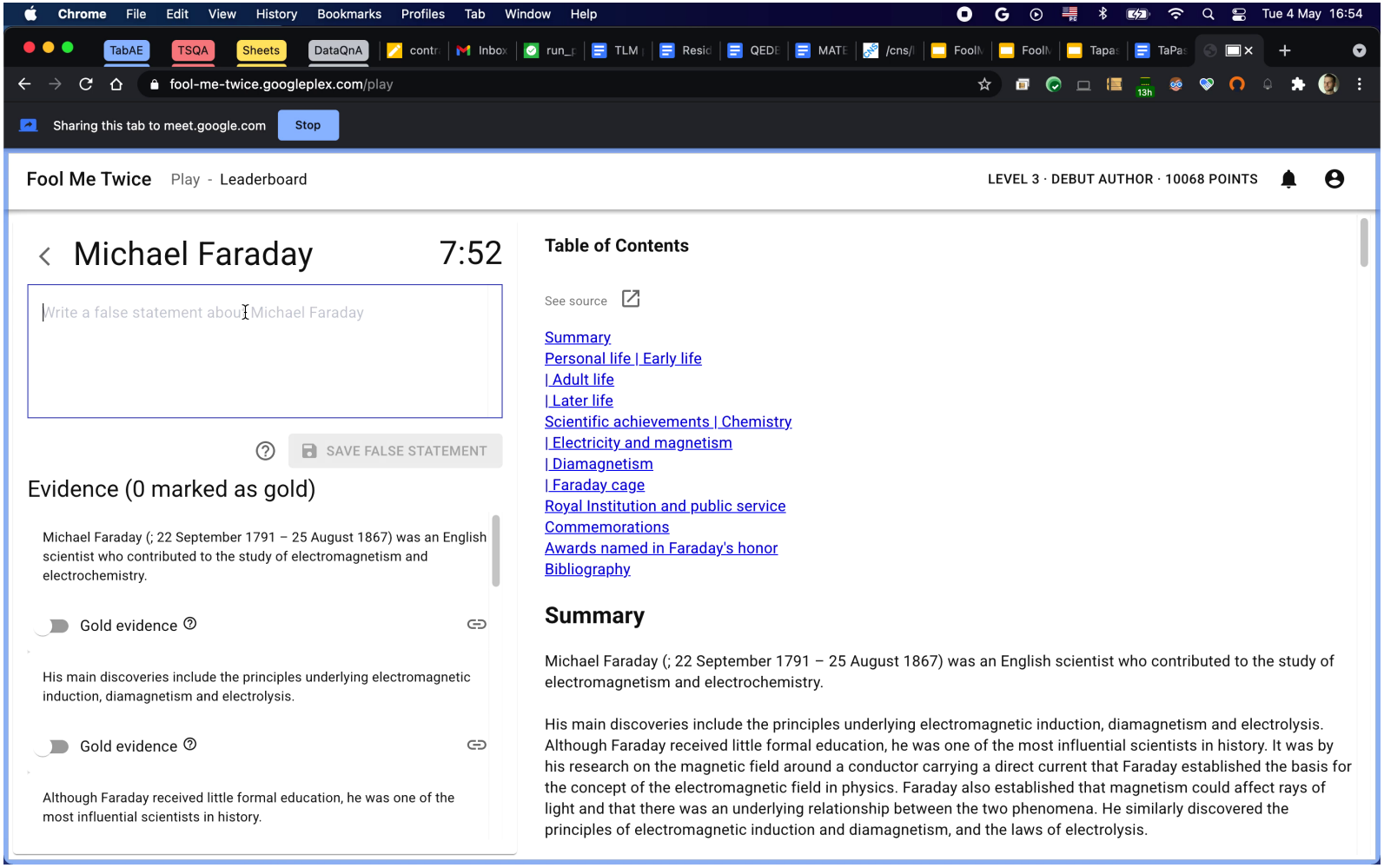
FEVER categories

- Examples from FEVER (Thorne et al., 2018)
- Supported:
  - Woody Allen is a person.
  - The Shining was directed.
  - François de Belleforest wrote.
- Not Enough Info:
  - Lisa Kudrow was in a car.
  - Tipper Gore was curated to Al Gore.
  - International Relations includes animals.
- Refuted:
  - Tipper Gore was created in 1048.
  - Alpha House is inspired by nobody.
  - Toy Story is incapable of being a film.

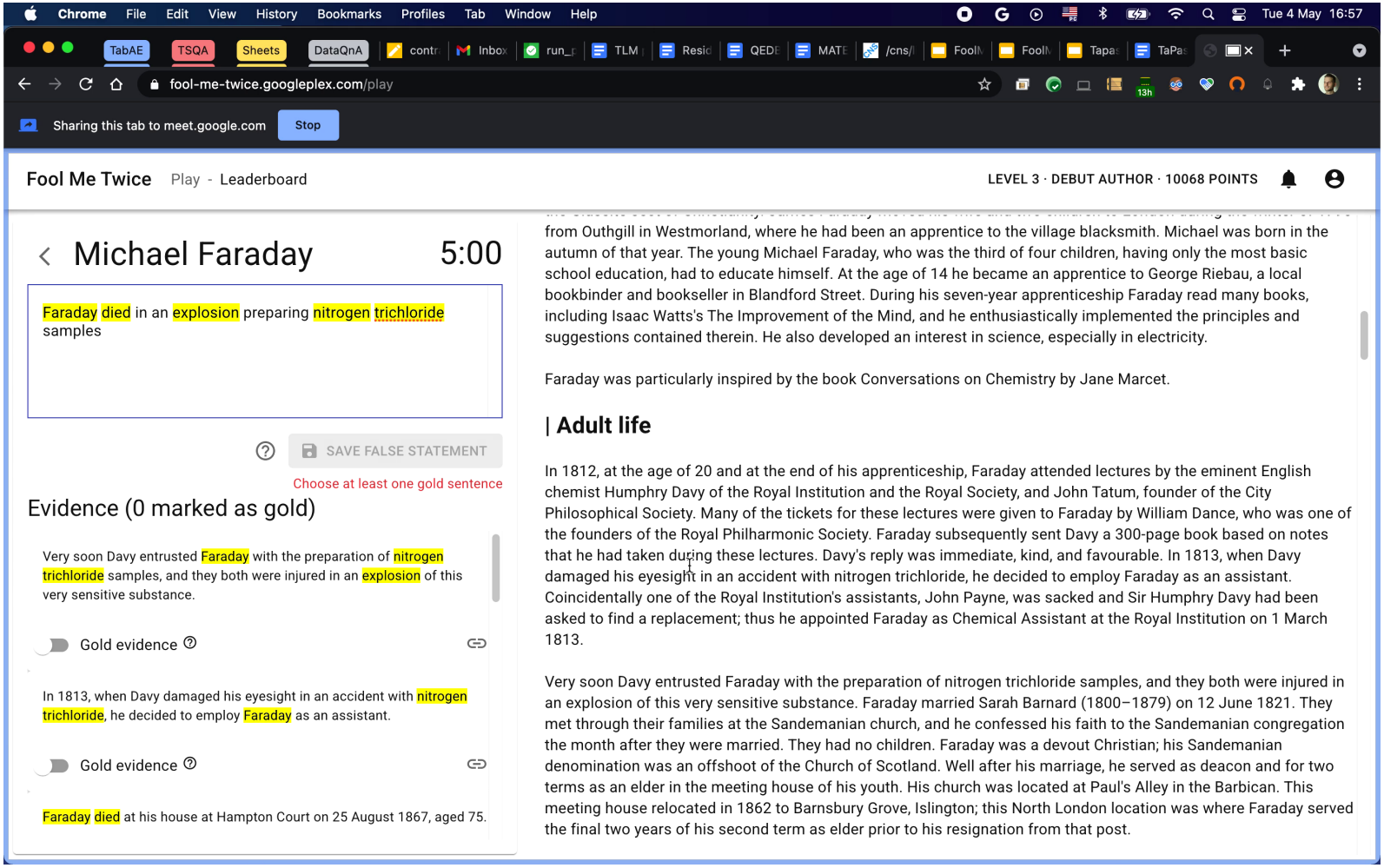
You don't always need the evidence (Poliak, 2018)

Fool Me Twice

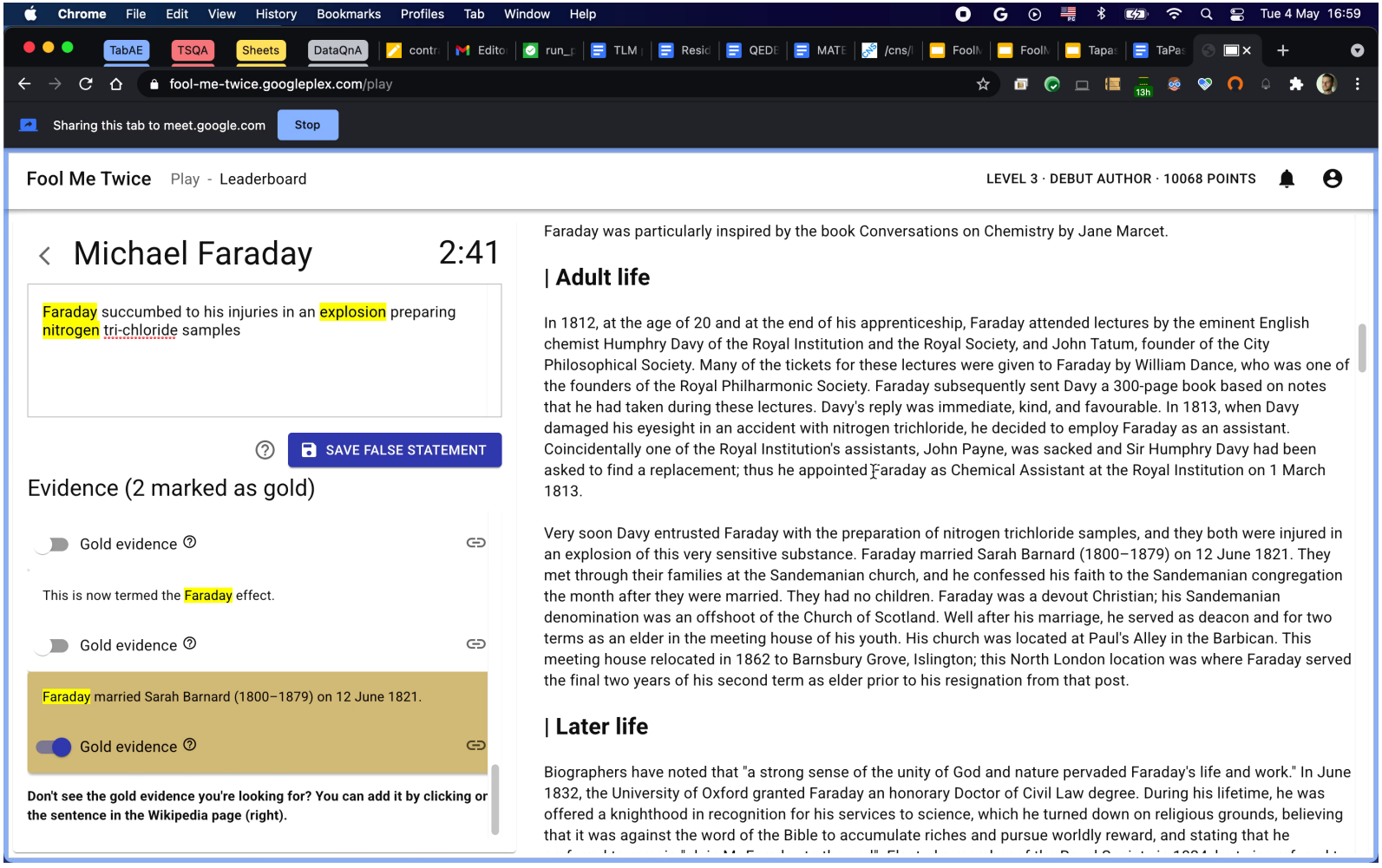
How to gather claims



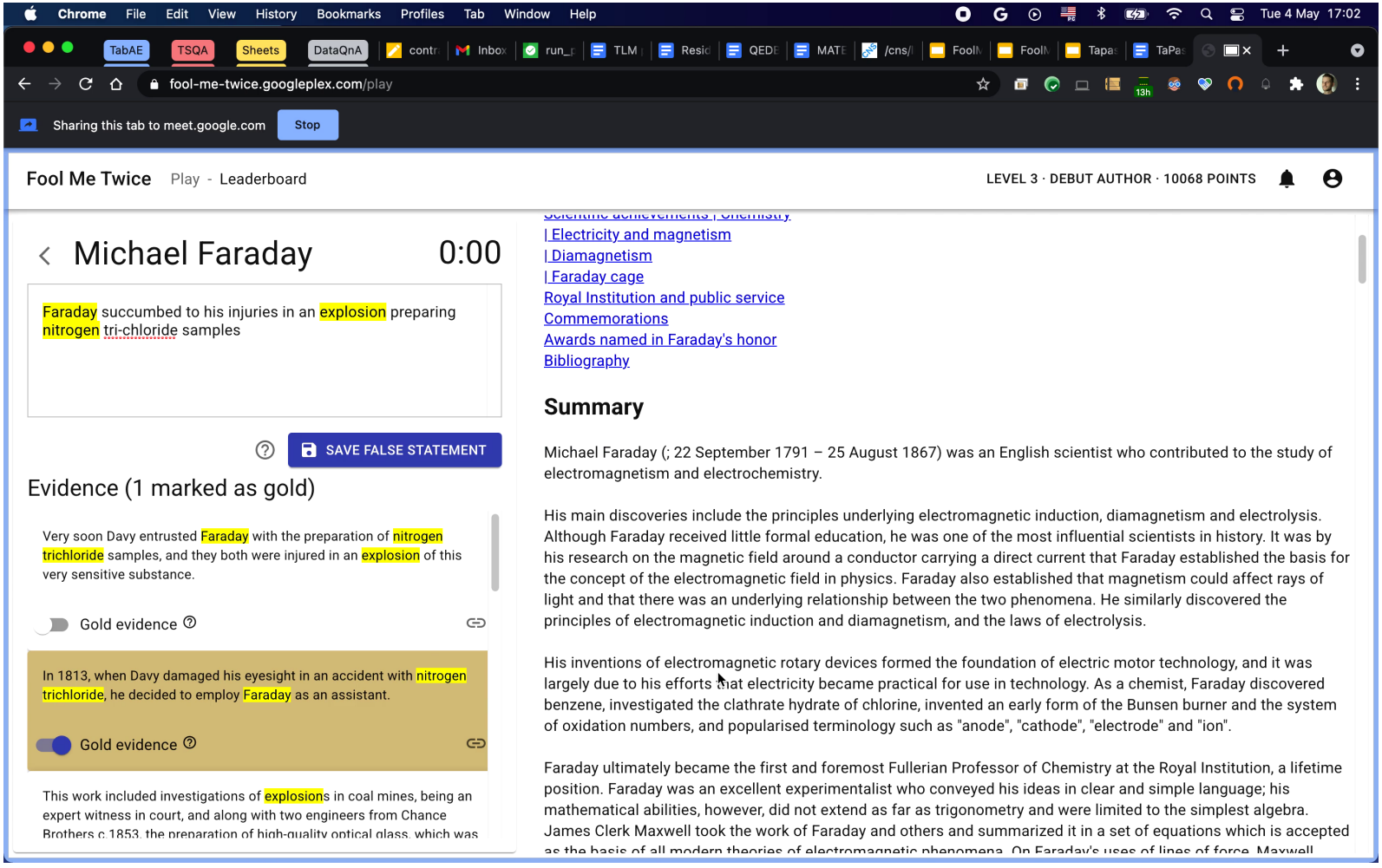
How to gather claims



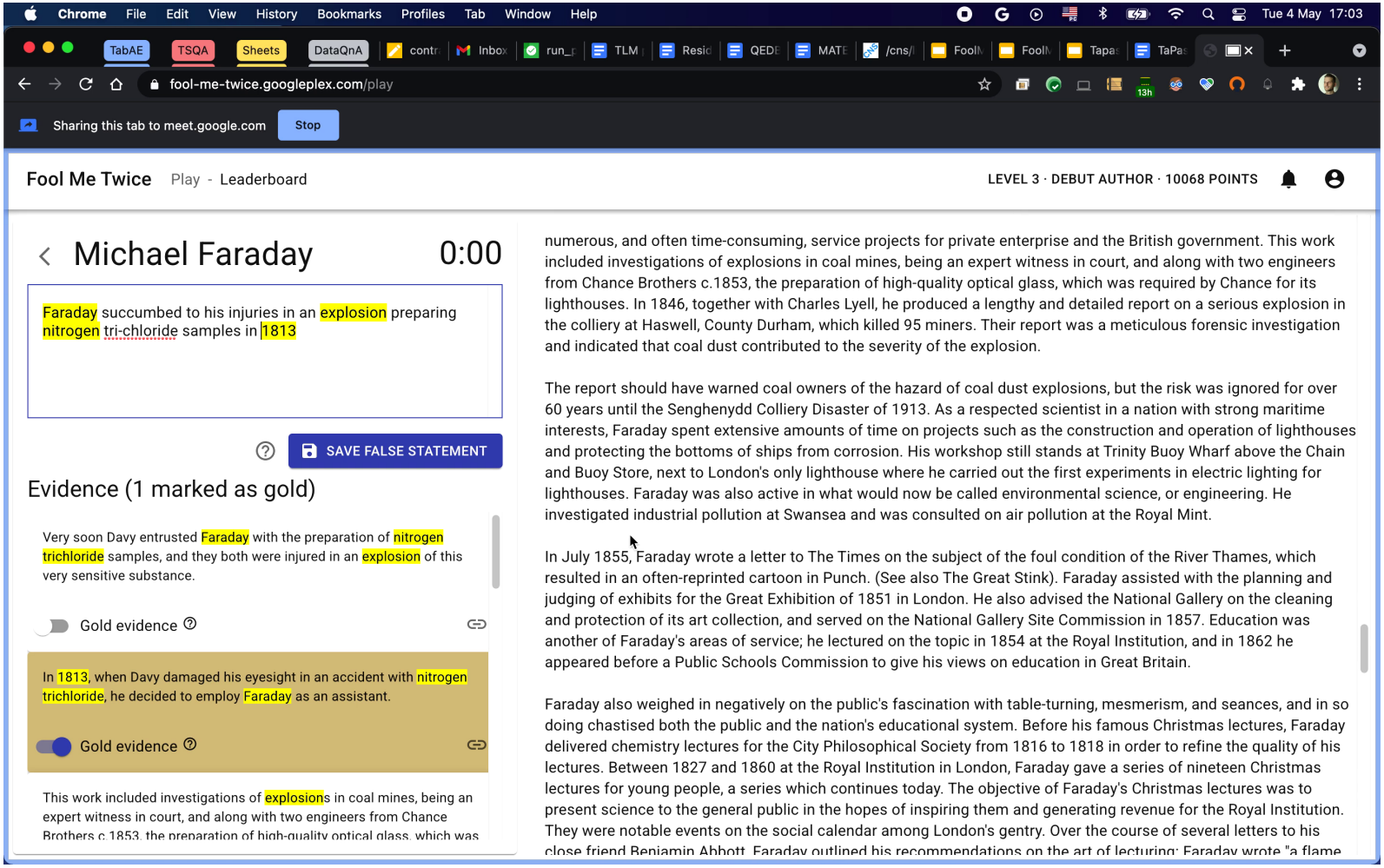
How to gather claims



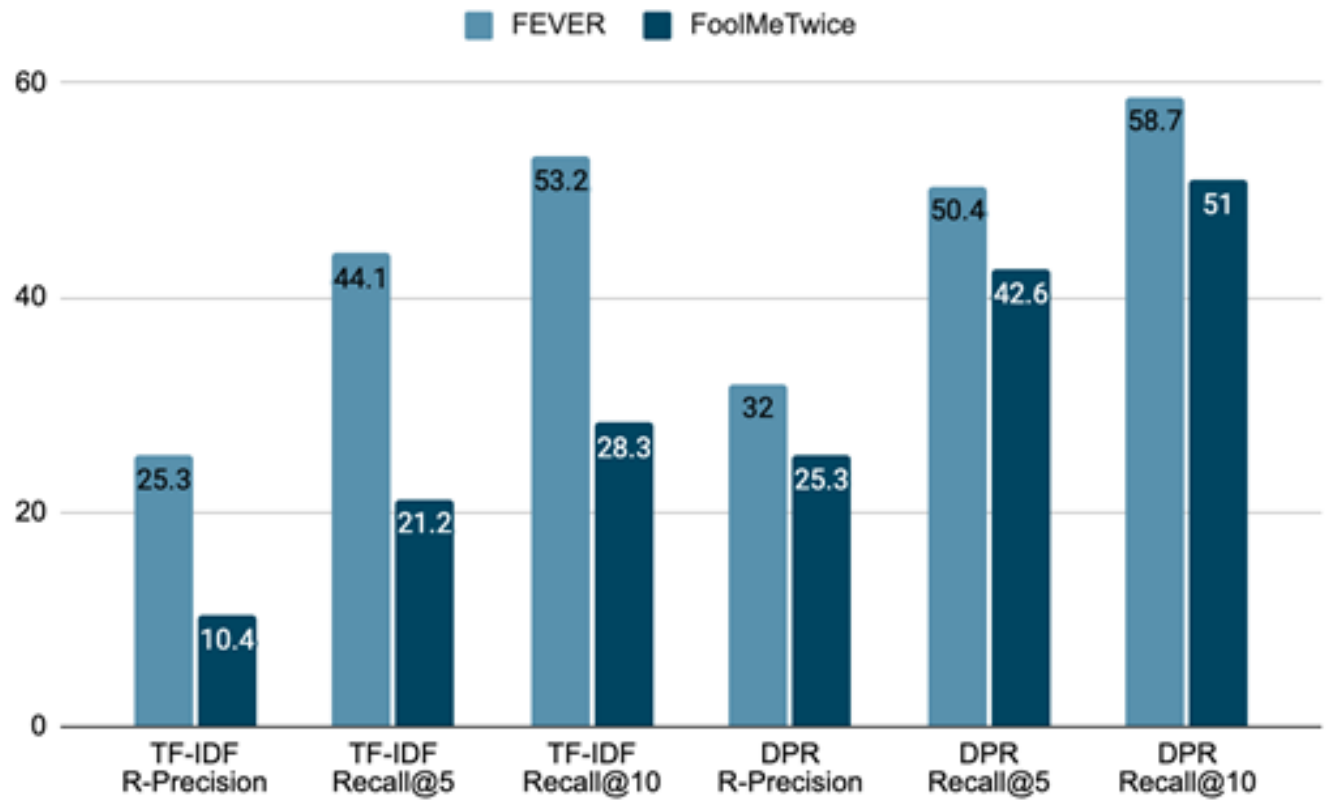
How to gather claims



How to gather claims



It is Harder (a bit)



Helping Humans Do the Task

Claim: Only one spacecraft has visited the planet and Neptune has more than 13 moons.

Helping Humans Do the Task

Claim: Only one spacecraft has visited the planet and Neptune has more than 13 moons.

ChatGPT Explanation:

Yes, the claim is true. According to the evidence from Wikipedia, only one spacecraft, Voyager 2, has visited Neptune. It flew by the planet on August 25, 1989, and remains the only spacecraft to have visited Neptune. Additionally, Neptune has more than 13 moons. It has a total of 14 known moons, with Triton being the largest and most massive moon.

Helping Humans Do the Task

Claim: Only one spacecraft has visited the planet and Neptune has more than 13 moons.

ChatGPT Explanation:

Yes, the claim is true. According to the evidence from Wikipedia, only one spacecraft, Voyager 2, has visited Neptune. It flew by the planet on August 25, 1989, and remains the only spacecraft to have visited Neptune. Additionally, Neptune has more than 13 moons. It has a total of 14 known moons, with Triton being the largest and most massive moon.

Retrieved Passage #1:

The spacecraft performed a near-encounter with the moon Nereid before it came within 4,400 km of Neptune's atmosphere on 25 August, then passed close to the planet's largest moon Triton later the same day. The spacecraft verified the existence of a magnetic field surrounding the planet and discovered that the field was offset from the centre and tilted in a manner similar to the field around Uranus...

Helping Humans Do the Task

Claim: Only one spacecraft has visited the planet and Neptune has more than 13 moons.

ChatGPT Explanation:

Yes, the claim is true. According to the evidence from Wikipedia, only one spacecraft, Voyager 2, has visited Neptune. It flew by the planet on August 25, 1989, and remains the only spacecraft to have visited Neptune. Additionally, Neptune has more than 13 moons. It has a total of 14 known moons, with Triton being the largest and most massive moon.

Retrieved Passage #1:

The spacecraft performed a near-encounter with the moon Nereid before it came within 4,400 km of Neptune's atmosphere on 25 August, then passed close to the planet's largest moon Triton later the same day. The spacecraft verified the existence of a magnetic field surrounding the planet and discovered that the field was offset from the centre and tilted in a manner similar to the field around Uranus...

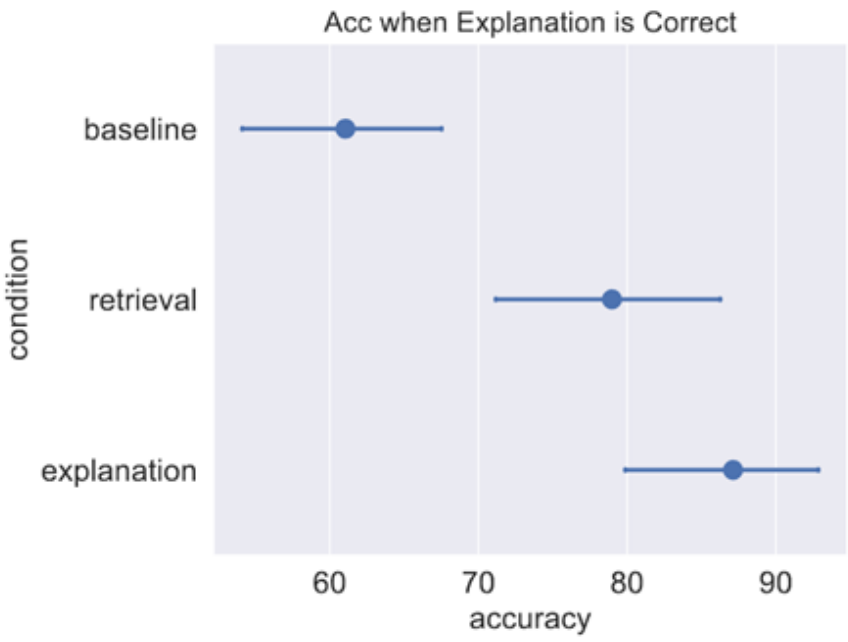
Contrastive Explanation - Supporting:

The claim is true based on the evidence from Wikipedia. The evidence states that only one spacecraft, Voyager 2, has visited Neptune. It flew by the planet on August 25, 1989, and remains the only spacecraft to have visited Neptune. Additionally, the evidence mentions that Neptune has 14 known moons, including Triton, which is the largest Neptunian moon. Therefore, the claim that only one spacecraft has visited the planet and Neptune has more than 13 moons is supported by the evidence.

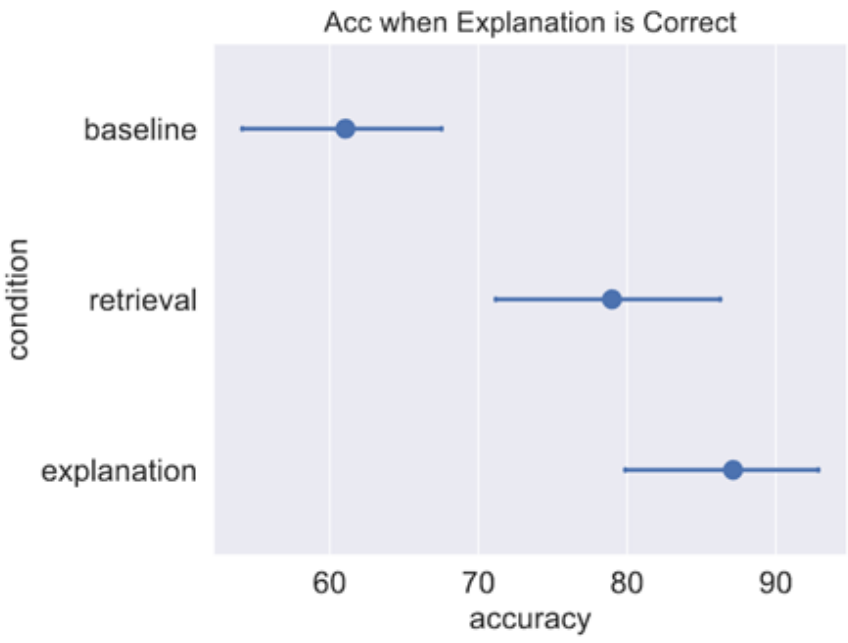
Contrastive Explanation - Refuting:

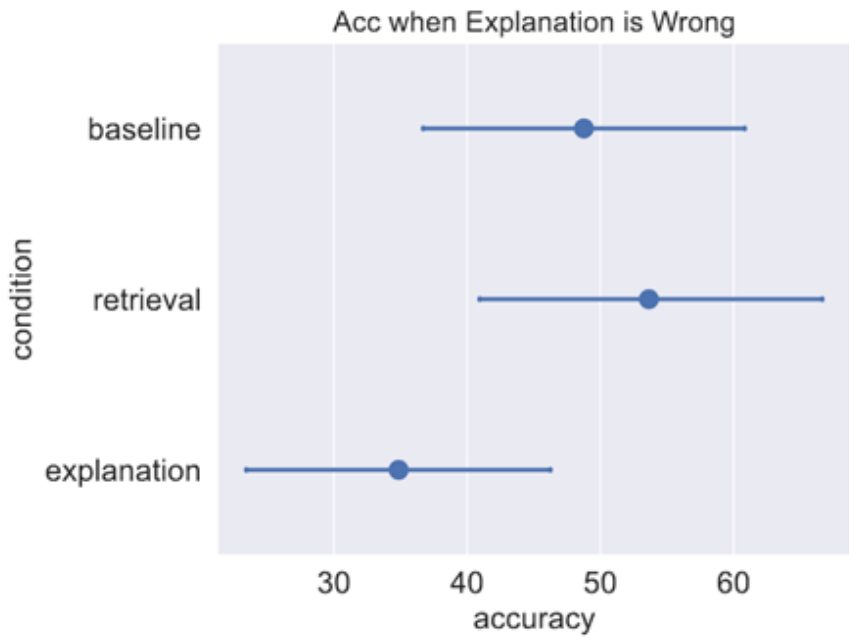
The claim is false. According to the evidence from Wikipedia, the spacecraft Voyager 2 visited Neptune on August 25, 1989. The evidence states that there have been discussions and proposals for future missions to Neptune, including an orbital mission and a flyby spacecraft. Therefore, it is clear that more than one spacecraft has visited Neptune. Furthermore, the evidence also mentions that Neptune has 14 known moons, contradicting the claim that it has more than 13 moons.

Results


Acc when Explanation is Correct
baseline
retrieval
explanation
condition
60
70
80
90
accuracy

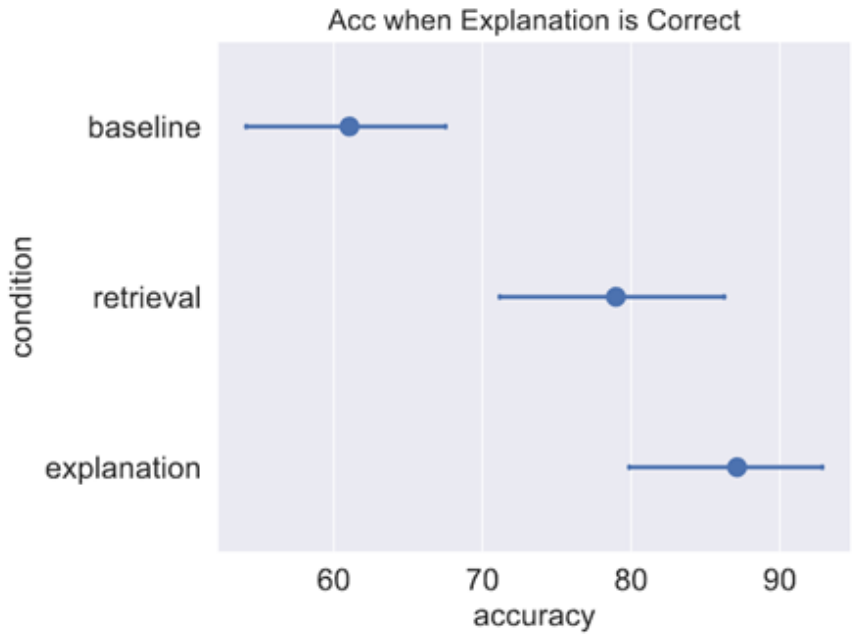
Results


Acc when Explanation is Correct
baseline
retrieval
explanation
condition
60
70
80
90
accuracy


Acc when Explanation is Wrong
baseline
retrieval
explanation
condition
30
40
50
60
accuracy

Why not just show IR all the time?

Why not just show IR all the time?



Breaking Things Down

- What's the solution?
- If the models were never wrong, we wouldn't have this problem
- If the models could know when they were wrong, we wouldn't have this problem
- If we knew when the models would be wrong, we wouldn't have this problem

CAIMIRA

Model that can predict when a model is likely to get a question right

Breaking Things Down

- What's the solution?
- If the models were never wrong, we wouldn't have this problem
- If the models could know when they were wrong, we wouldn't have this problem
- If we knew when the models would be wrong, we wouldn't have this problem

CAIMIRA

Model that can predict when a model is likely to get a question right

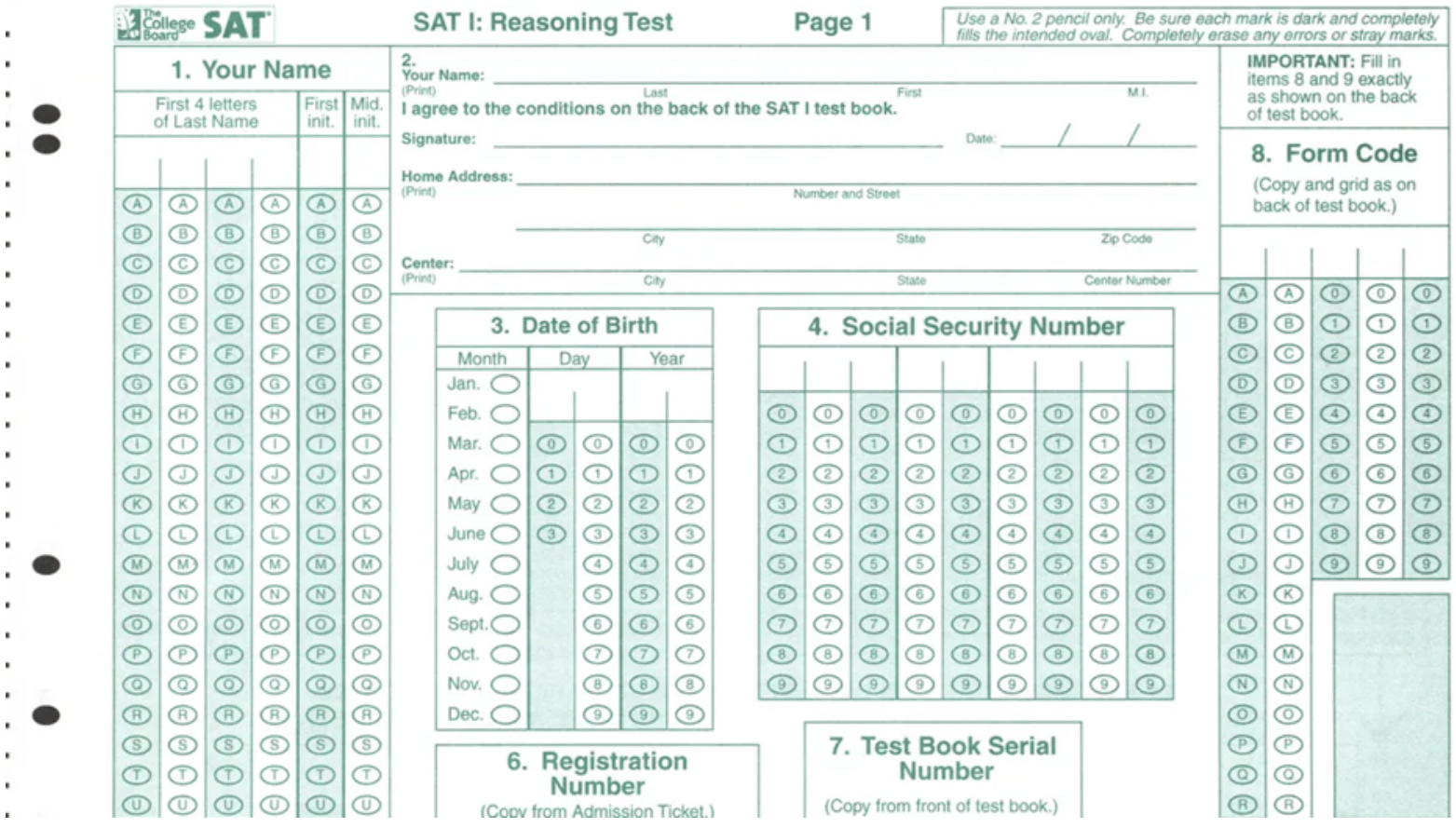
CAIMIRA

Do great minds think alike? Investigating Human-AI Complementarity in Question Answering with CAIMIRA

Maharshi Gor[1] **Hal Daumé III**[1,2] **Tianyi Zhou**[1] **Jordan Boyd-Graber**[1]

[1]University of Maryland [2]Microsoft Research
mgor@cs.umd.edu

Item Response Theory



The College Board SAT

SAT I: Reasoning Test Page 1

Use a No. 2 pencil only. Be sure each mark is dark and completely fills the intended oval. Completely erase any errors or stray marks.

1. Your Name

First 4 letters of Last Name | First init. | Mid. init.

2.
Your Name: (Print) Last First M.I.

I agree to the conditions on the back of the SAT I test book.

Signature: Date: / /

Home Address: (Print) Number and Street

City State Zip Code

Center: (Print) City State Center Number

IMPORTANT: Fill in items 8 and 9 exactly as shown on the back of test book.

8. Form Code

(Copy and grid as on back of test book.)

3. Date of Birth

Month: Jan. Feb. Mar. Apr. May June July Aug. Sept. Oct. Nov. Dec.

Day: 0 1 2 3 4 5 6 7 8 9

Year: 0 1 2 3 4 5 6 7 8 9

4. Social Security Number

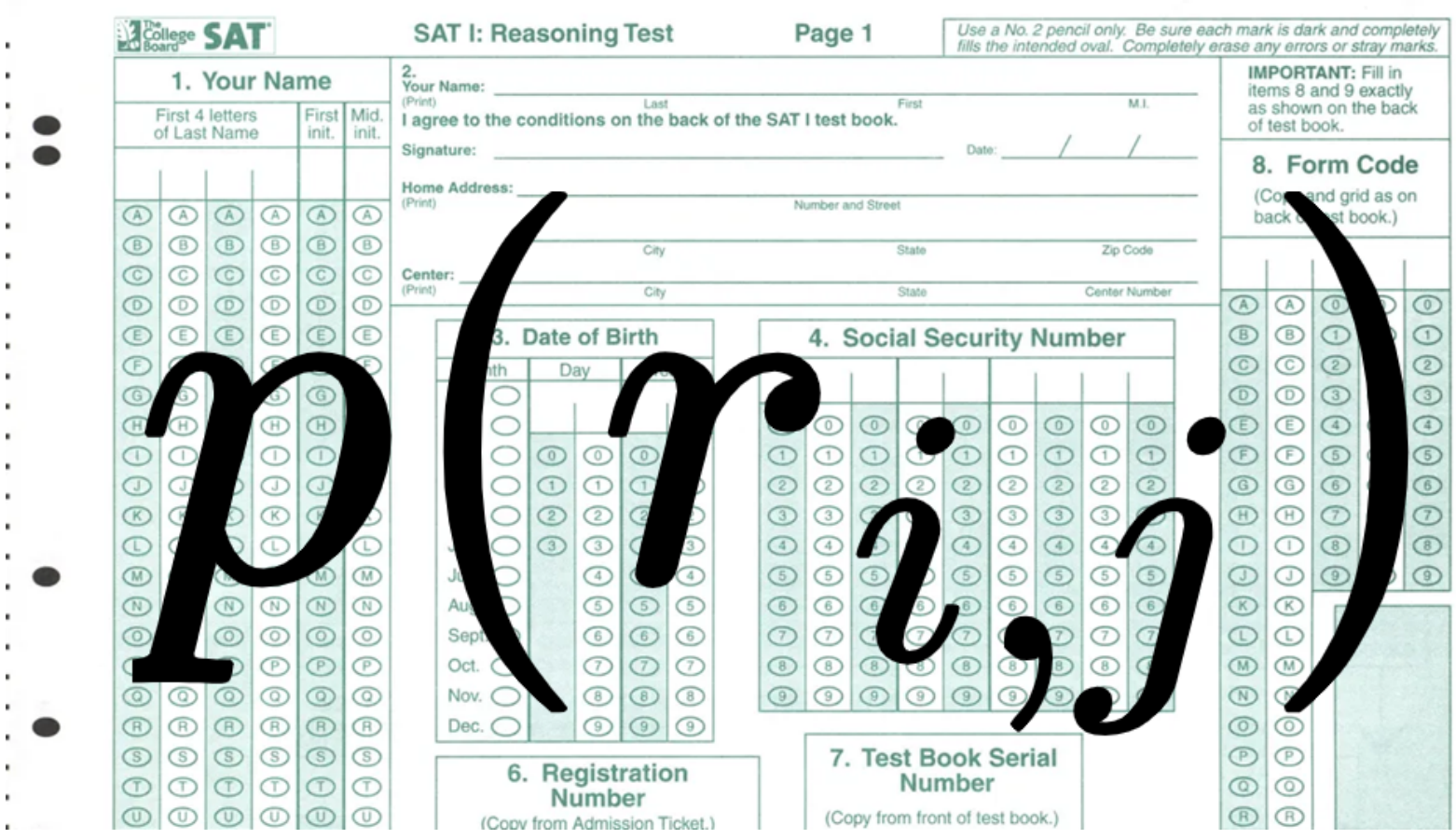
6. Registration Number

(Copy from Admission Ticket.)

7. Test Book Serial Number

(Copy from front of test book.)

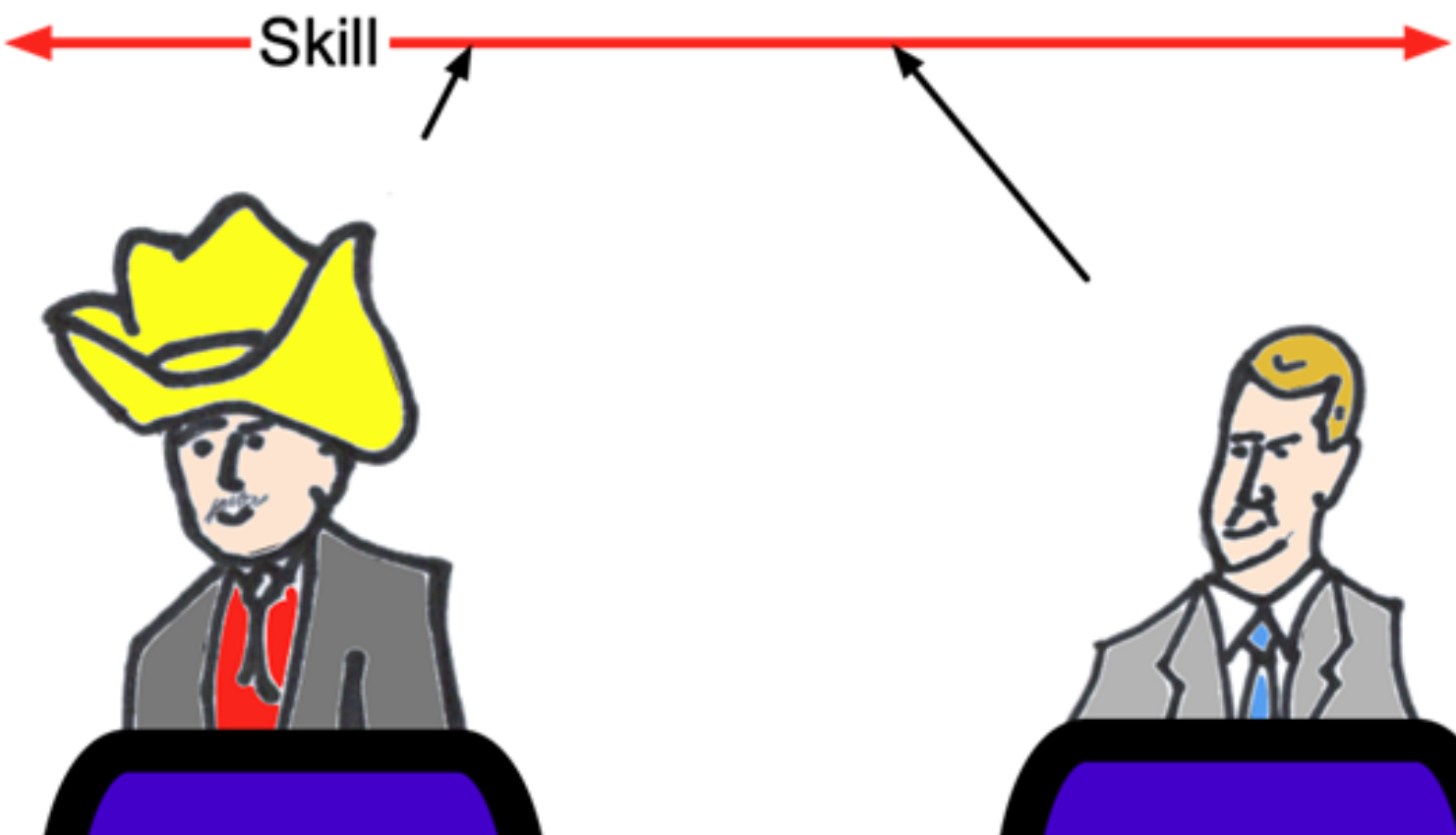
Item Response Theory



Item Response Theory

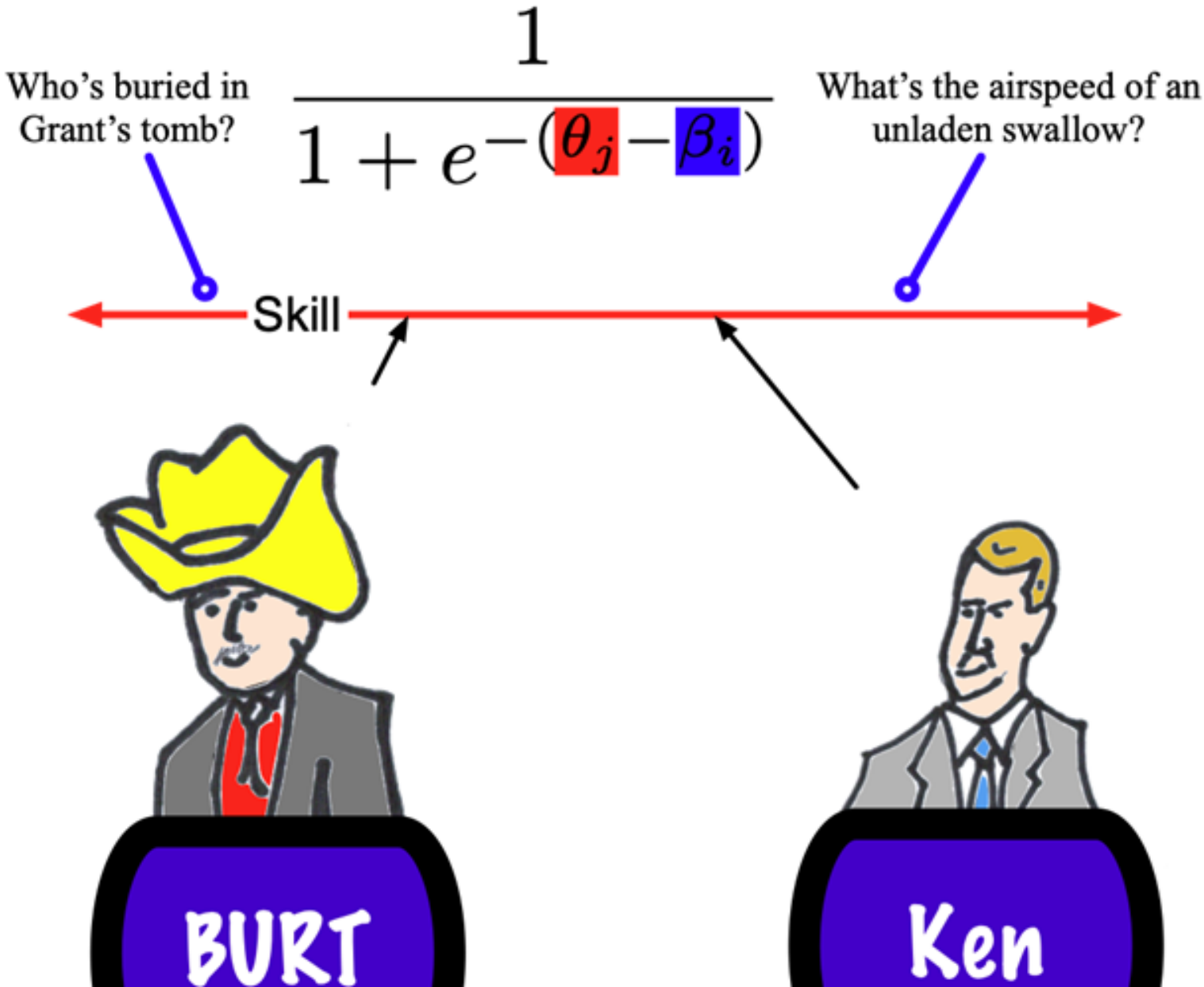
Item Response Theory

$$\frac{1}{1 + e^{-(\theta_j - \beta_i)}}$$



Item Response Theory

$$\frac{1}{1 + e^{-(\theta_j - \beta_i)}}$$



Aside: Family Tree

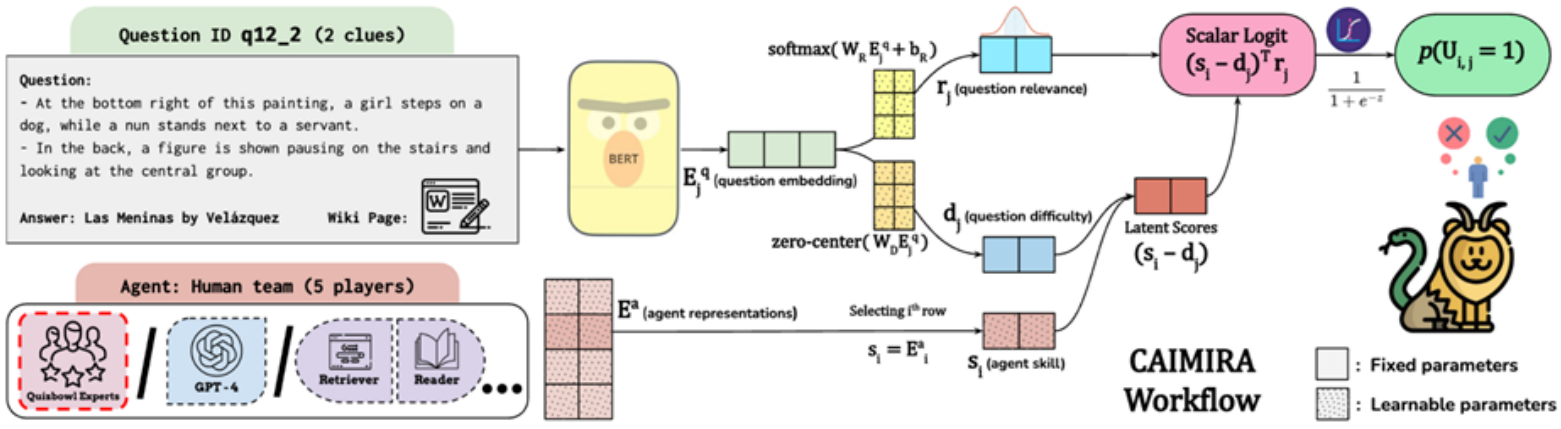
Zermelo

Bradley-Terry Models

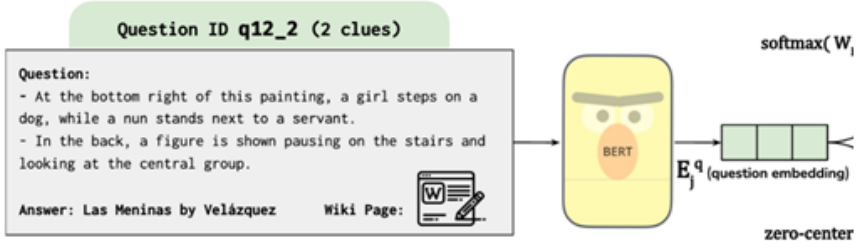
Ideal Point Models

ELO Models

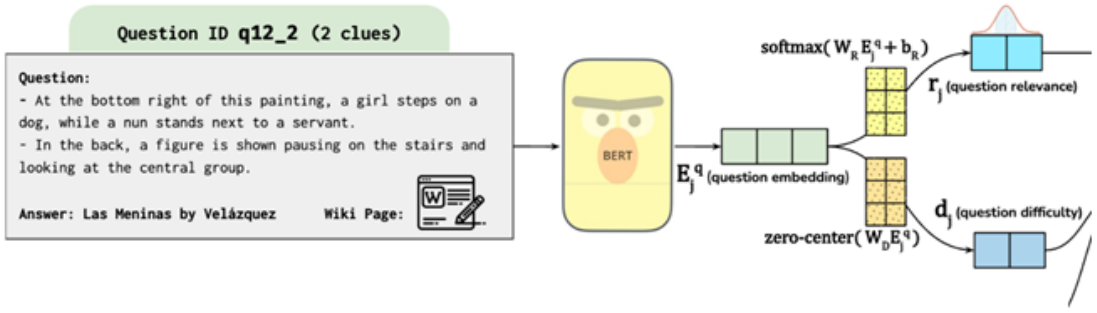
What more do we need?



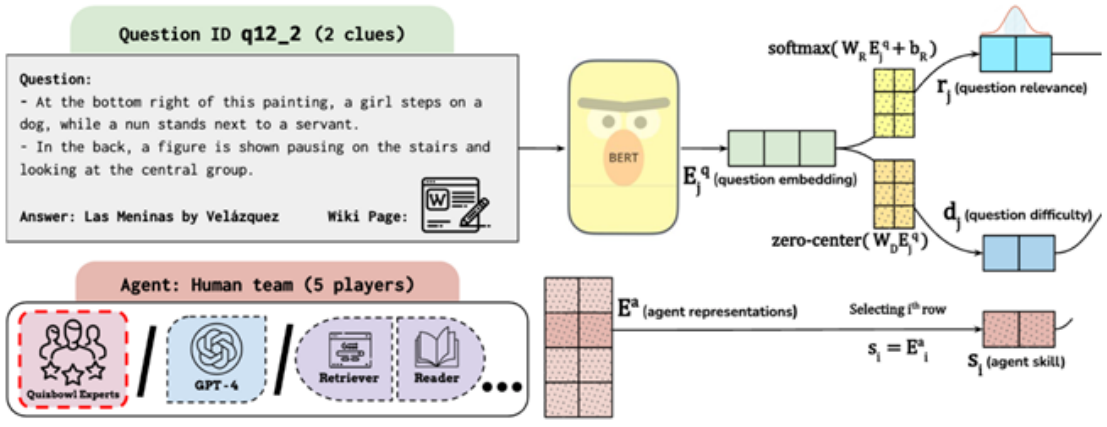
What more do we need?



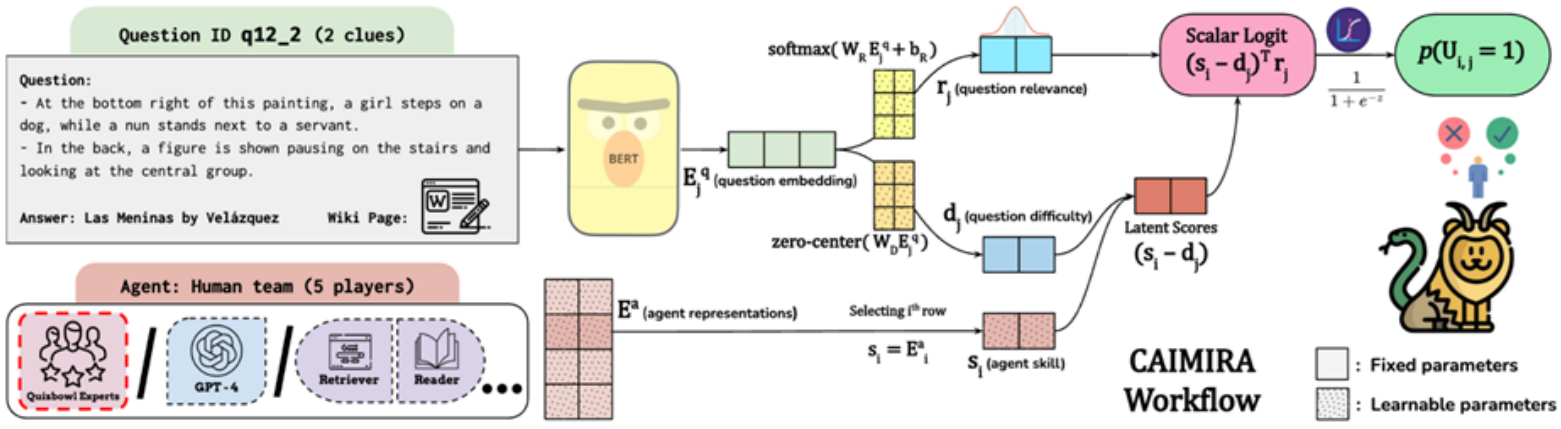
What more do we need?



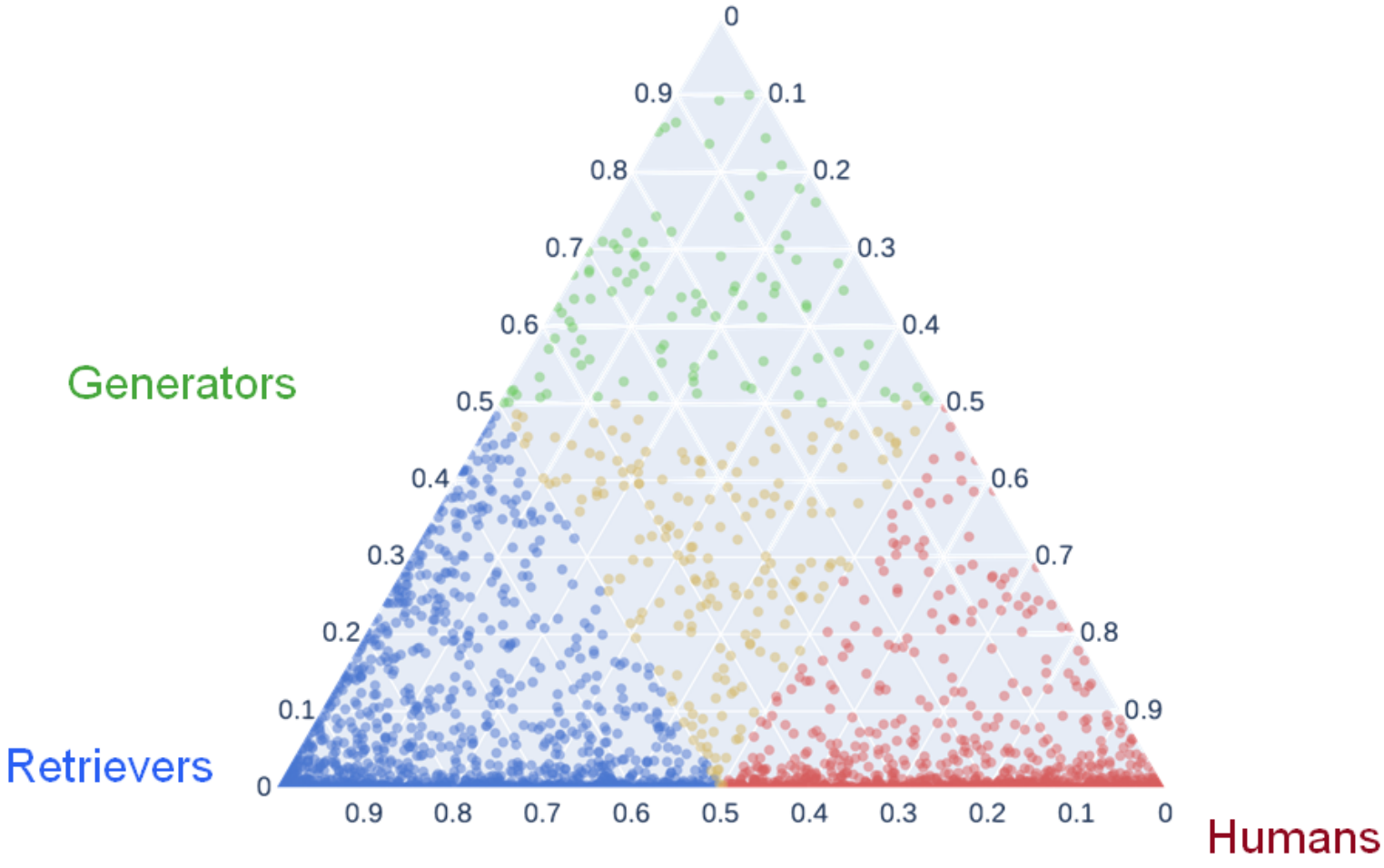
What more do we need?



What more do we need?



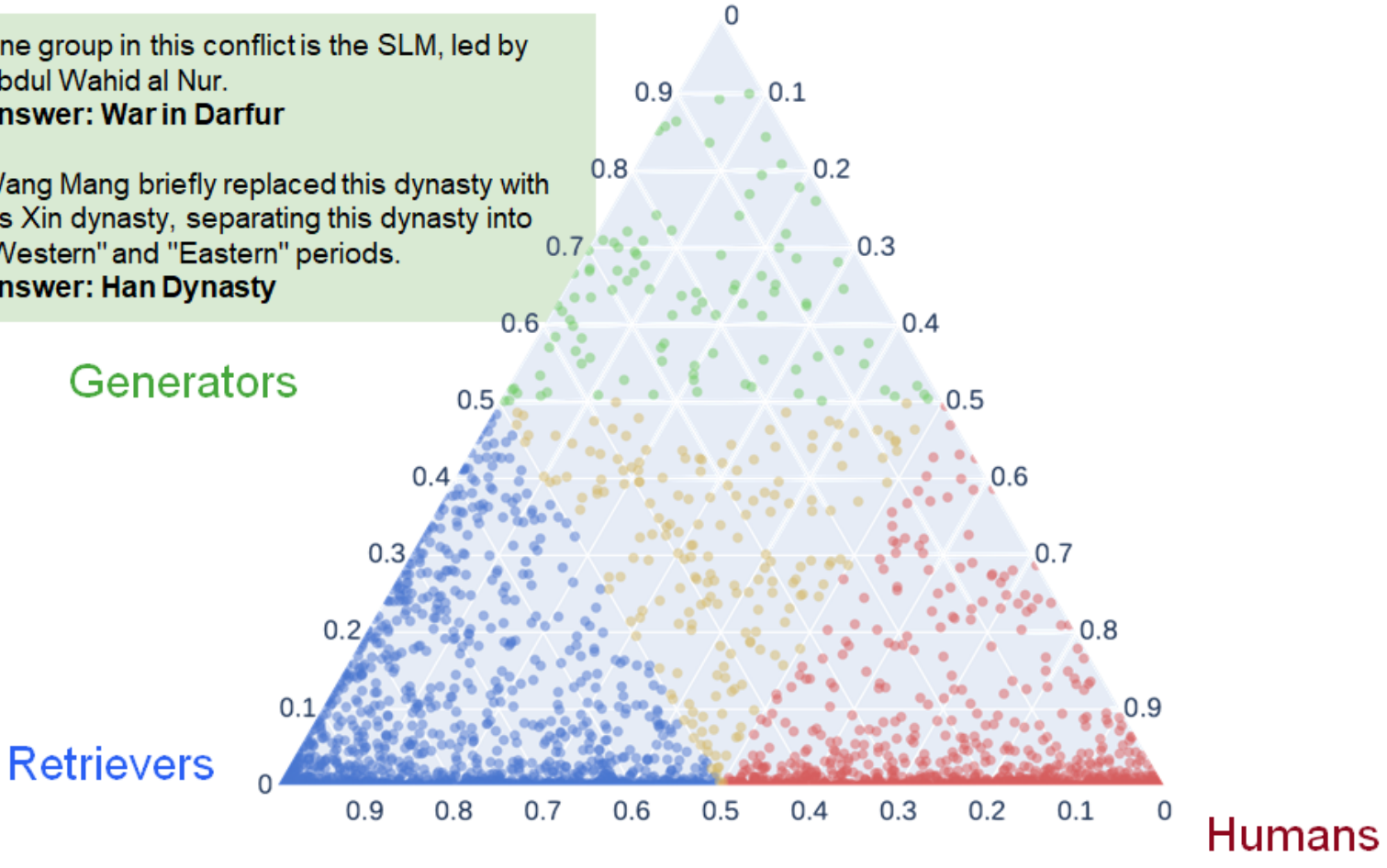
What's hard for Computers



What's hard for Computers

One group in this conflict is the SLM, led by Abdul Wahid al Nur.
Answer: War in Darfur

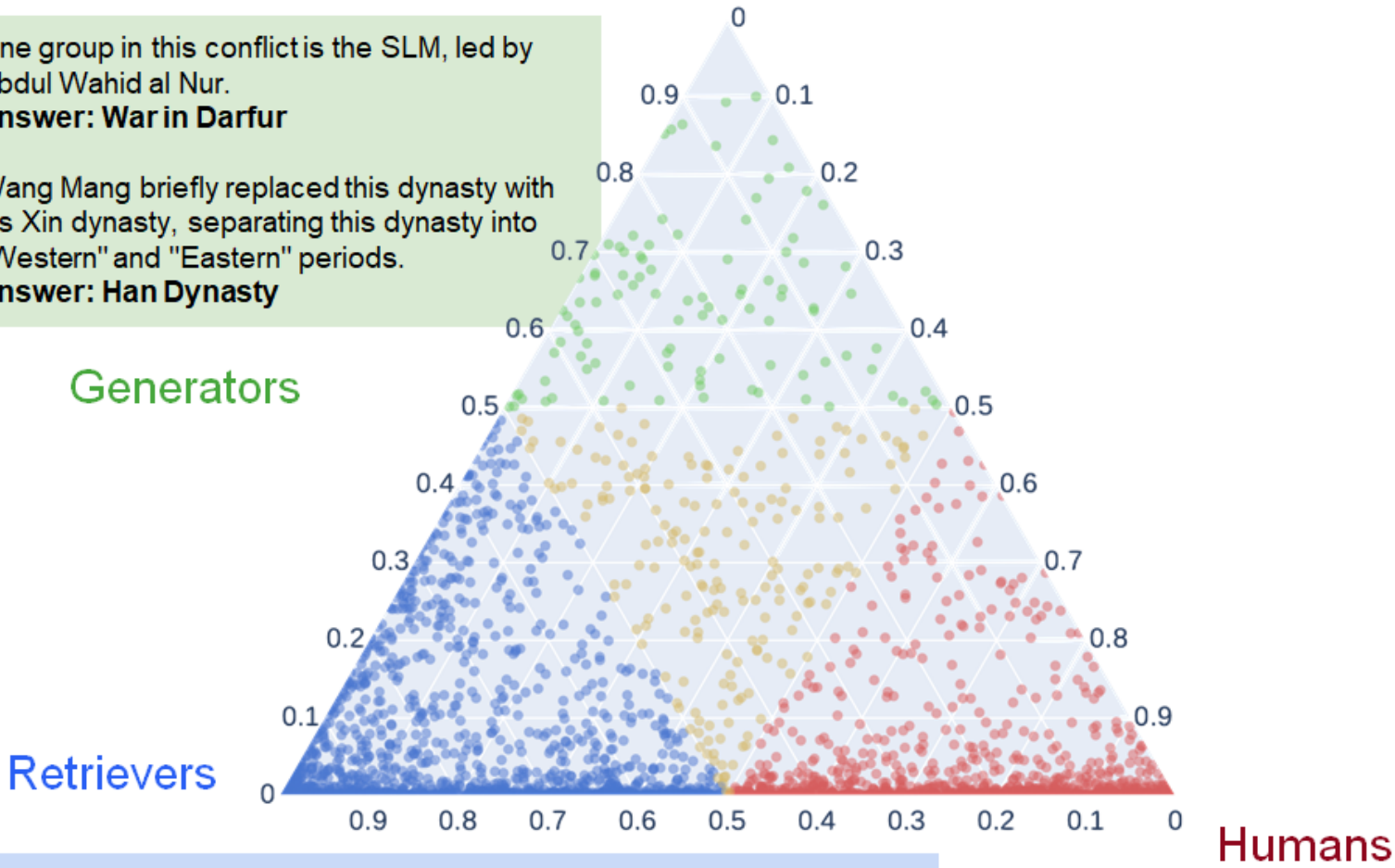
Wang Mang briefly replaced this dynasty with his Xin dynasty, separating this dynasty into "Western" and "Eastern" periods.
Answer: Han Dynasty



What's hard for Computers

One group in this conflict is the SLM, led by Abdul Wahid al Nur.
Answer: War in Darfur

Wang Mang briefly replaced this dynasty with his Xin dynasty, separating this dynasty into "Western" and "Eastern" periods.
Answer: Han Dynasty



This nation's government, controlled by the Islamic Brotherhood and headed by Mohamed Morsi, came to power after the overthrow of Hosni Mubarak.
Answer: Egypt
When this athlete was the 2004 NBA Rookie of the Year, he became the youngest person ever to win that award.
Answer: LeBron Raymone James

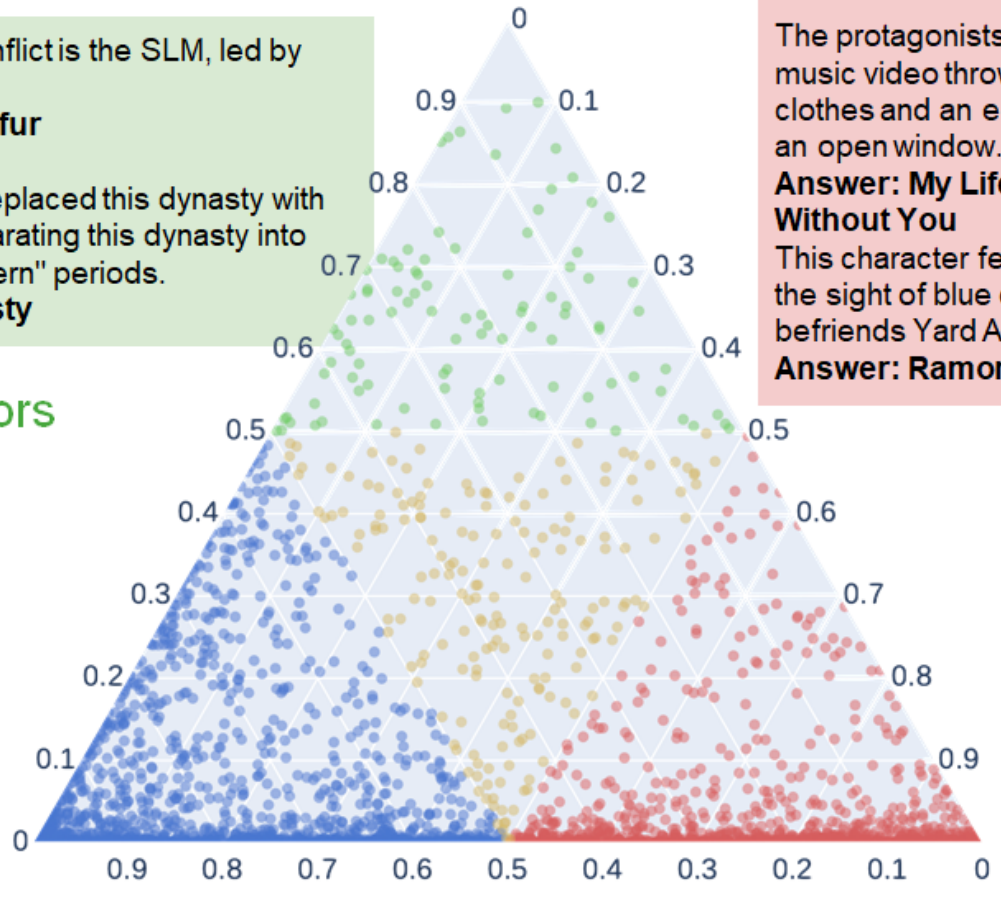
What's hard for Computers

One group in this conflict is the SLM, led by Abdul Wahid al Nur.
Answer: War in Darfur

Wang Mang briefly replaced this dynasty with his Xin dynasty, separating this dynasty into "Western" and "Eastern" periods.
Answer: Han Dynasty

The protagonists of this song's music video throw magazines, clothes and an empty fishbowl out an open window.
Answer: My Life Would Suck Without You

This character feels nauseous at the sight of blue oatmeal and befriends Yard Ape.
Answer: Ramona Quimby



This nation's government, controlled by the Islamic Brotherhood and headed by Mohamed Morsi, came to power after the overthrow of Hosni Mubarak.
Answer: Egypt

When this athlete was the 2004 NBA Rookie of the Year, he became the youngest person ever to win that award.
Answer: LeBron Raymone James

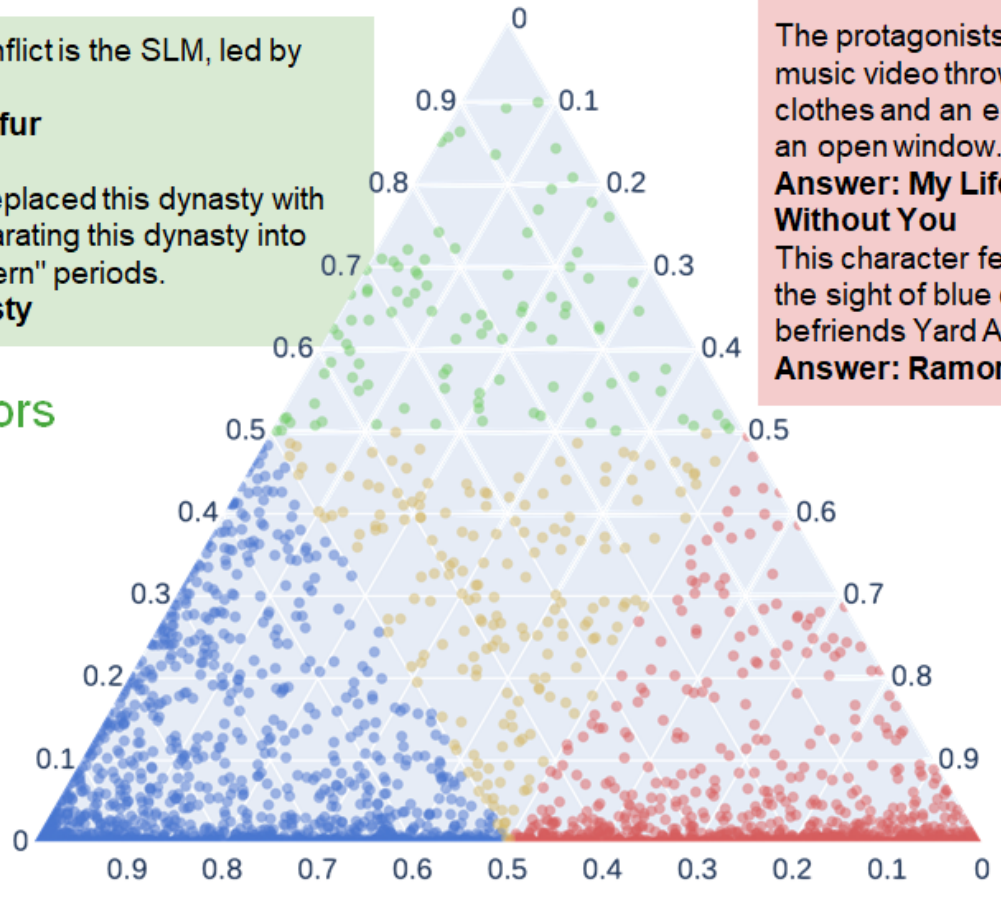
What's hard for Computers

One group in this conflict is the SLM, led by Abdul Wahid al Nur.
Answer: War in Darfur

Wang Mang briefly replaced this dynasty with his Xin dynasty, separating this dynasty into "Western" and "Eastern" periods.
Answer: Han Dynasty

The protagonists of this song's music video throw magazines, clothes and an empty fishbowl out an open window.
Answer: My Life Would Suck Without You

This character feels nauseous at the sight of blue oatmeal and befriends Yard Ape.
Answer: Ramona Quimby



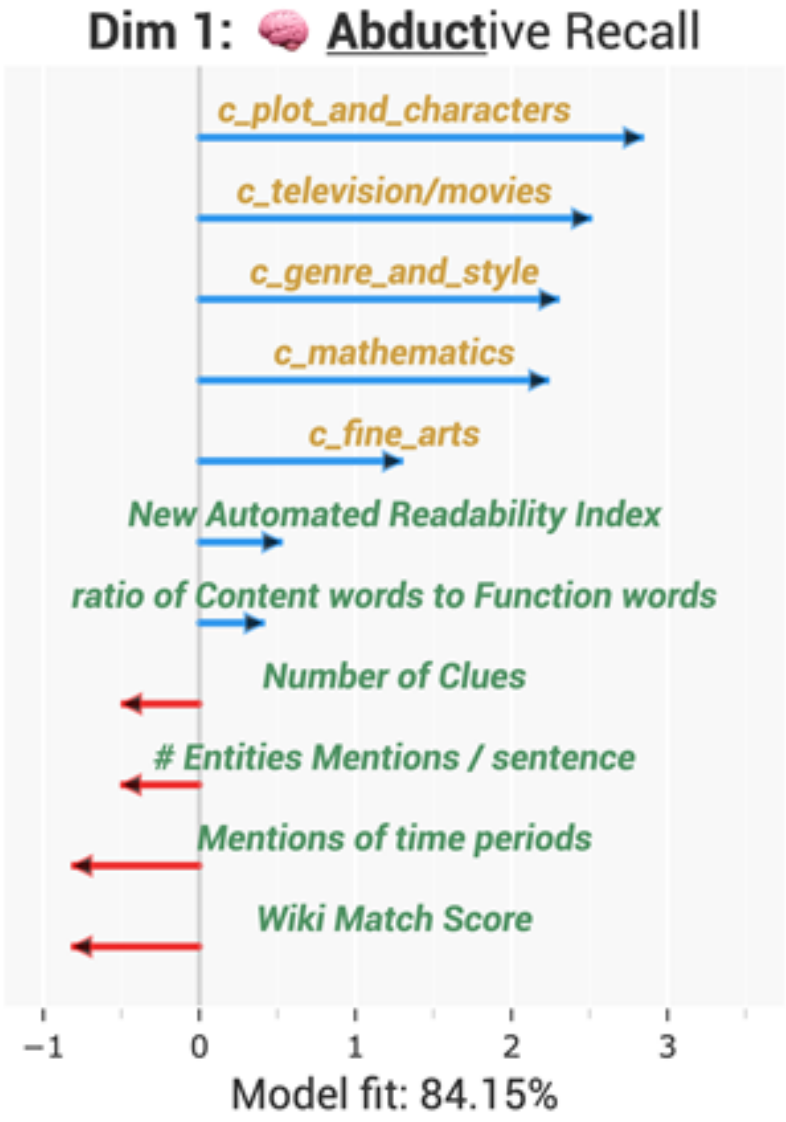
This nation's government, controlled by the Islamic Brotherhood and headed by Mohamed Morsi, came to power after the overthrow of Hosni Mubarak.
Answer: Egypt

When this athlete was the 2004 NBA Rookie of the Year, he became the youngest person ever to win that award.
Answer: LeBron Raymone James

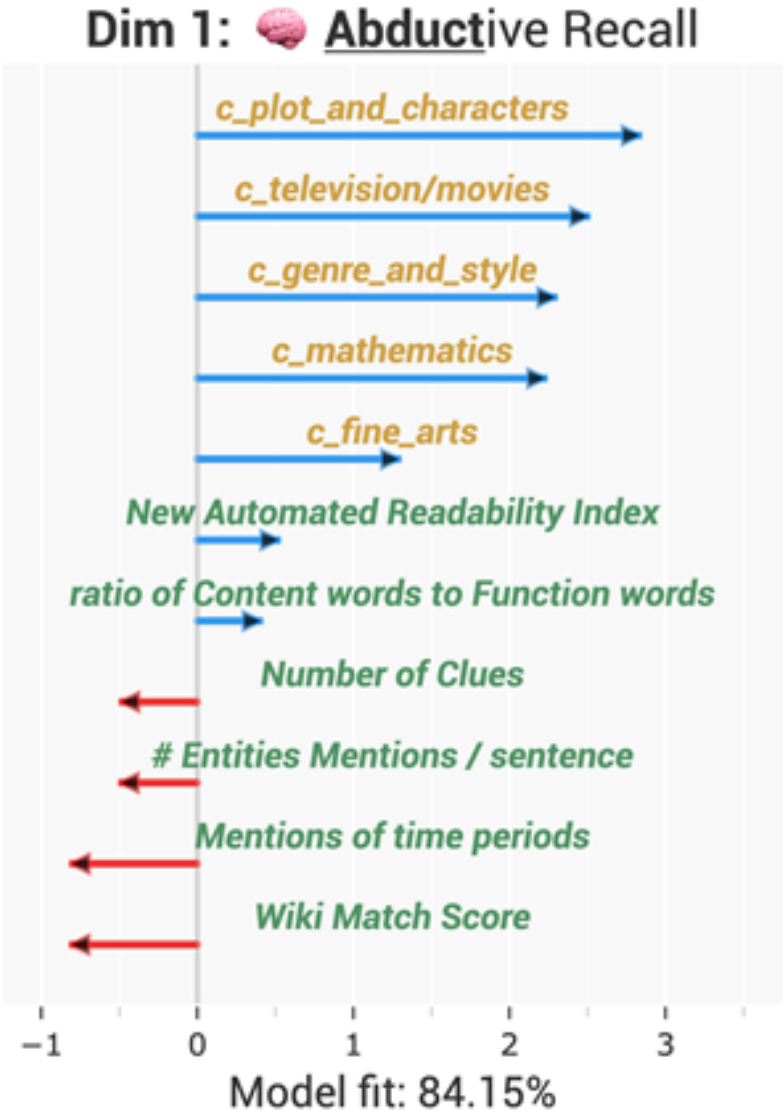
Making Dimensions Interpretable

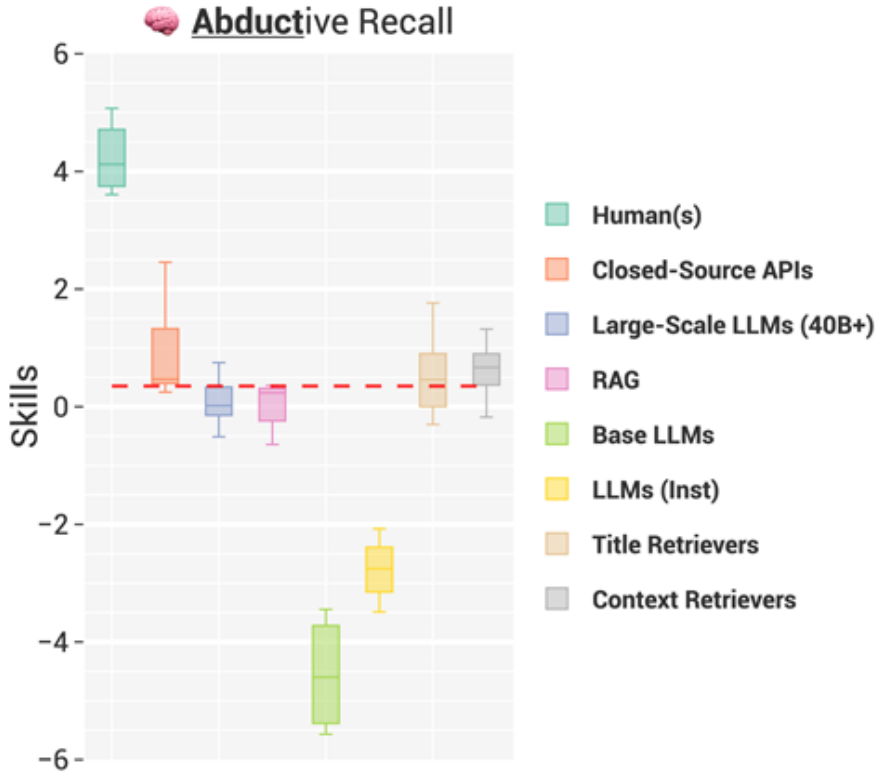
- We encode features to predict when a latent dimension will be active $r_{i,j}$
- Only allow model to use a few dimensions
- Objective function encourages model to find dimensions that discriminate agents
- Posthoc analysis allows us to label the dimensions

Hard for Computers: Abductive Inference

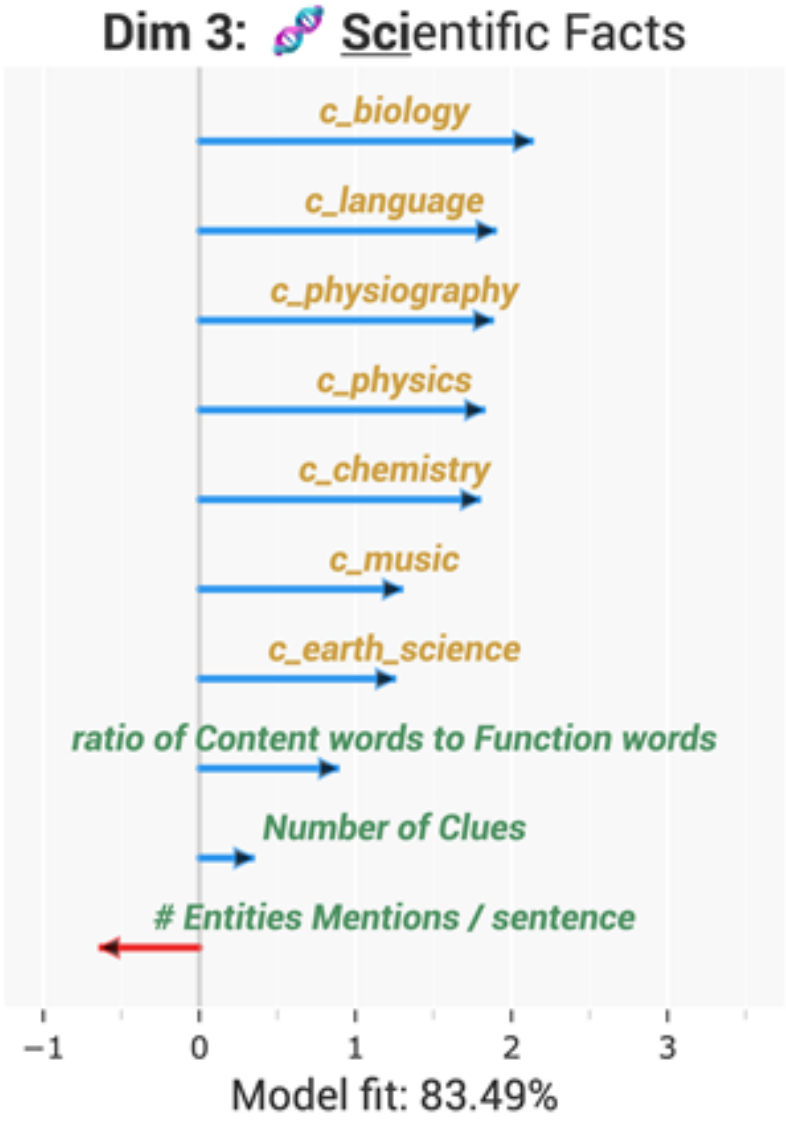


Hard for Computers: Abductive Inference

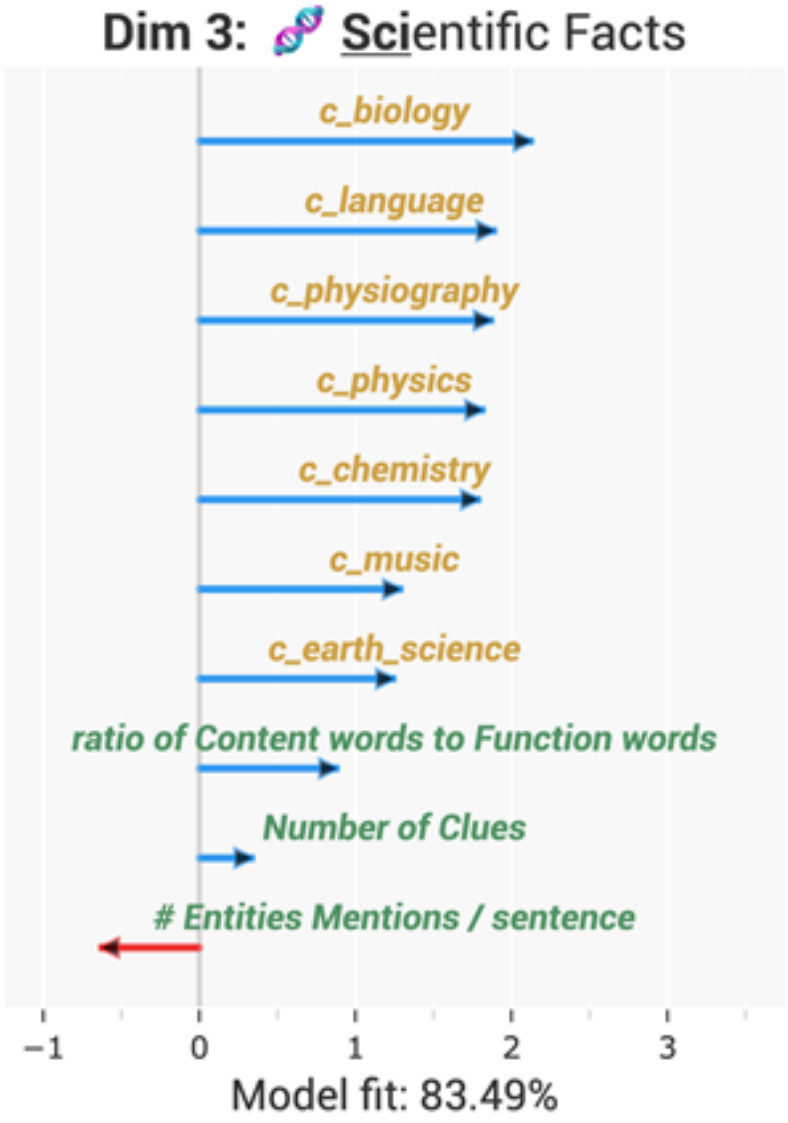


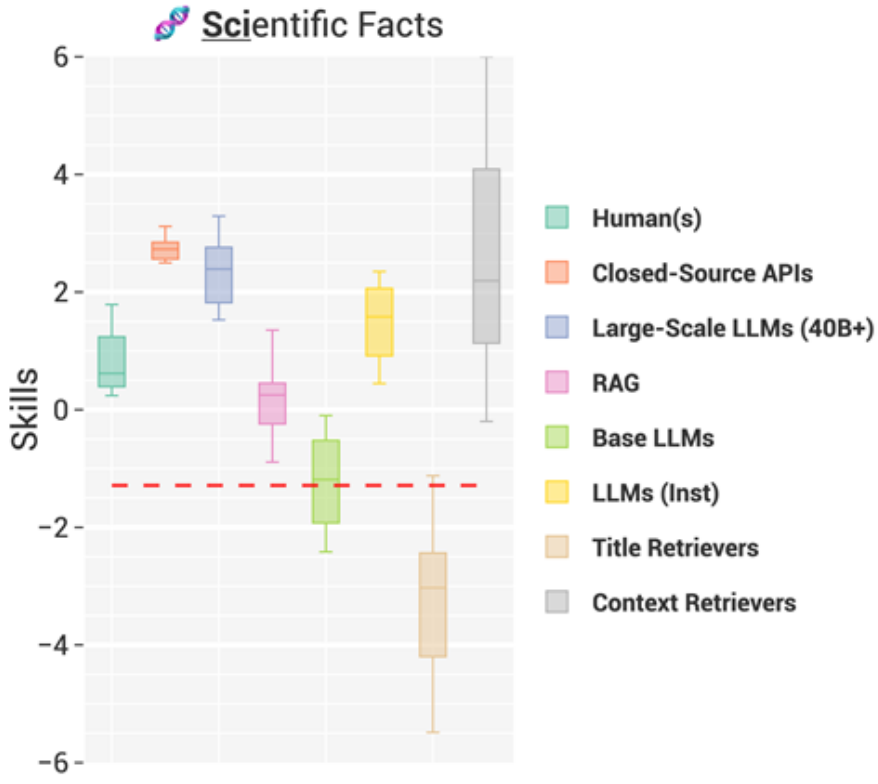


Hard for Humans: Science



Hard for Humans: Science





Adversarial Datasets

- Many benchmarks are "saturated"
- Newer datasets claim to be "adversarial"
  - Hard for computers, "easy" for humans
  - No real metric / definition
- Can we use the lessons of the previous paper to inform how to write hard examples
- Can we *measure* how well we did?

Adversarial Datasets

- Many benchmarks are "saturated"
- Newer datasets claim to be "adversarial"
  - Hard for computers, "easy" for humans
  - No real metric / definition
- Can we use the lessons of the previous paper to inform how to write hard examples
- Can we *measure* how well we did?
- Language game: increasing the difficulty level
- But need to measure!

Expanding IRT: Discriminability

How much of a skill gap is needed?

$$p\left(r_{i,j}\right) = \frac{1}{\exp\left\{\gamma_j(\beta_i - \theta_j)\right\}} \tag{1}$$

Expanding IRT: Discriminability

How much of a skill gap is needed?

$$p\left(r_{i,j}\right) = \frac{1}{\exp\left\{\gamma_j(\beta_i - \theta_j)\right\}} \tag{1}$$

Adversarial Score

- Use IRT again
- Model probability of skilled human getting it right, compare to models (want it to be high!)

$$\mu_j = \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{H_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled human rep. prob.}} - \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{M_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled model rep. prob.}}, \qquad (2)$$

Adversarial Score

- Use IRT again
- Model probability of skilled human getting it right, compare to models (want it to be high!)

$$\mu_j = \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{H_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled human rep. prob.}} - \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{M_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled model rep. prob.}}, \qquad (2)$$

Why not use raw accuracy?

- Want patterns, not luck
- IRT can find (and downweight) bad questions

Adversarial Score

- Use IRT again
- Model probability of skilled human getting it right, compare to models (want it to be high!)

$$\mu_j = \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{H_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled human rep. prob.}} - \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{M_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled model rep. prob.}}, \tag{2}$$

- Skilled humans should agree on the answer

$$\delta_j = \underset{i \sim H_{(1)}}{\text{MD}}\left[\sigma_{\text{2pl}}\left(\beta_i^{H_{(1)}}, \theta_j, \gamma_j\right)\right]. \tag{3}$$

Adversarial Score

- Use IRT again
- Model probability of skilled human getting it right, compare to models (want it to be high!)

$$\mu_j = \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{H_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled human rep. prob.}} - \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{M_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled model rep. prob.}}, \tag{2}$$

- Skilled humans should agree on the answer

$$\delta_j = \underset{i \sim H_{(1)}}{\text{MD}}\left[\sigma_{\text{2pl}}\left(\beta_i^{H_{(1)}}, \theta_j, \gamma_j\right)\right]. \tag{3}$$

- Question should be informative (Fischer information to agent skill):

$$\text{IIF}_j(\theta) = \gamma_j^2 \cdot p_j(\theta) \cdot (1 - p_j(\theta)), \text{ where} \tag{4}$$

$$p_j(\theta) = \sigma_{\text{2pl}}(\theta, \theta_j, \gamma_j). \tag{5}$$

Adversarial Score

- Use IRT again
- Model probability of skilled human getting it right, compare to models (want it to be high!)

$$\mu_j = \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{H_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled human rep. prob.}} - \underbrace{\sigma_{\text{2pl}}\left(\beta_*^{M_{(0)}}, \theta_j, \gamma_j\right)}_{\text{Skilled model rep. prob.}}, \tag{2}$$

- Skilled humans should agree on the answer

$$\delta_j = \underset{i \sim H_{(1)}}{\text{MD}}\left[\sigma_{\text{2pl}}\left(\beta_i^{H_{(1)}}, \theta_j, \gamma_j\right)\right]. \tag{3}$$

- Question should be informative (Fischer information to agent skill):

$$\text{IIF}_j(\theta) = \gamma_j^2 \cdot p_j(\theta) \cdot (1 - p_j(\theta)), \text{ where} \tag{4}$$

$$p_j(\theta) = \sigma_{\text{2pl}}(\theta, \theta_j, \gamma_j). \tag{5}$$

Adversarial Score

Is this a viable incentive structure?

- Can human authors interpret incentive?
  - Computers should get questions wrong, smart humans should get them right
  - Answers should be unique and easily verifiable
  - Reward knowledge and skill
  - Avoid ambiguity
- Posthoc (no realtime feedback): Prizes given based on metric
- Professional trivia writers

Adversarial Strategies

What is the name of the American actor who stood up for his wife with a "slap that was heard around the world" during a popular awards show?

Adversarial Strategies

What is the name of the American actor who stood up for his wife with a "slap that was heard around the world" during a popular awards show?

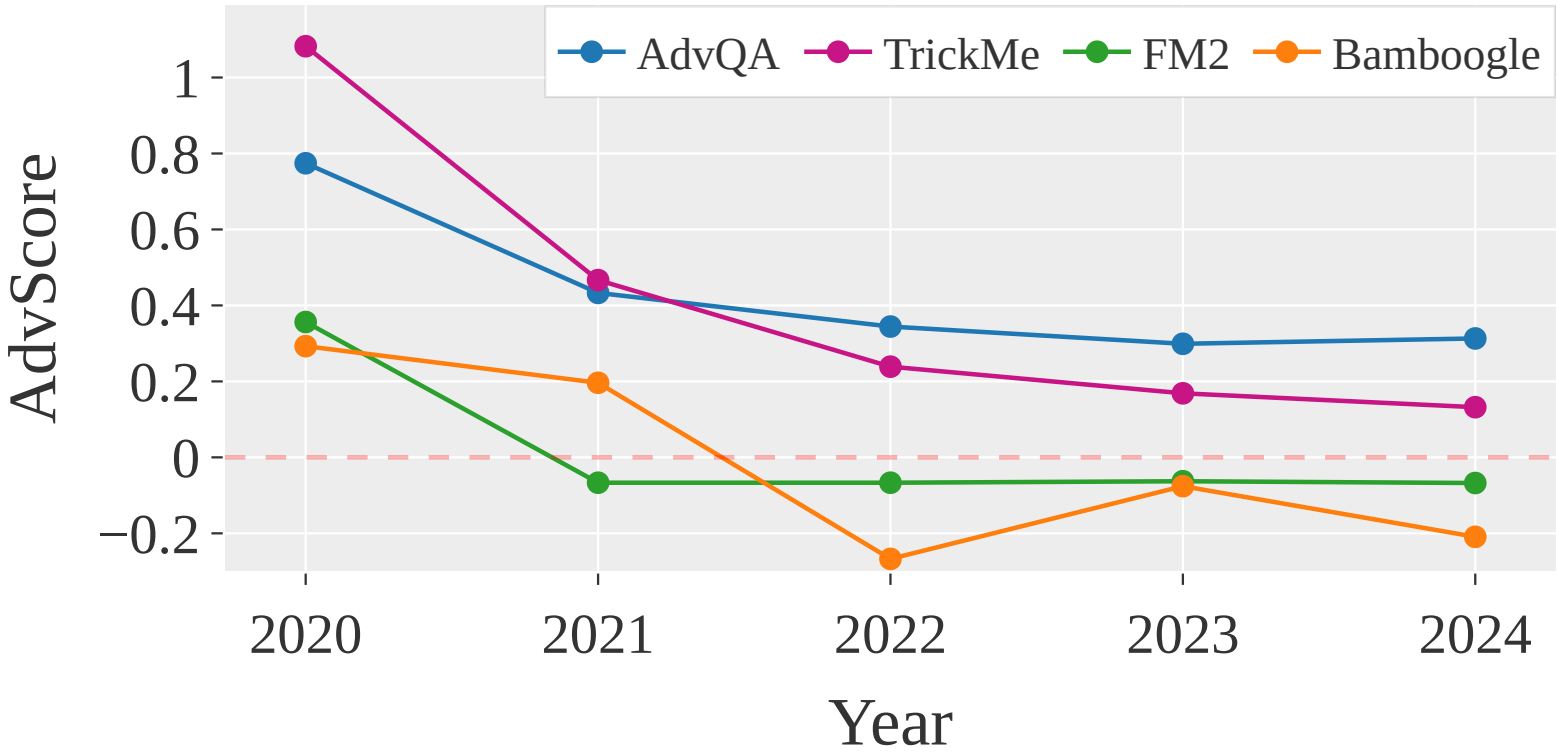
Adversarial Strategies

What post-apocalyptic film directed by a Korean but not the director of Parasite is an allegory set on a train featuring the machinations of a rich businessman against the occupants of other cars?

Adversarial Strategies

What post-apocalyptic film directed by a Korean but not the director of Parasite is an allegory set on a train featuring the machinations of a rich businessman against the occupants of other cars?

Which Datasets are Adversarial?



- Not all datasets remain adversarial forever
- What helps make datasets adversarial?
  - Bamboogle: Automatically generated
  - TrickMe: Human in the loop interface (expert), IR models
  - FM2: Human in the loop interface (crowdworker), IR models
  - AdvCal: Huaman in the loop (expert), LLM model + category

The Value of Repeated Games

- We're only learning per example
- Can we learn more efficiently?
- What if we got feedback per token?

The Value of Repeated Games

- We're only learning per example
- Can we learn more efficiently?
- What if we got feedback per token?
- And improve calibration?

Mimicking How Humans Learn Calibration

- Questions get easier (for humans)
- Humans evaluate whether they know enough to answer
- If they answer too early, they get "locked out" from the rest of the questions


PIENSE
THINK
सोचिए
ΣΚΕΨΟΥ
DENKE
PENSER
$300,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$1,000
$1,000,000
Who is Bram Stoker?
$ 17,973
$200,000
WHO IS BRAM STOKER?
$5600

This is **not** Jeopardy

- Jeopardy: must decide to answer **once**, after complete question
- Quiz Bowl: decide after each word

Sample Question

With Leo Szilard, he invented a doubly-eponymous

Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of

Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so

Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients

Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by

Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the

Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Albert Einstein

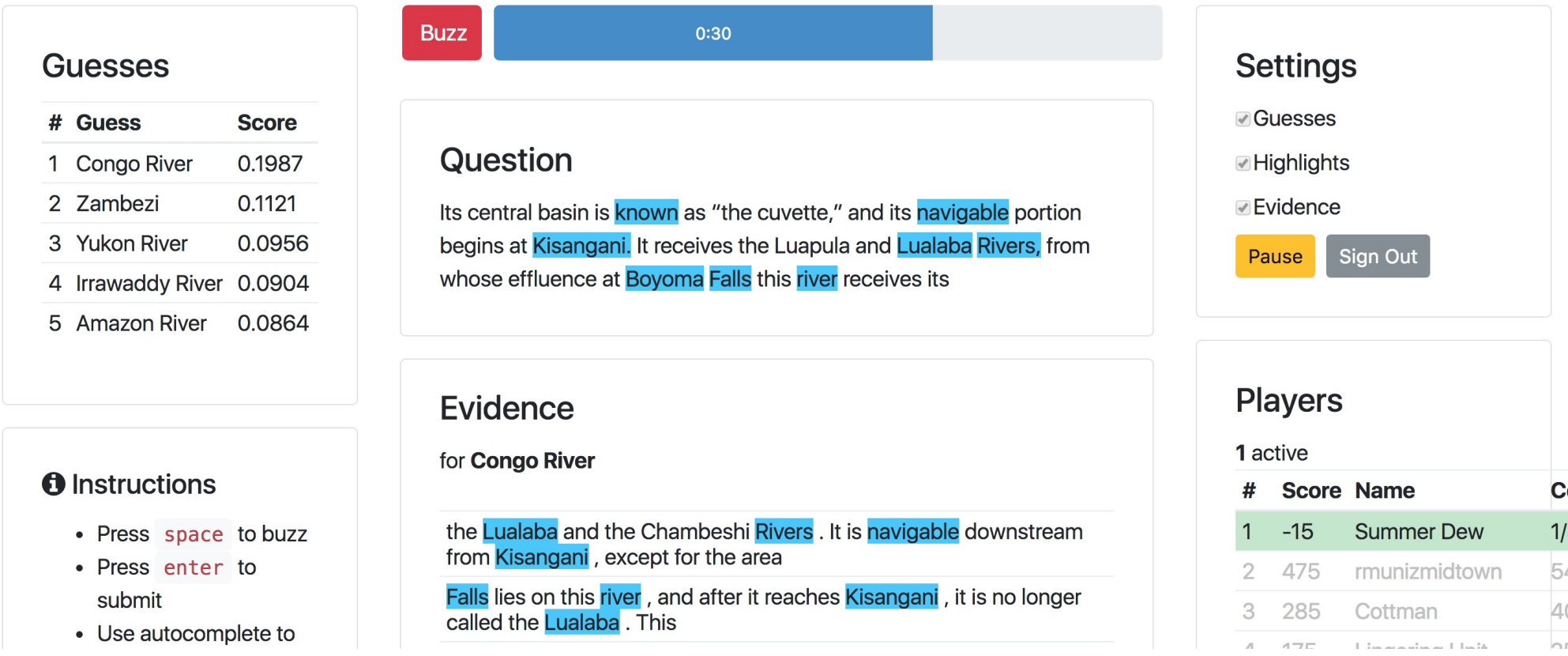
Sample Question

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into

Faster = Smarter, More Calibrated

1. University of Chicago
2. Colorado School of Mines
3. Cornell University
4. UIUC
5. Brigham Young University
6. California Institute of Technology
7. Peking University
8. Harvey Mudd College
9. Darmstadt University
10. University of Colorado

Centaur Chess


Guesses
Guess Score
1 Congo River 0.1987
2 Zambezi 0.1121
3 Yukon River 0.0956
4 Irrawaddy River 0.0904
5 Amazon River 0.0864
Instructions
Press space to buzz
Press enter to submit
Use autocomplete to
Buzz
0:30
Question
Its central basin is known as "the cuvette," and its navigable portion begins at Kisangani. It receives the Luapula and Lualaba Rivers, from whose effluence at Boyoma Falls this river receives its
Evidence
for Congo River
the Lualaba and the Chambeshi Rivers . It is navigable downstream from Kisangani , except for the area
Falls lies on this river , and after it reaches Kisangani , it is no longer called the Lualaba . This
Settings
Guesses
Highlights
Evidence
Pause
Sign Out
Players
1 active
Score Name
1 -15 Summer Dew
2 475 rmunizmidtown
3 285 Cottman

Interface

Guesses

#	Guess	Score
1	Congo River	0.1987
2	Zambezi	0.1121
3	Yukon River	0.0956

Question

Its central basin is known as “the cuvette,” and its navigable portion begins at Kisangani. It receives the Luapula and Lualaba Rivers, from whose effluence at Boyoma Falls this river receives its

Highlighting

Evidence

for **Congo River**

the Lualaba and the Chambeshi Rivers . It is navigable downstream

---

Falls lies on this river , and after it reaches Kisangani , it is no longer

---

from Kisangani , except for the area

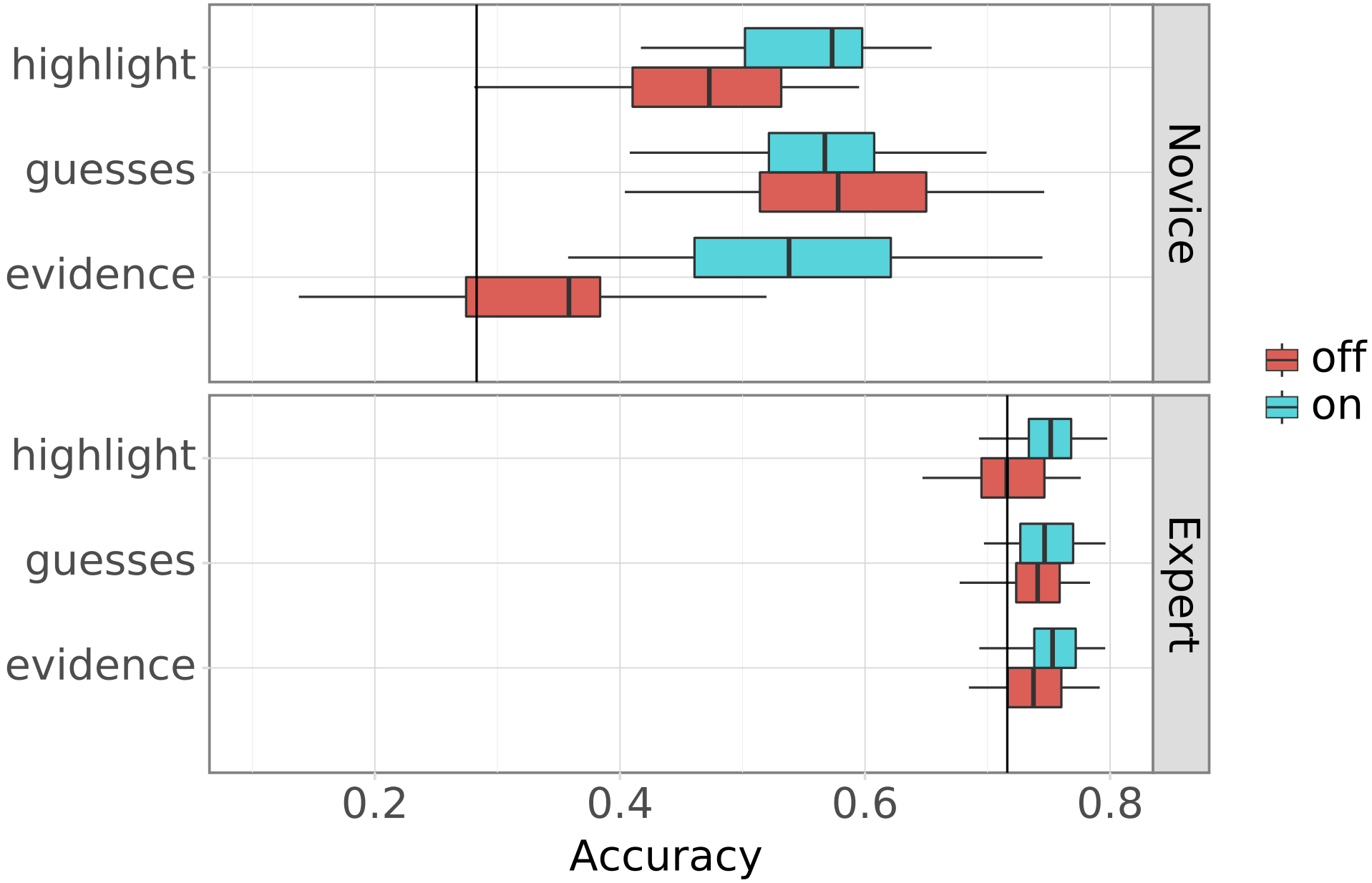
Experts vs. Novices

Experts

Trivia experts, familiar with task, enjoy the task

Mechanical Turkers

Mechanical Turkers: easily overwhelmed, need the help



Evidence helps novices, experts are expert

Regression Analysis

For each triple (player, question, interpretations),
predict the outcome (correct answer or not)
with logistic regression. Using features:

- player ID
- question ID
- buzzing position
- enabled interpretations: individual and combinations

Regression Analysis

For each triple (player, question, interpretations),
predict the outcome (correct answer or not)
with logistic regression. Using features:

- player ID
- question ID
- buzzing position
- enabled interpretations: individual and combinations

Coefficients tell story!

- **Big, Positive**: Help
- **Big, Negative**: Hurt
- **Small**: Neutral

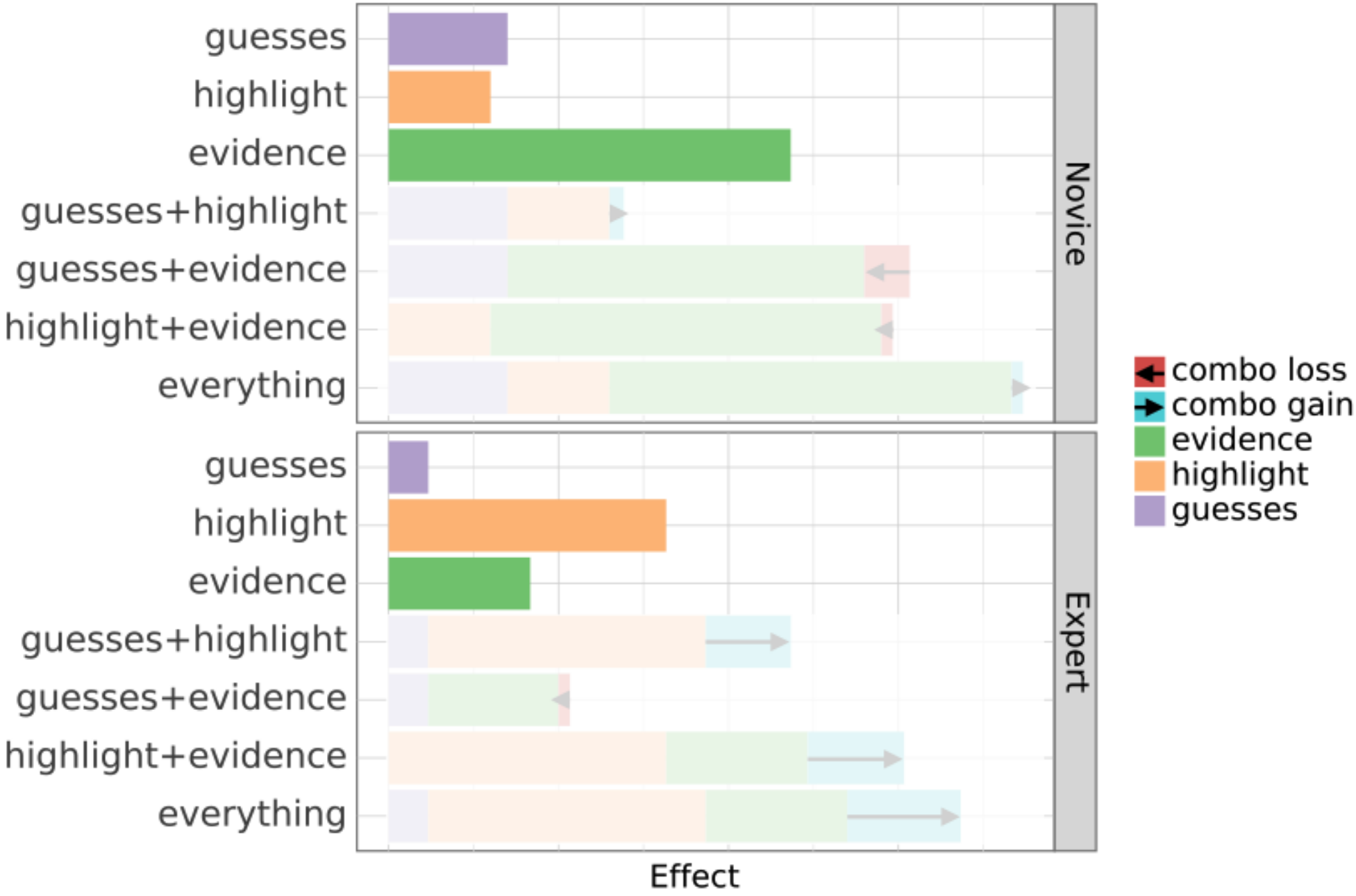
Regression Analysis

For each triple (player, question, interpretations),
predict the outcome (correct answer or not)
with logistic regression. Using features:

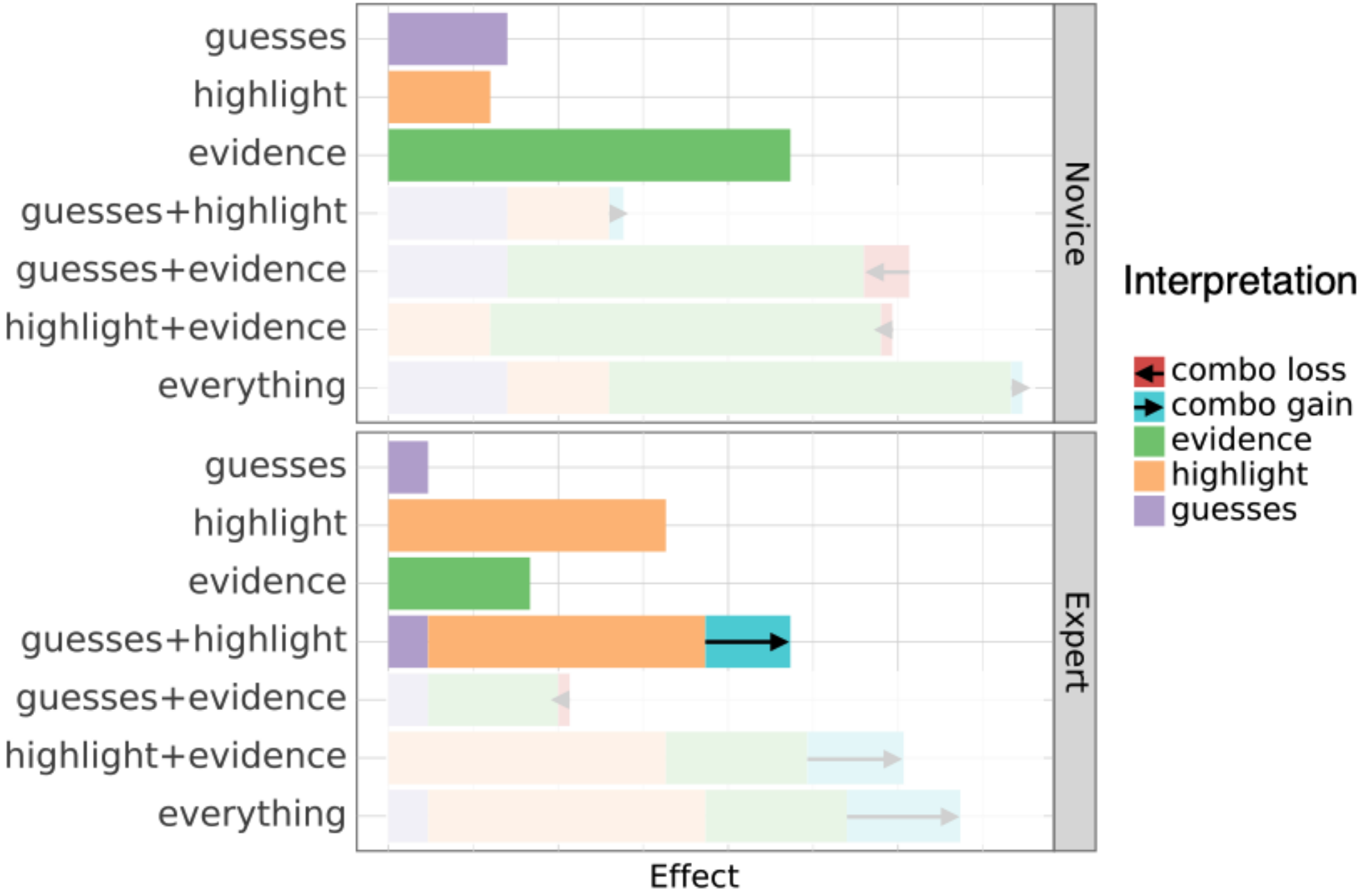
- player ID
- question ID
- buzzing position
- enabled interpretations: individual and combinations

Coefficients tell story!

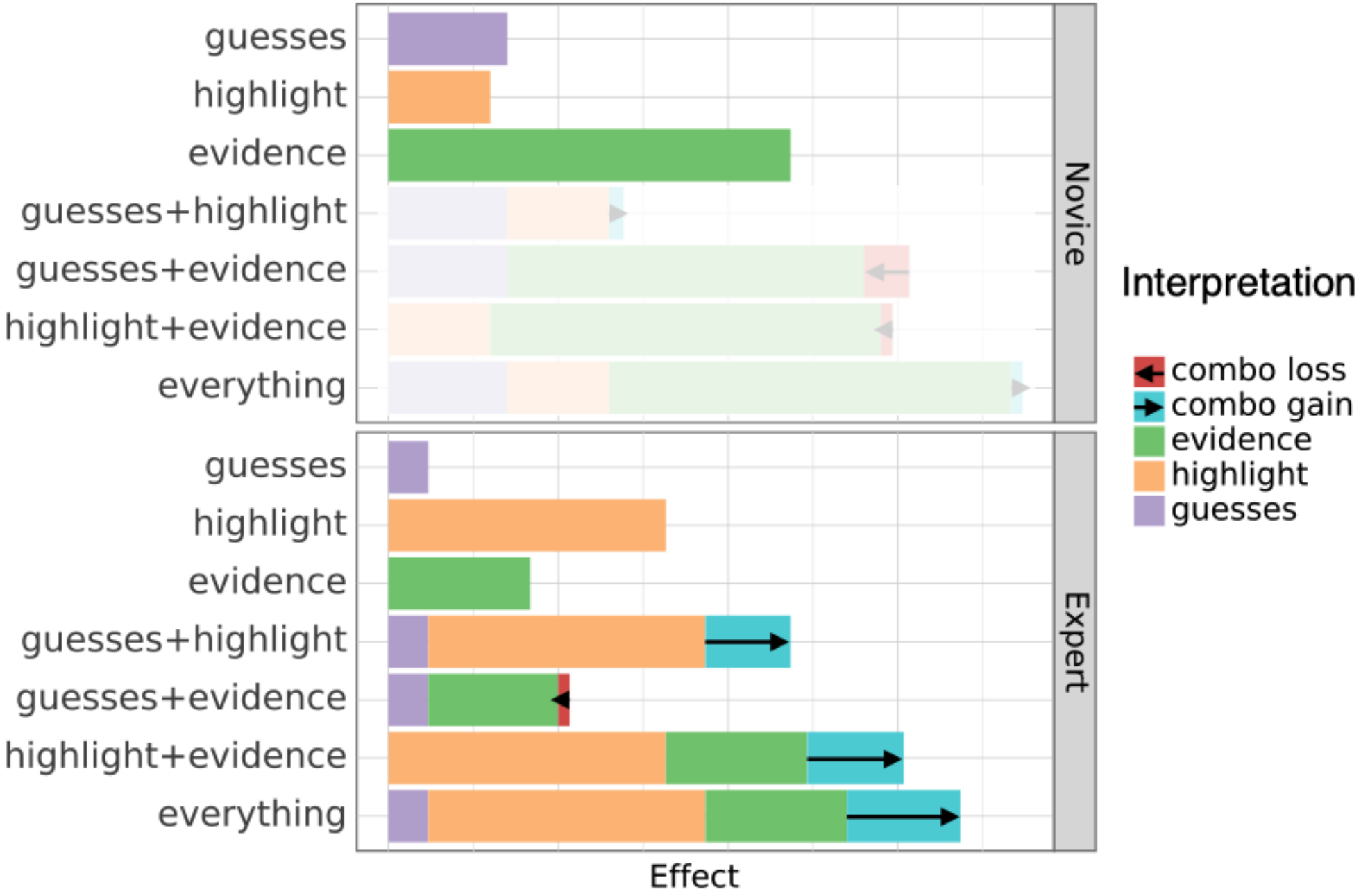
- **Big, Positive**: Help
- **Big, Negative**: Hurt
- **Small**: Neutral



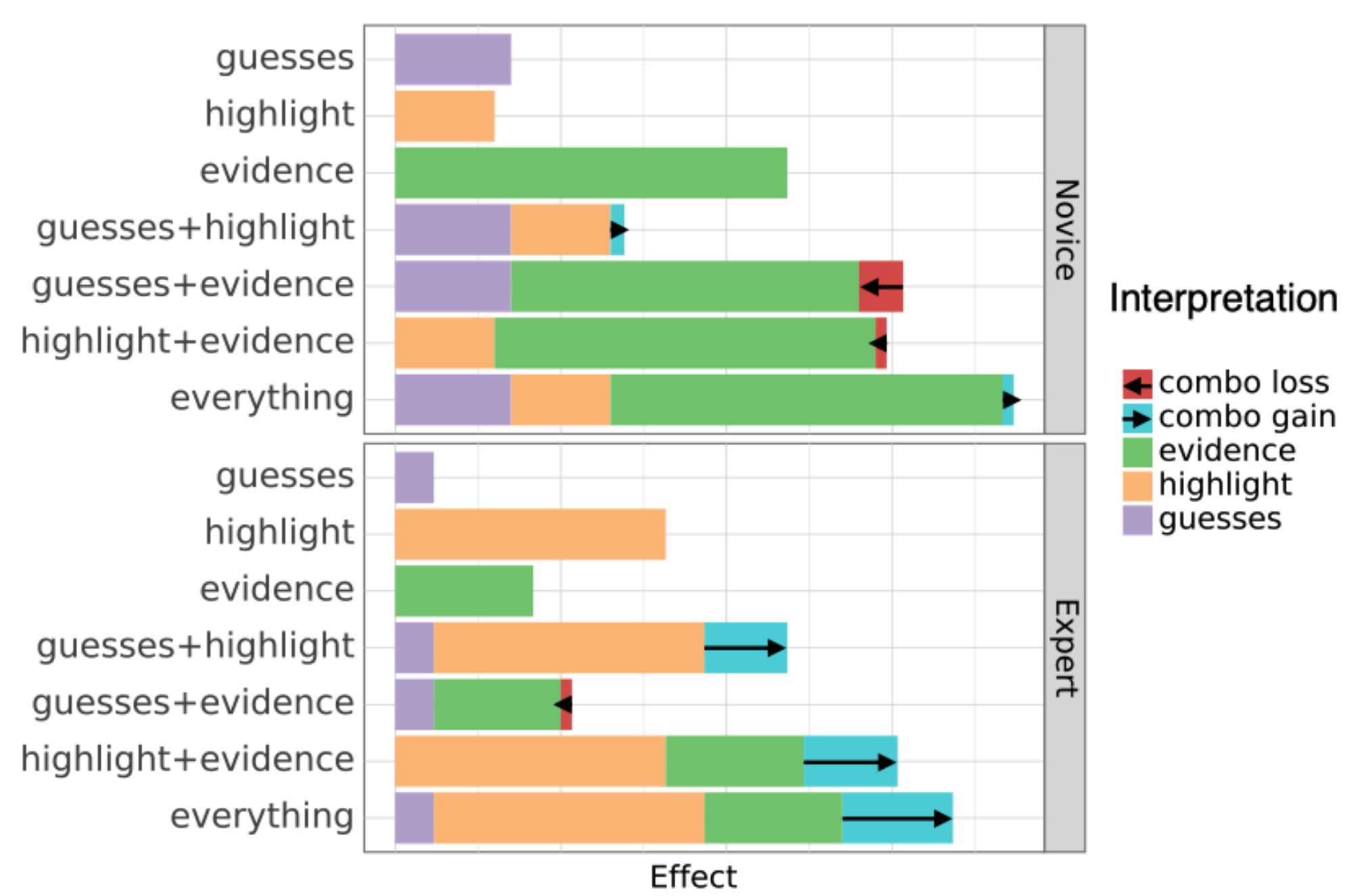
Everything helps: Evidence for novies, Highlight for experts



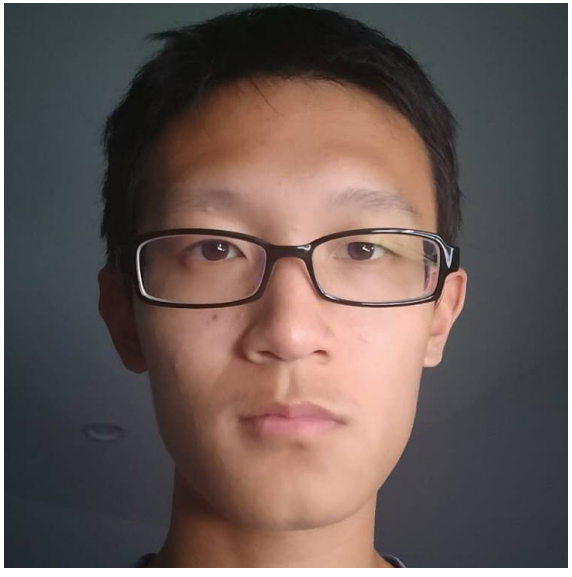
Synergistic effects



Highlight and evidence help experts most



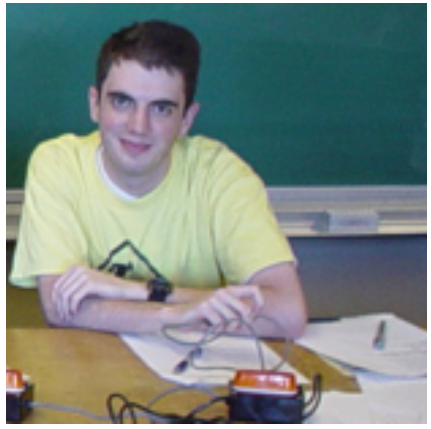
For novices, less synergy



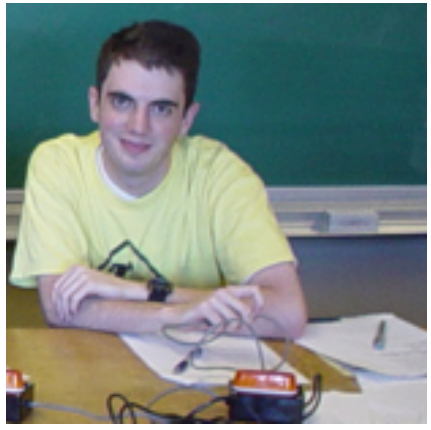
Learning to Explain Selectively

Shi Feng and **Jordan Boyd-Graber**.
Empirical Methods in Natural Language Processing, 2022

Measuring Interpretability



Measuring Interpretability



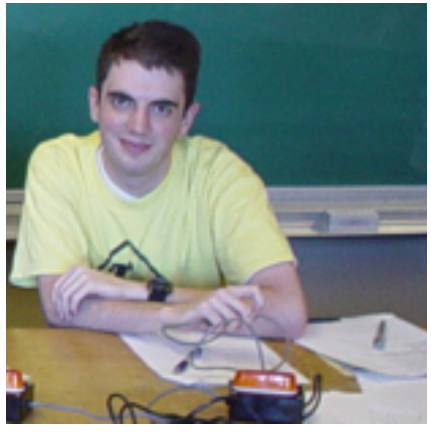
Solo

6

Measuring Interpretability

Mauser Rifle
Aristophanes
Battle of Marathon
Battle of Waterloo
Quadratic Equation
Hypotenuse
Raphael

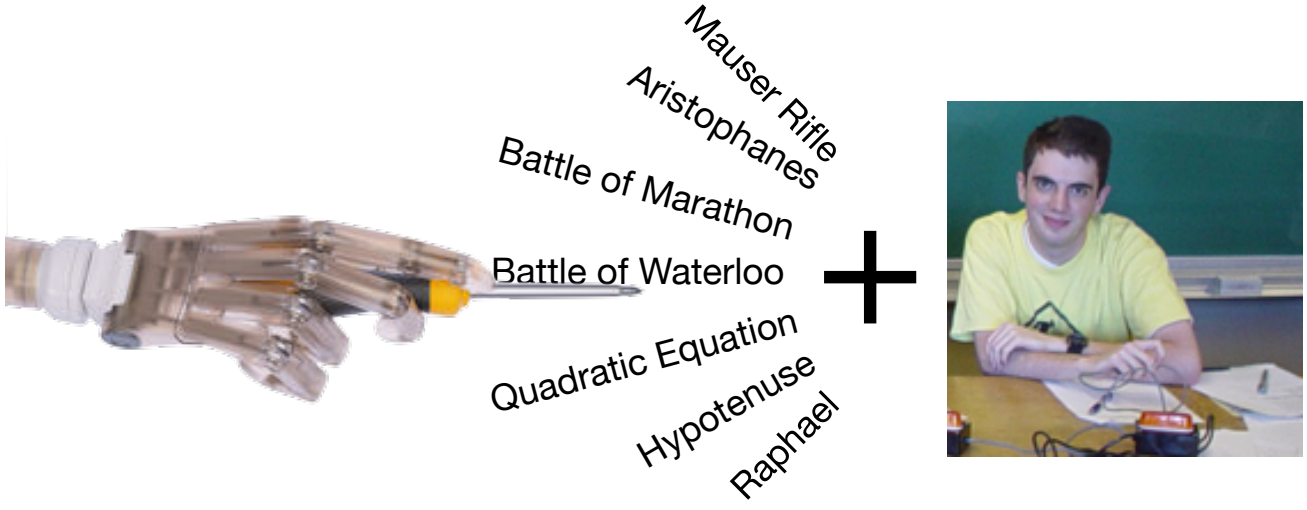
+



Solo

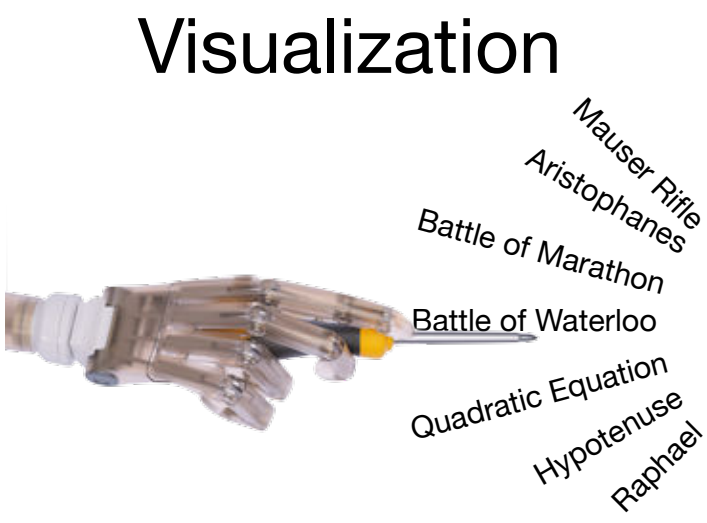
6

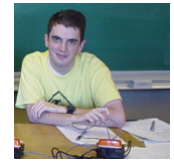
Measuring Interpretability



Exp A	Exp B	Exp C	Solo
5.5	7.3	6.1	6

Improvement through Bandit Algorithms

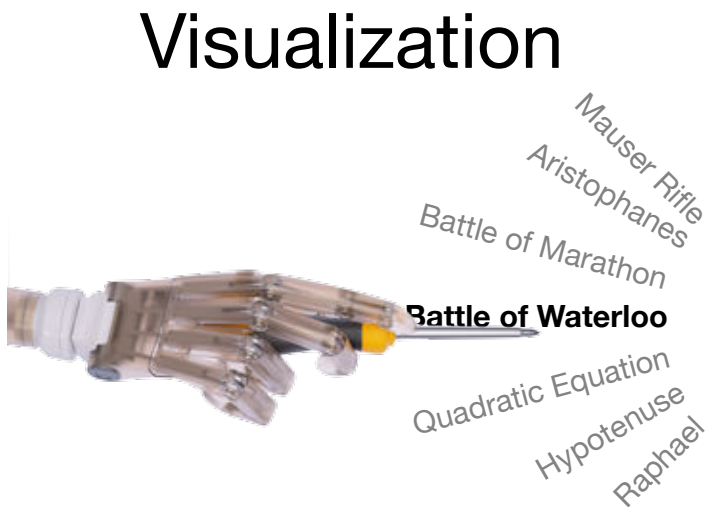


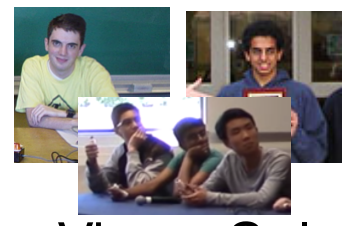


Viz	Solo
7.3	6

Improvement through Bandit Algorithms

Improvement through Bandit Algorithms

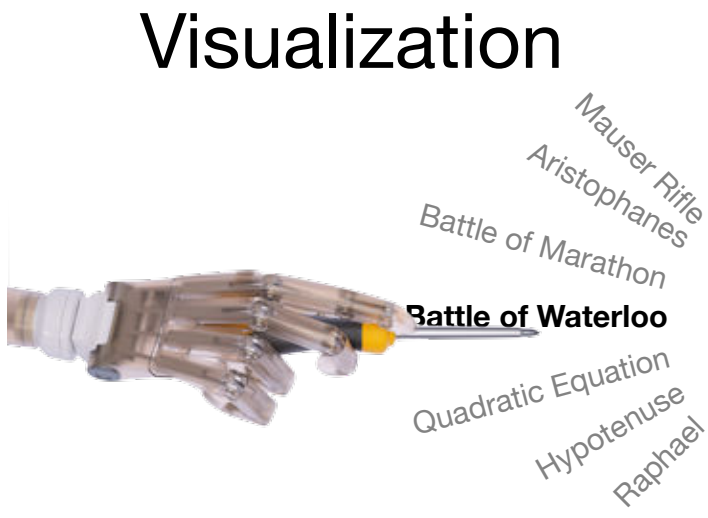


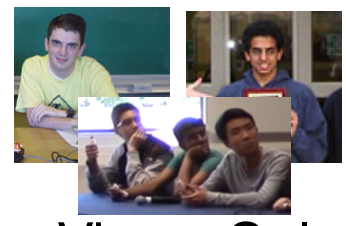


Viz Solo

7.3 6

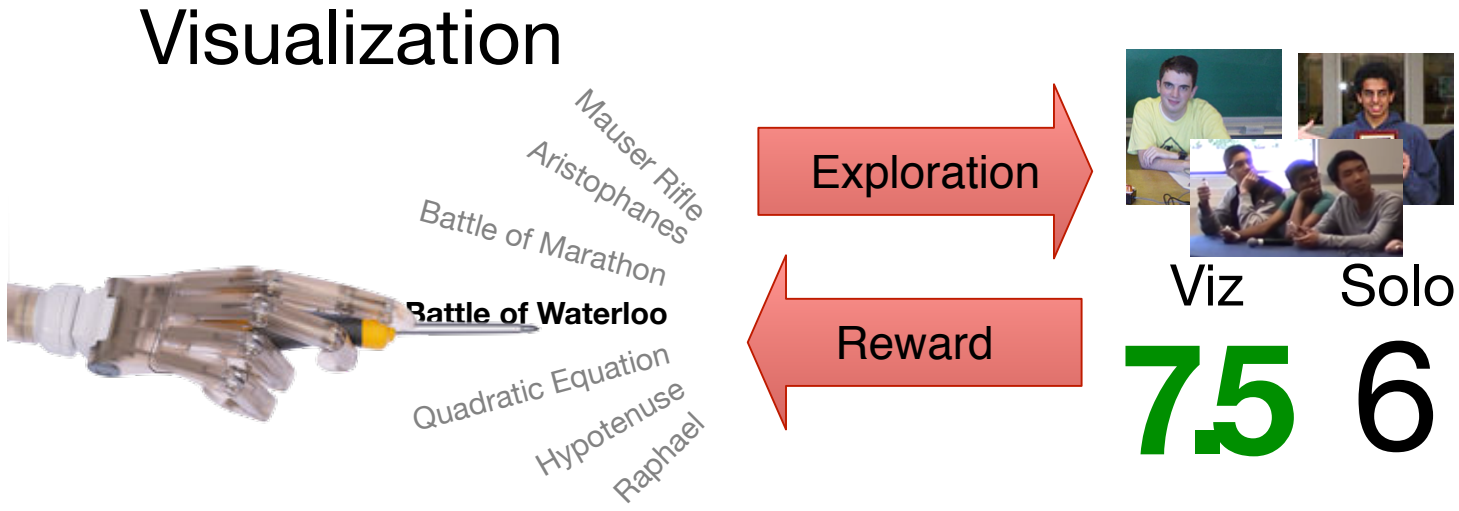
Improvement through Bandit Algorithms



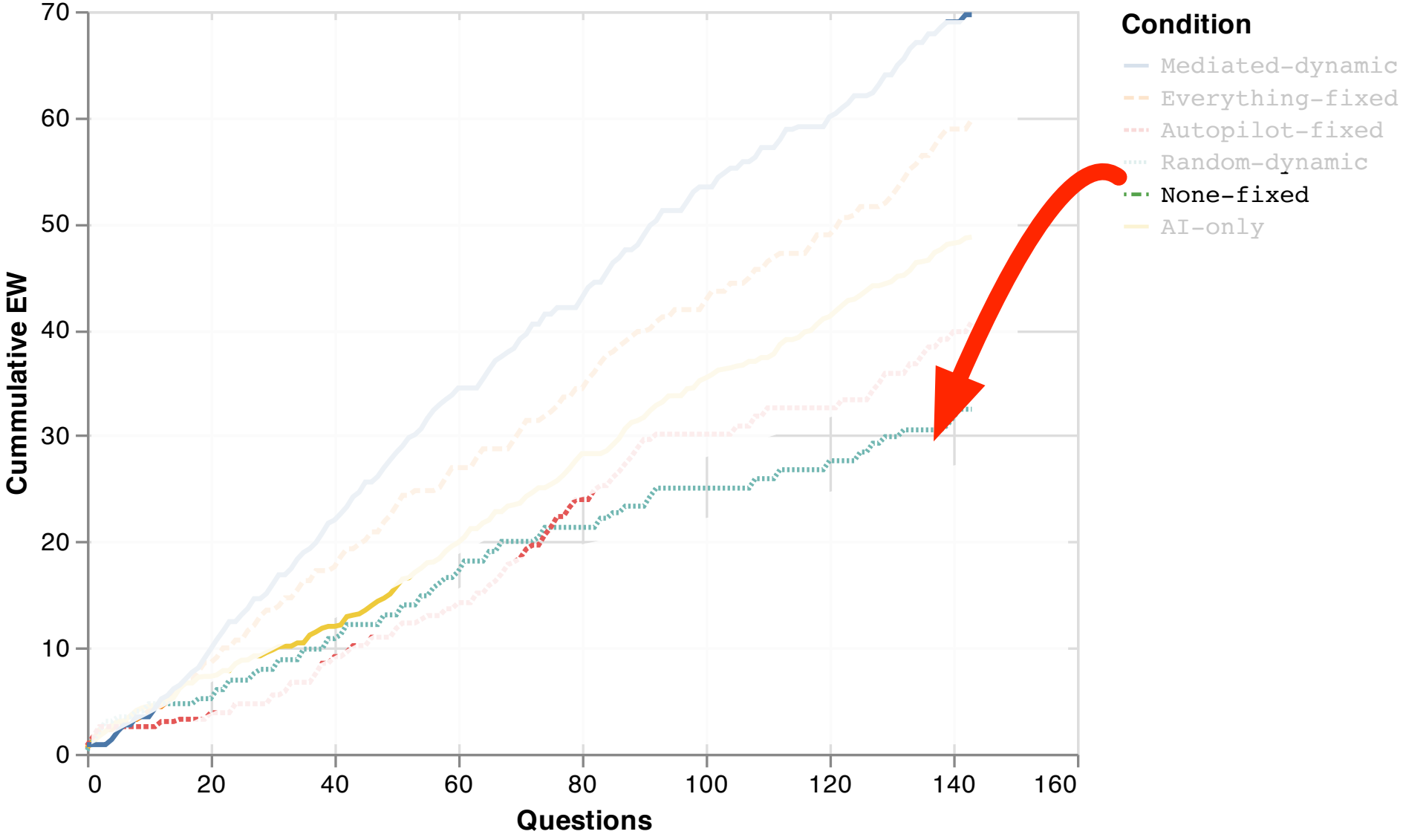


Viz	Solo
7.5	6

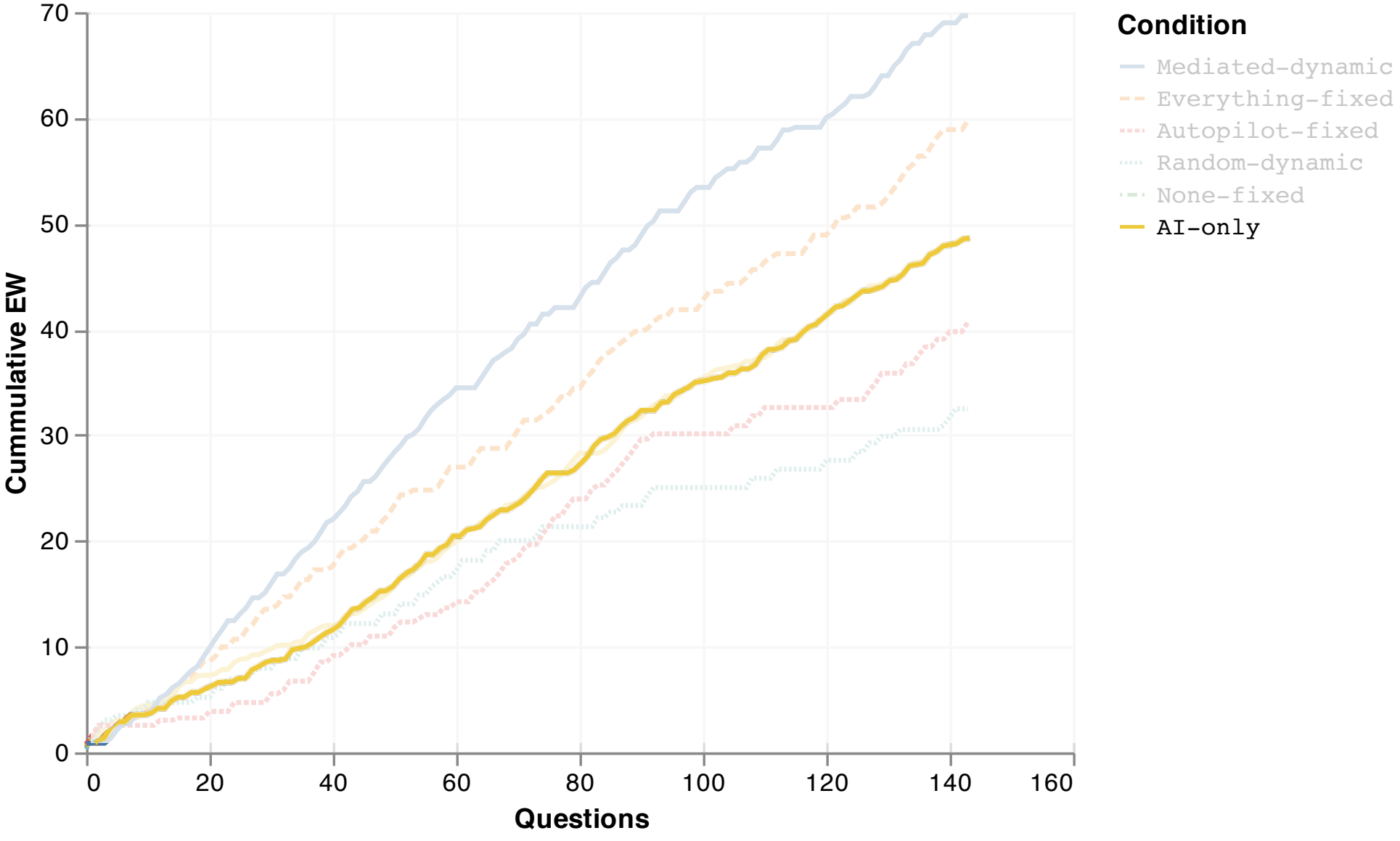
Improvement through Bandit Algorithms



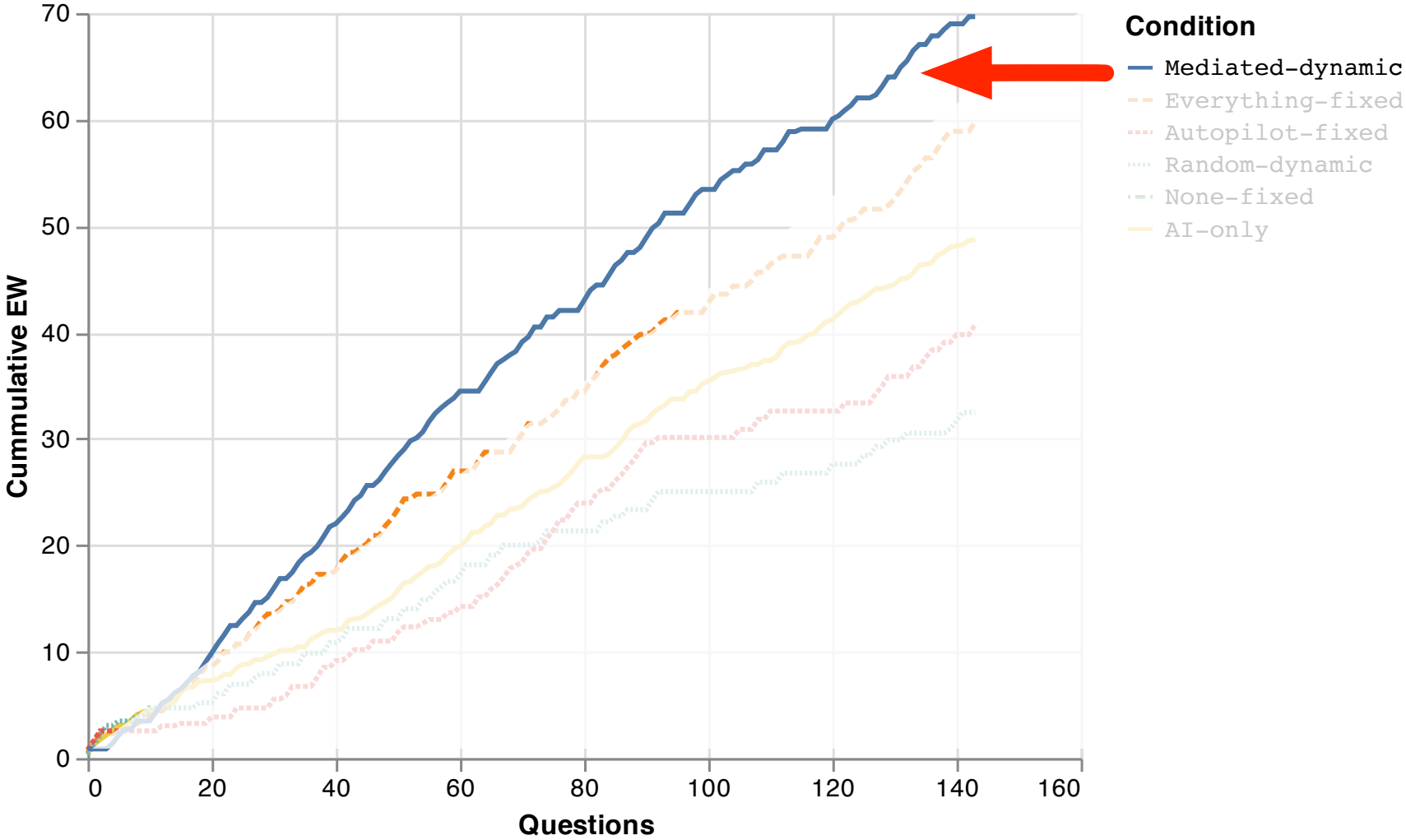
Bandit actions [**?**]: turn each of the explanations (Guess, Highlight, Evidence) on or off.



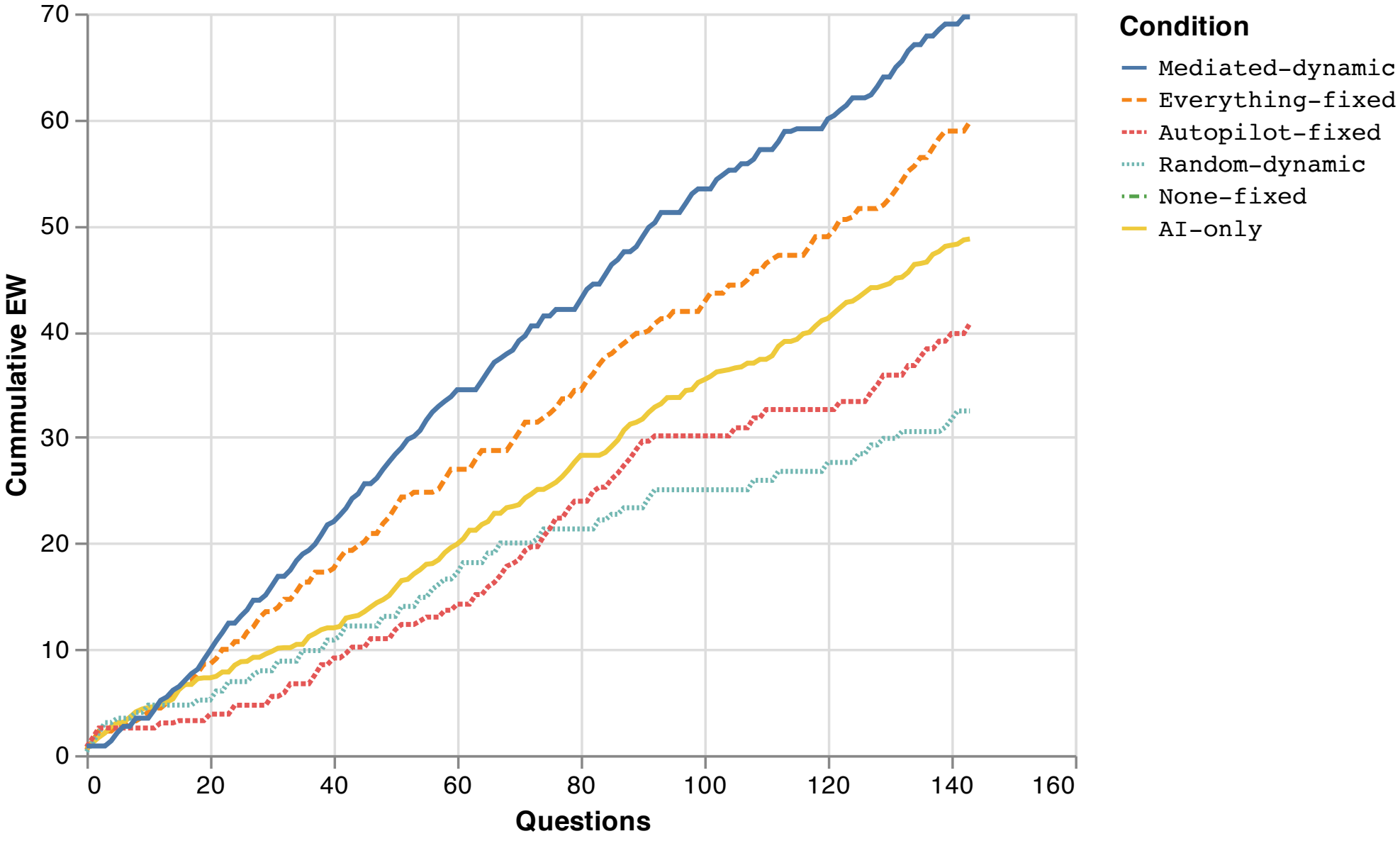
Human alone without an AI teammate



AI alone without a human teammate



Dynamic assistance to human

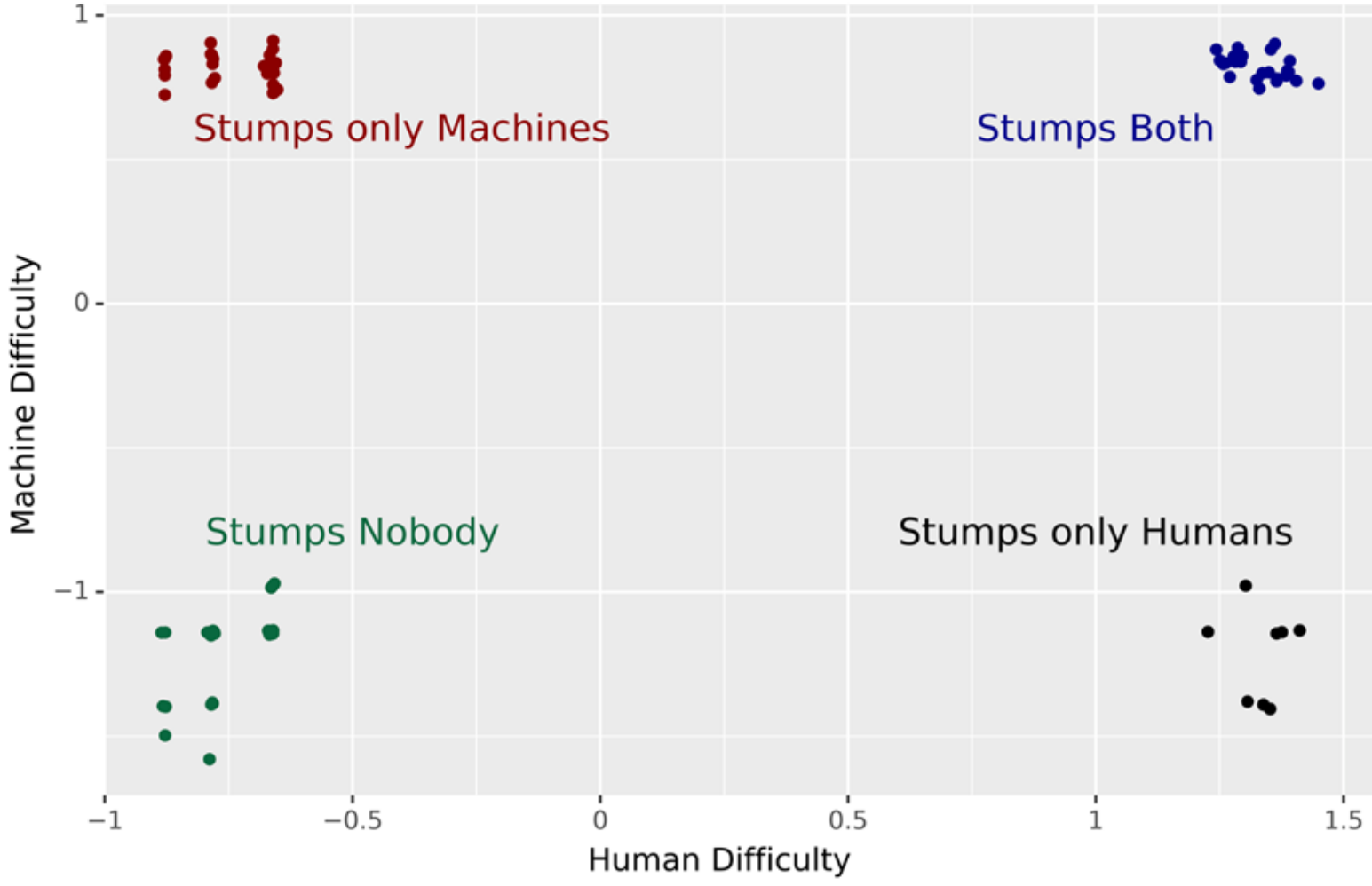


Better than showing everything!

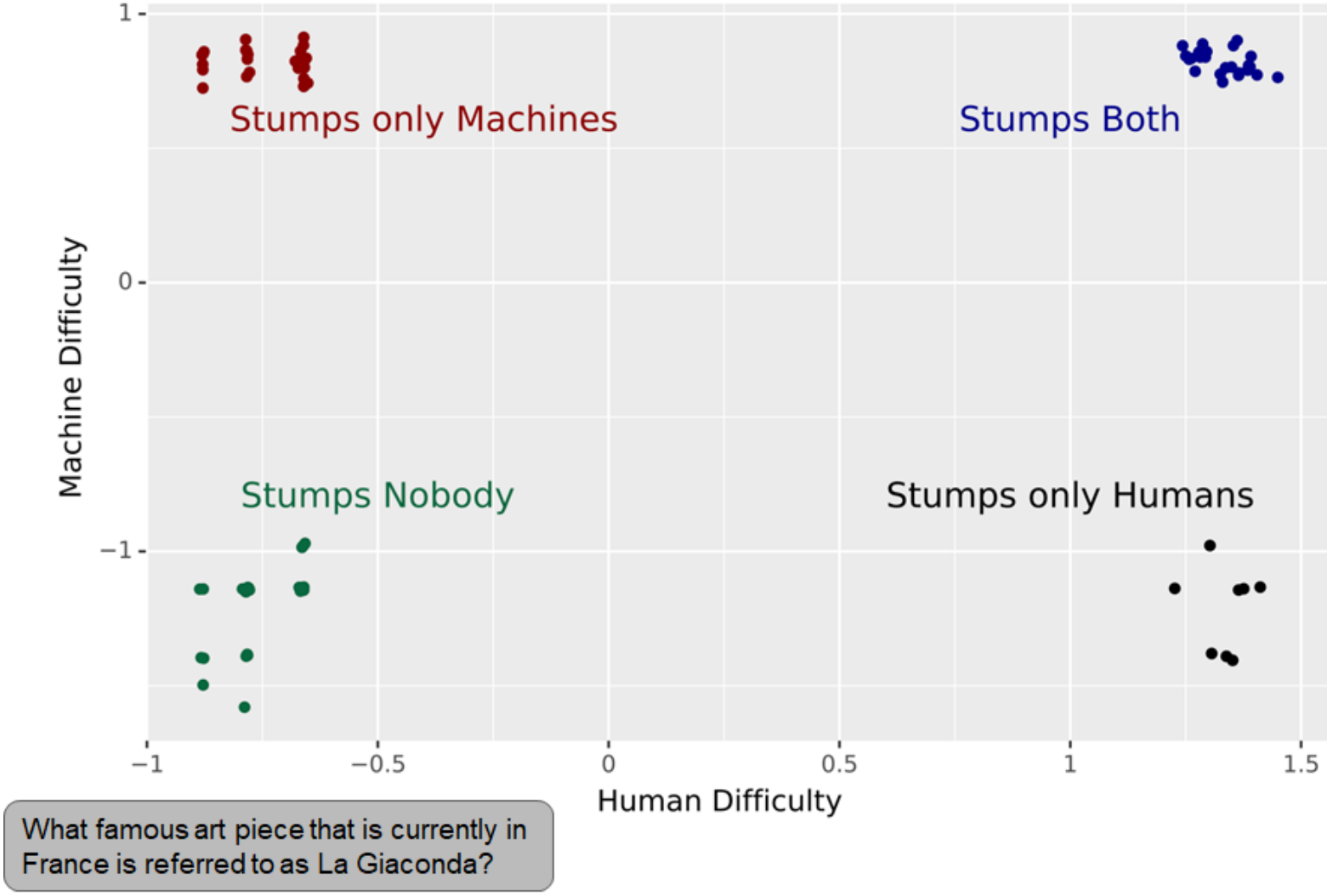
Calibration is hard

- If they knew when they were making stuff up, this wouldn't be a problem But LLMs are notoriously bad at knowing when they don't know
  - Depends on length of generation
  - Depends on frequency of response
  - Depends on reasoning
  - Depends on tokenization
- Even our metrics for knowing when the uncertainty is bad are flawed
- And doing a good job of detection requires deeper access to the model

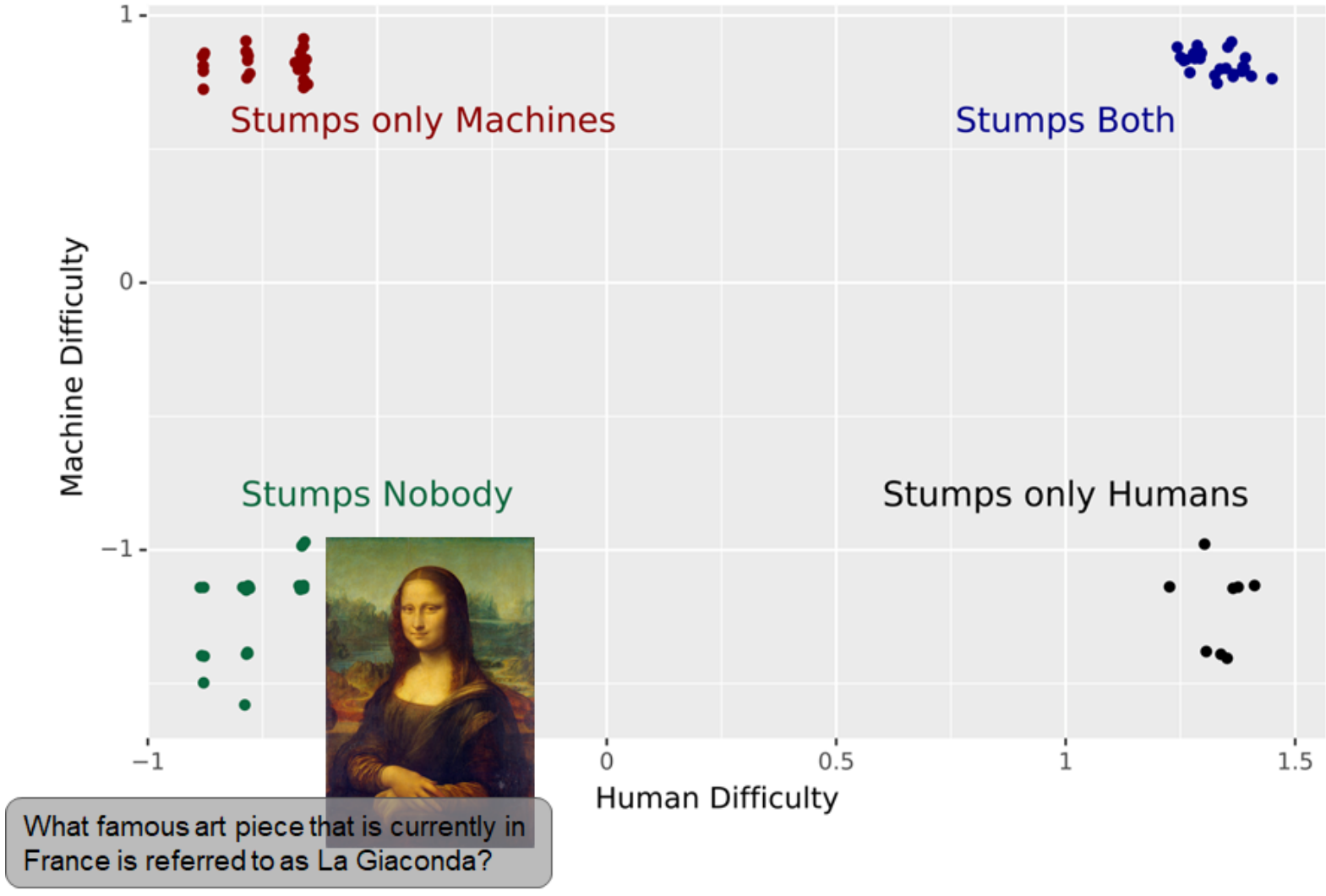
What makes for Adversarial Example



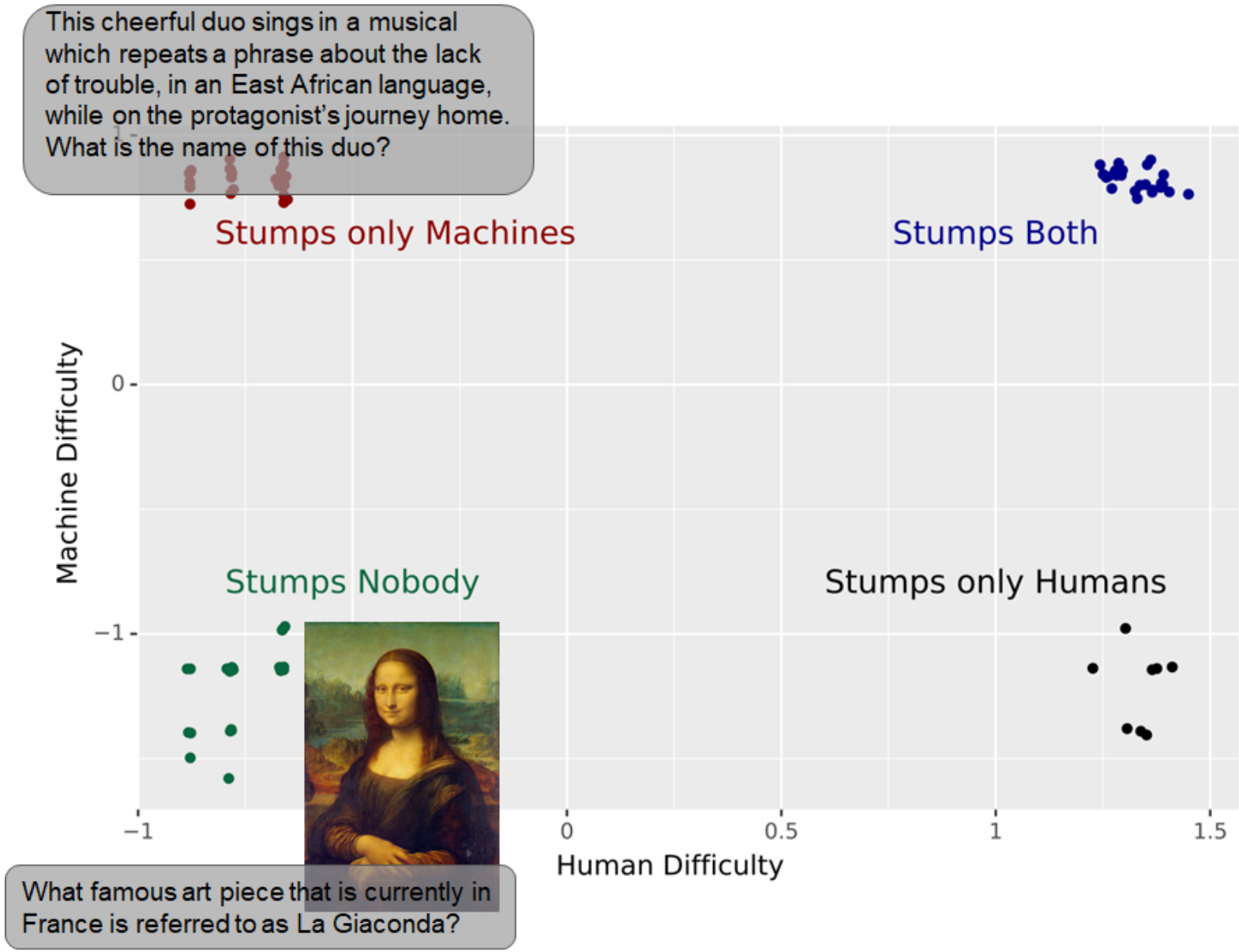
What makes for Adversarial Example



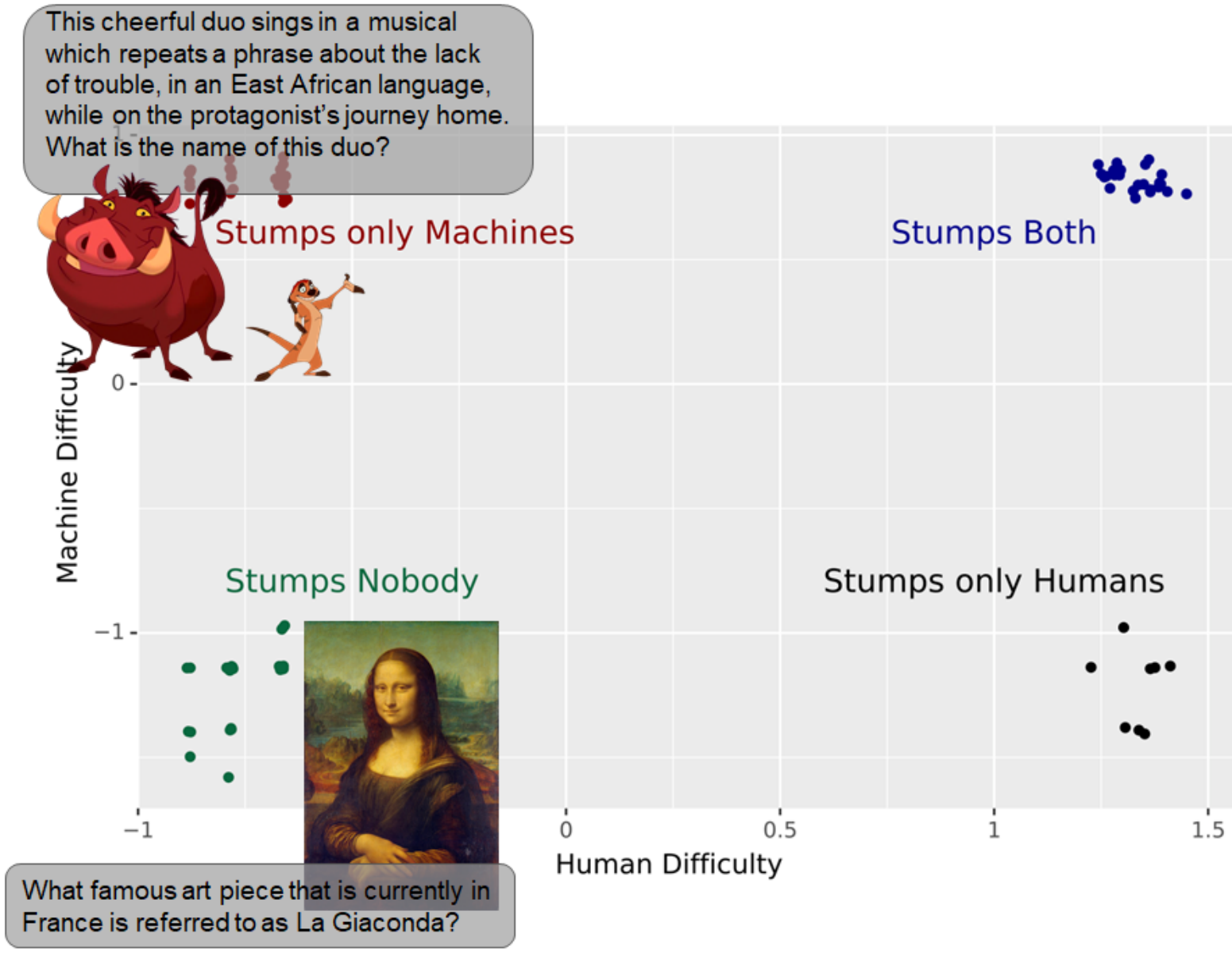
What makes for Adversarial Example



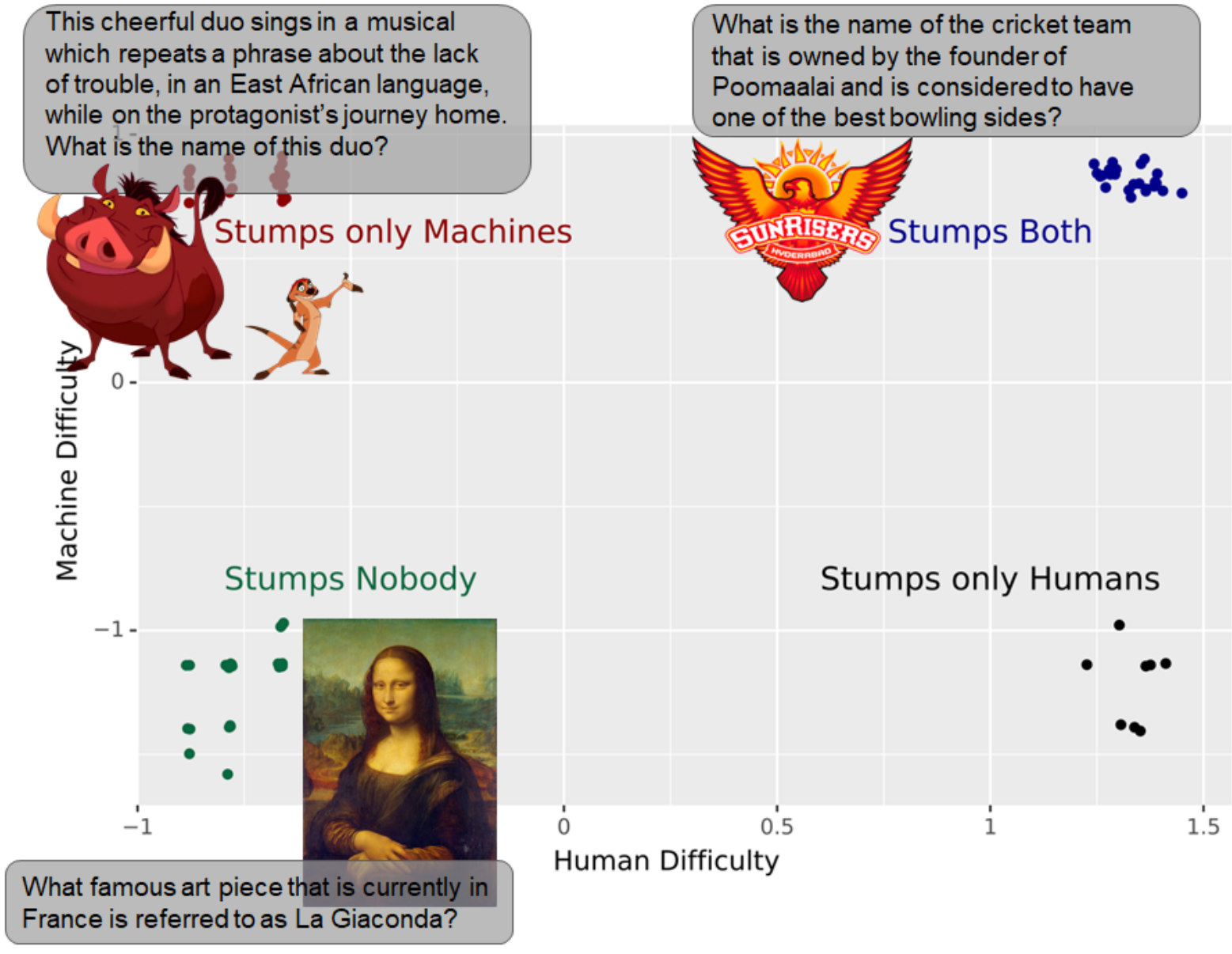
What makes for Adversarial Example



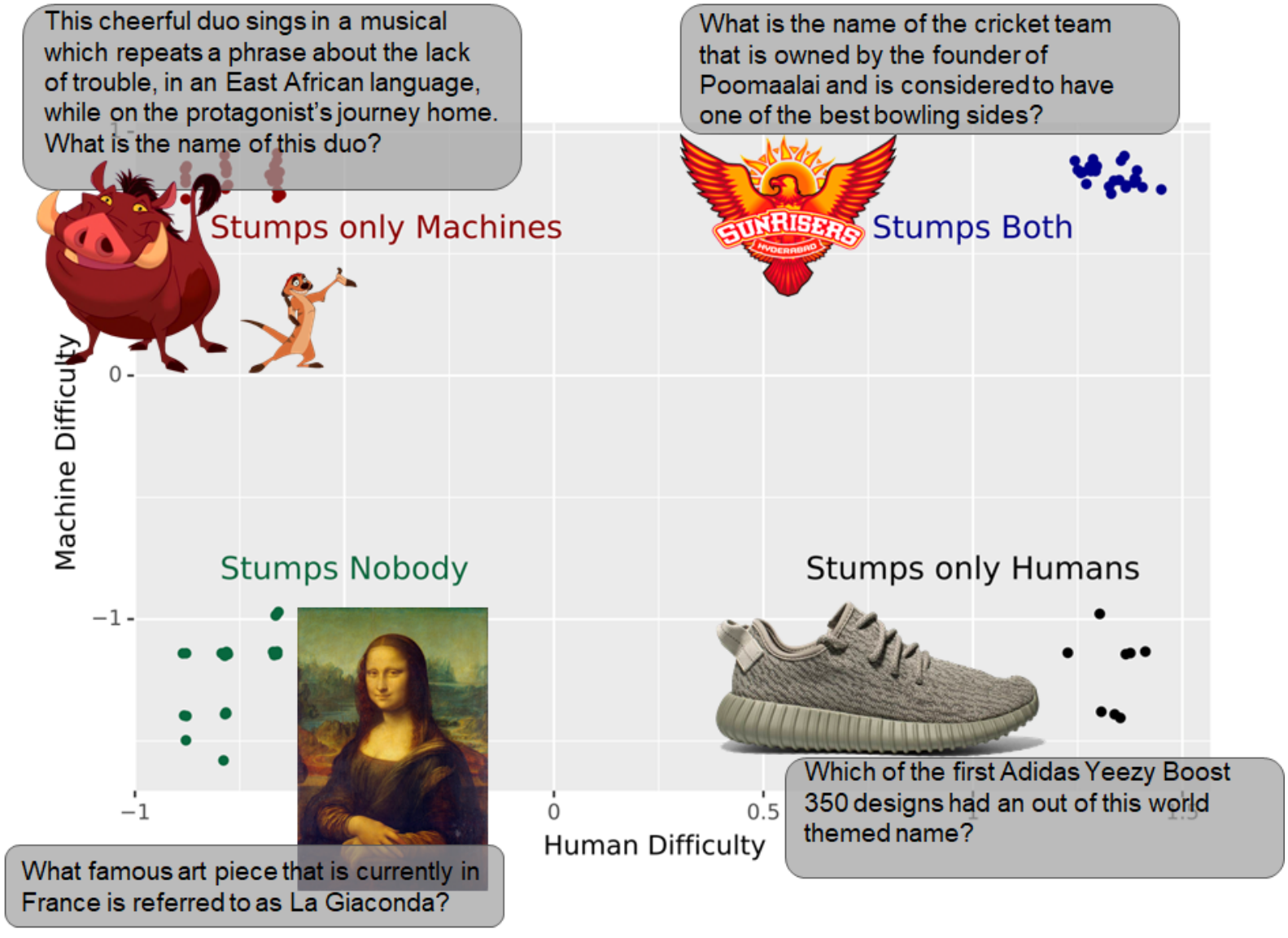
What makes for Adversarial Example



What makes for Adversarial Example

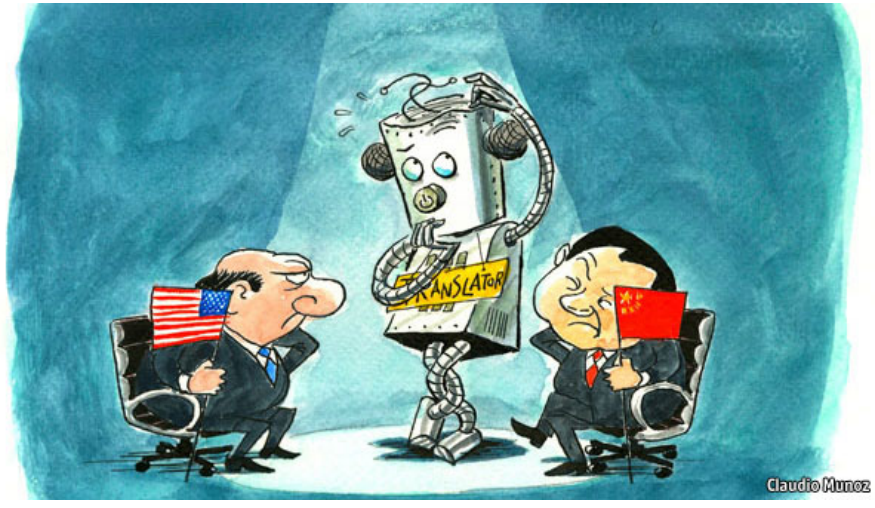


What makes for Adversarial Example



Simultaneous Interpretation is Hard!

- Exhausting for humans
- Computers not trusted
- Differential strengths
- Same word-by-word characteristic



Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation

Alvin Grissom II, He He, **Jordan Boyd-Graber**, John Morgan, and Hal Daumé III. *Empirical Methods in Natural Language Processing*, 2014

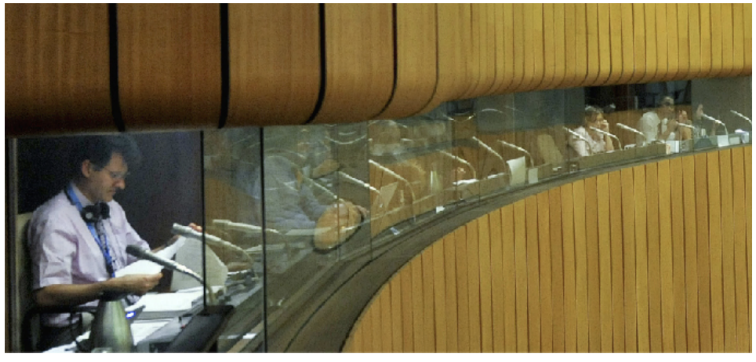
Interpretese vs. Translationese: The Uniqueness of Human Strategies in Simultaneous Interpretation

He He, **Jordan Boyd-Graber**, and Hal Daumé III. *North American Association for Computational Linguistics*, 2016

SimQA: Detecting Simultaneous MT Errors through Word-by-Word Question Answering

HyoJung Han, Marine Carpuat, **Jordan Boyd-Graber**. *Empirical Methods in Natural Language Processing*, 2022

STACL: Simultaneous Translation with Integrated Anticipation & Controllable Latency



Liang Huang

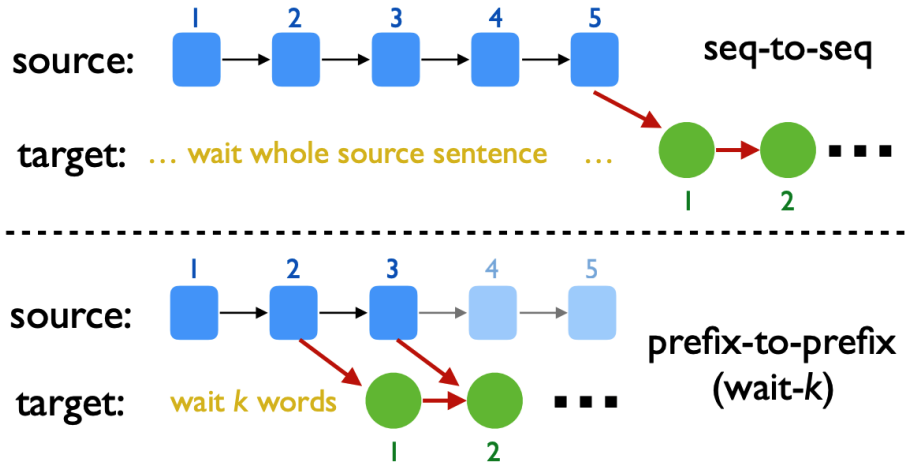
Principal Scientist, Baidu Research

Assistant Professor (on-leave), Oregon State University

Joint work between Baidu Research (Sunnyvale) and Baidu NLP (Beijing)

Prefix-to-Prefix Translation

- seq-to-seq is only suitable for conventional full-sentence MT
- we propose prefix-to-prefix, tailed to simultaneous MT
  - special case: wait-*k* policy: translation is always *k* words behind source sentence
  - training in this way enables anticipation



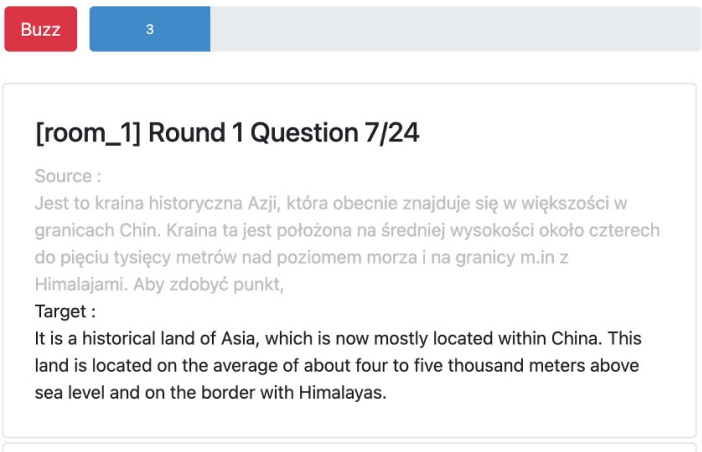
Bùshí	zǒngtǒng	zài	Mòsīkē	yǔ	Éluósī	zǒngtǒng	Pǔjīng	huìwù
布什	总统	在	莫斯科	与	俄罗斯	总统	普京	会晤
Bush	President	in	Moscow	with	Russian	President	Putin	meet

President Bush meets with Russian President Putin in Moscow

How to Evaluate

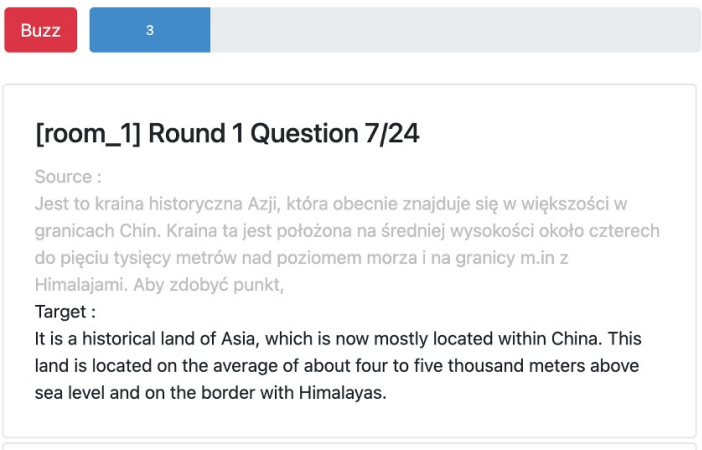
- You're a contestant on a Polish game show
- You have access to a simultaneous translation system
- Your job is to answer the question before your opponent (as quickly as possible)

How to Evaluate



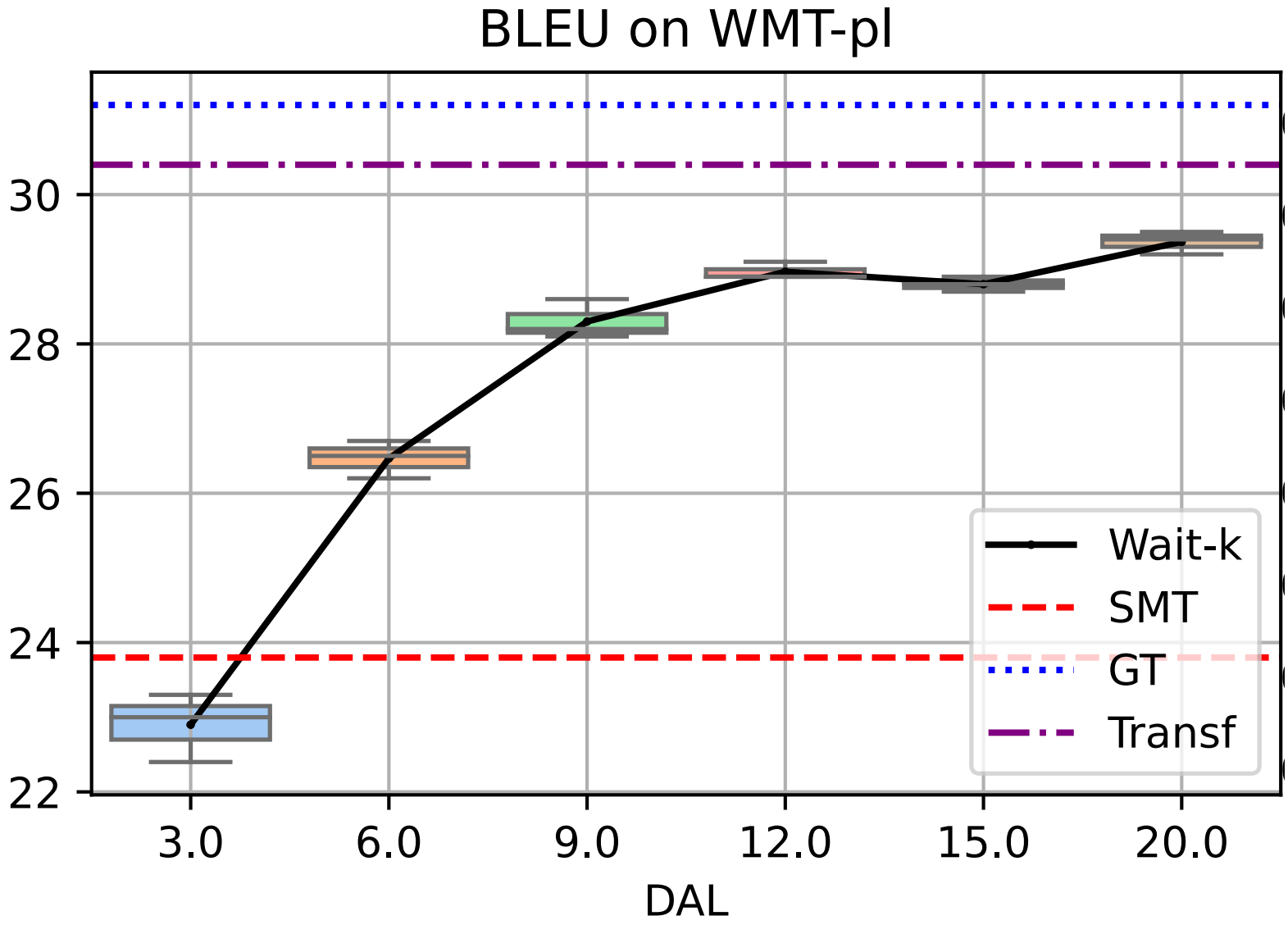
- You're a contestant on a Polish game show
- You have access to a simultaneous translation system
- Your job is to answer the question before your opponent (as quickly as possible)

How to Evaluate

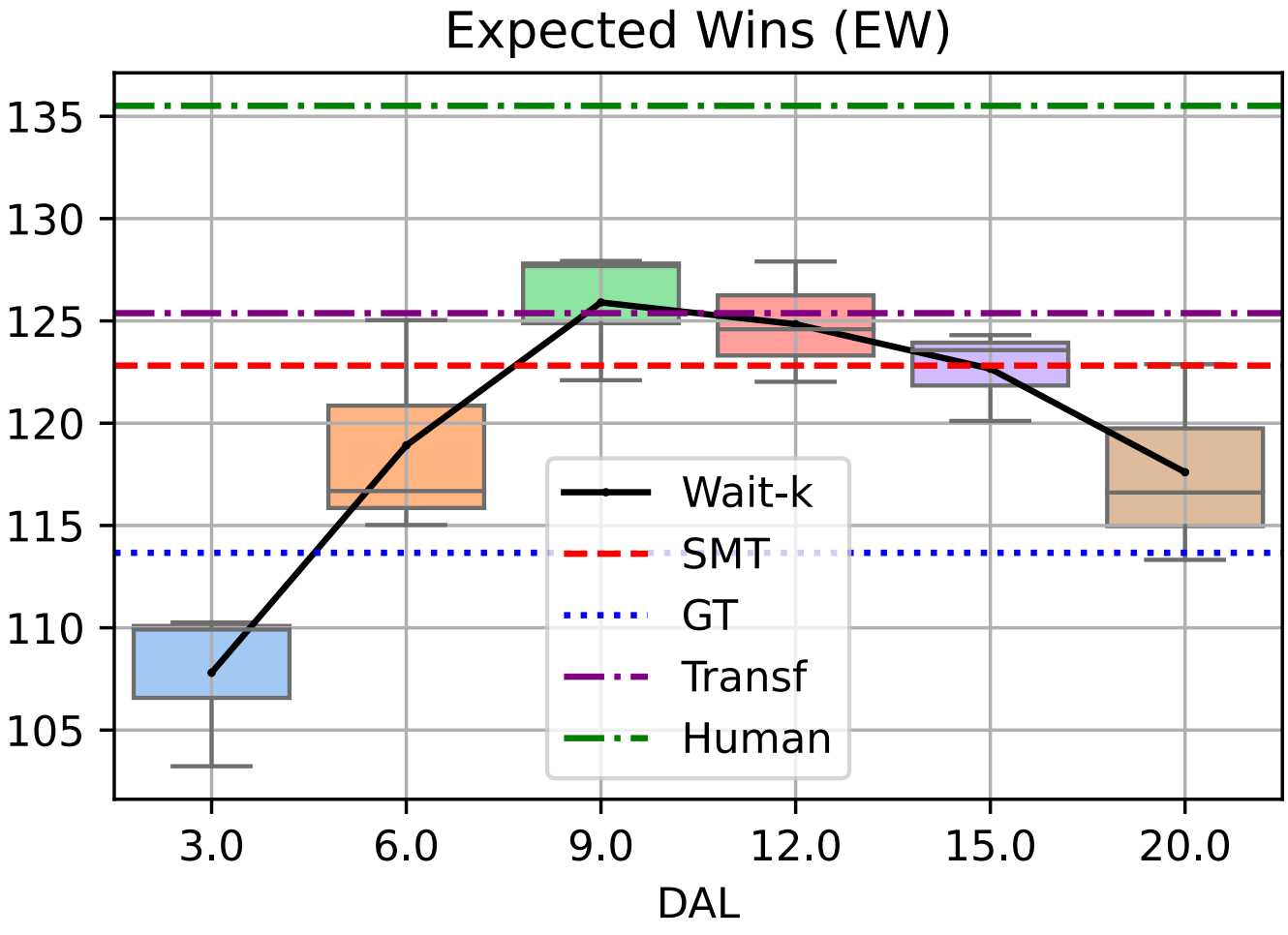


- You're a contestant on a Polish game show
- You have access to a simultaneous translation system
- Your job is to answer the question before your opponent (as quickly as possible)
- Keep question answerer the same, vary translation

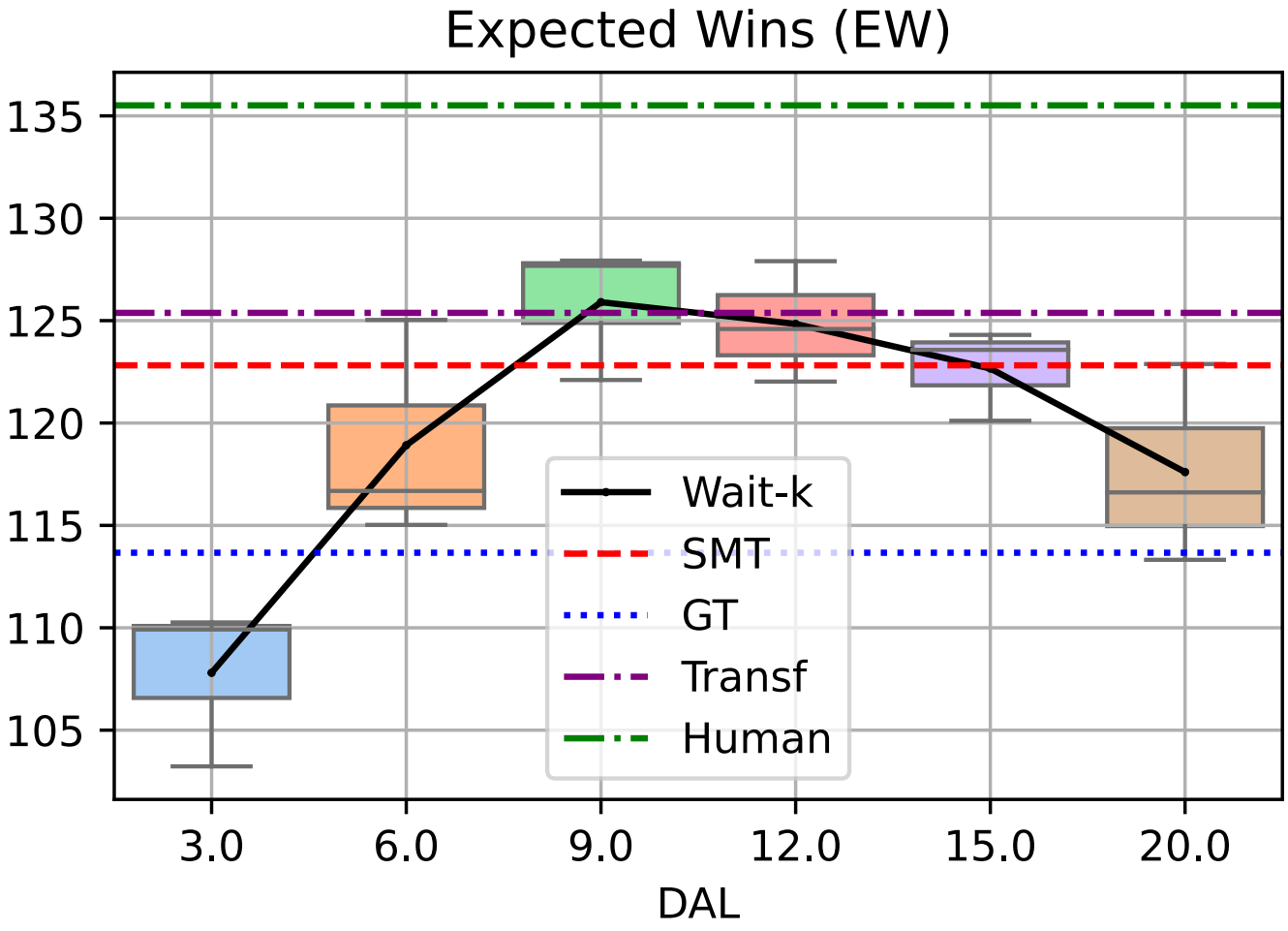
BLEU results for modern Simultaneous Translation Systems



Downstream QA Results

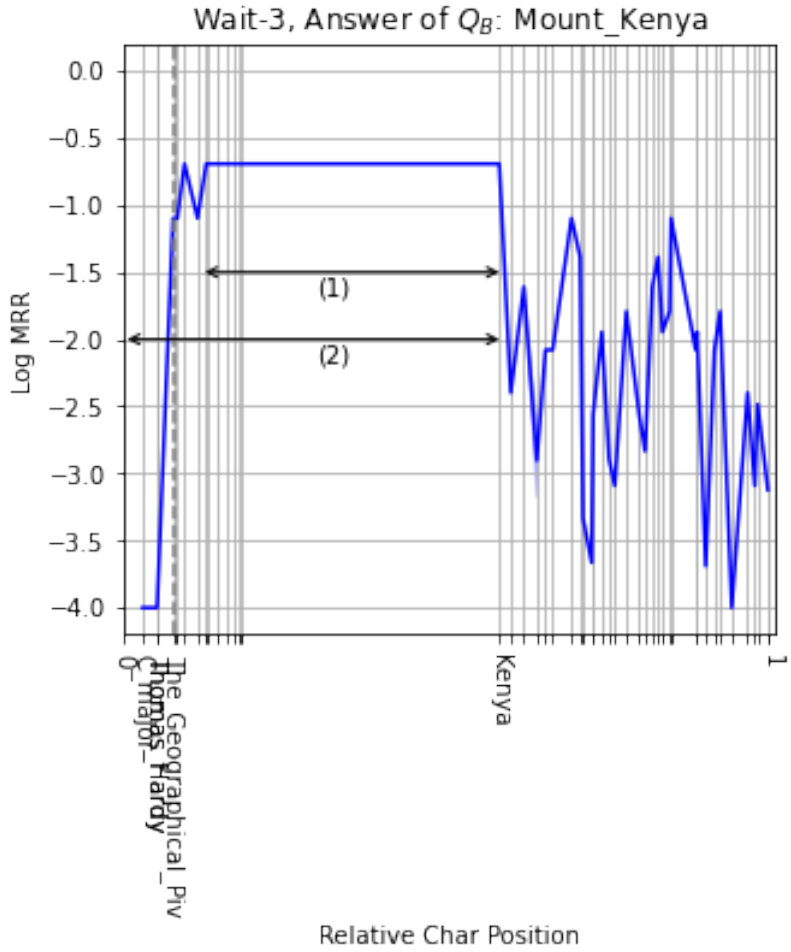


Downstream QA Results



Additional benefit: Only need to translate the answer

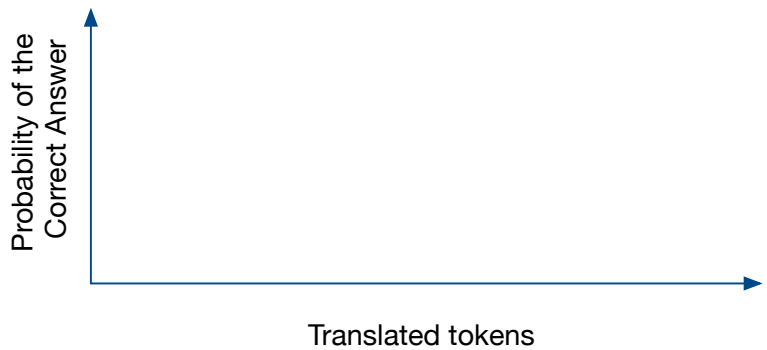
Undertranslation



When the translation doesn't help...

When are Mistakes / Hallucinations Harmful?

Question: Tę współrzędną wyznacza kąt dwuścienny między półpłaszczyzną południka zerowego a półpłaszczyzną południka …

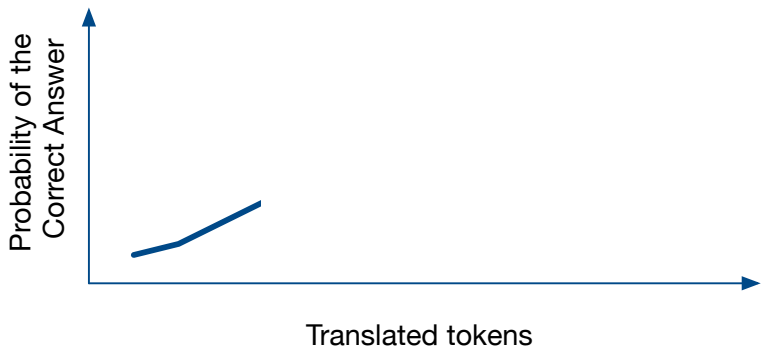


Guess: ??

When are Mistakes / Hallucinations Harmful?

This coordinate determines

Question: Tę współrzędną wyznacza kąt dwuścienny między półpłaszczyzną południka zerowego a półpłaszczyzną południka …

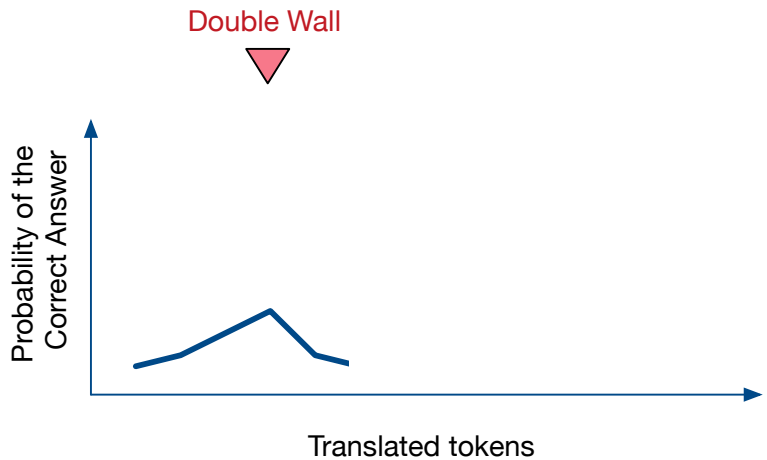


Guess: IP Address

When are Mistakes / Hallucinations Harmful?

This coordinate determines the double-wall

Question: Tę współrzędną wyznacza kąt dwuścienny między półpłaszczyzną południka zerowego a półpłaszczyzną południka …

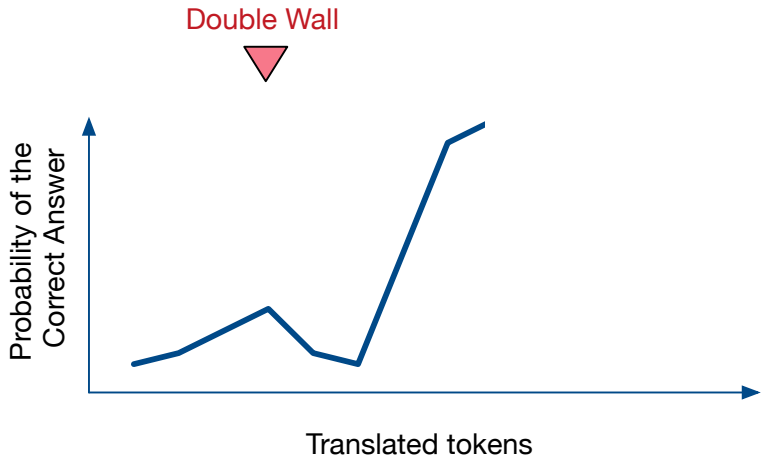


Guess: Spherical Coordinate

When are Mistakes / Hallucinations Harmful?

This coordinate determines the double-wall angle between the southern half of the meridian plane and the

Question: Tę współrzędną wyznacza kąt dwuścienny między półpłaszczyzną południka zerowego a półpłaszczyzną południka …

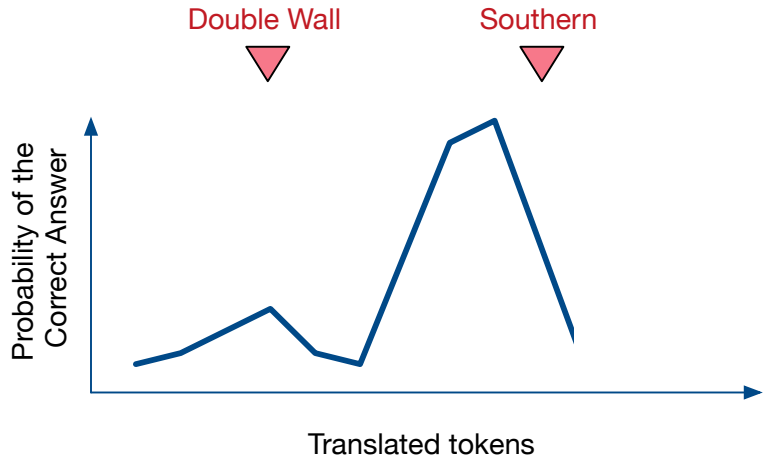


Guess: Longitude

When are Mistakes / Hallucinations Harmful?

This coordinate determines the double-wall angle between the southern half of the meridian plane and the southern

Question: Tę współrzędną wyznacza kąt dwuścienny między półpłaszczyzną południka zerowego a półpłaszczyzną południka …

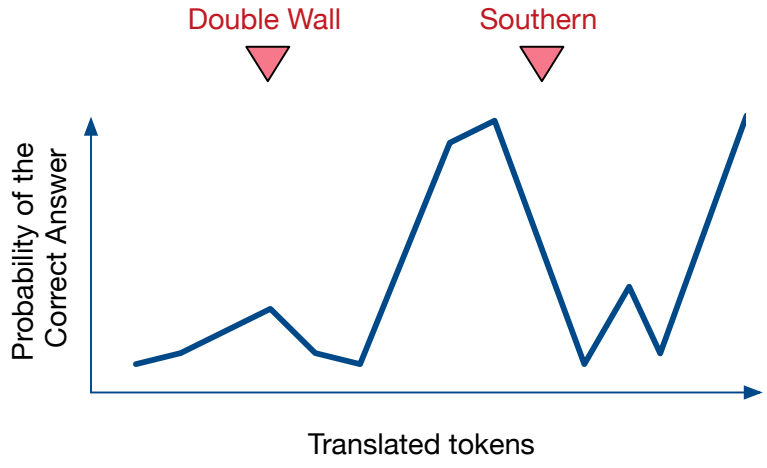


Guess: Spherical Coordinate

When are Mistakes / Hallucinations Harmful?

This coordinate determines the double-wall angle between the southern half of the meridian plane and the southern

Question: Tę współrzędną wyznacza kąt dwuścienny między półpłaszczyzną południka zerowego a półpłaszczyzną południka …



Guess: Longitude

Thanks

Collaborators

Hal Daumé III (UMD), David Blei (Columbia), Marine Carpuat (UMD), Leah Findlater (UW), Kevin Seppi (BYU), Eric Ringger (BYU)

Funders



IARPA

Supporters

International Academic Competitions

NAQT℠

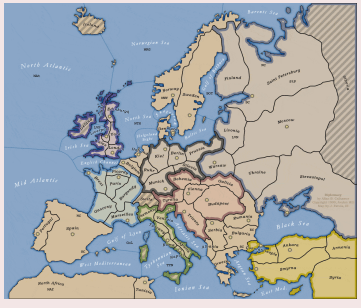
National Academic Quiz Tournaments, LLC

But wait, there's more!

Computational Social Science

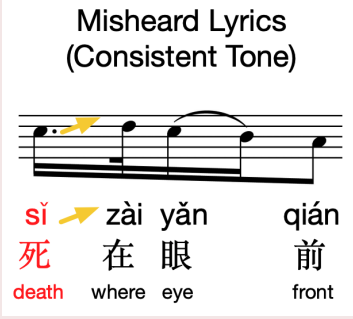
[?, ?]

Detecting Deception



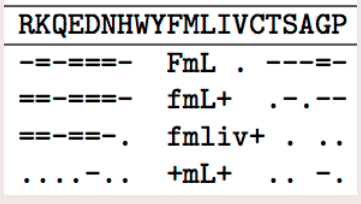
[?, ?]

Multilingual/Multicultural Models



[?, ?]

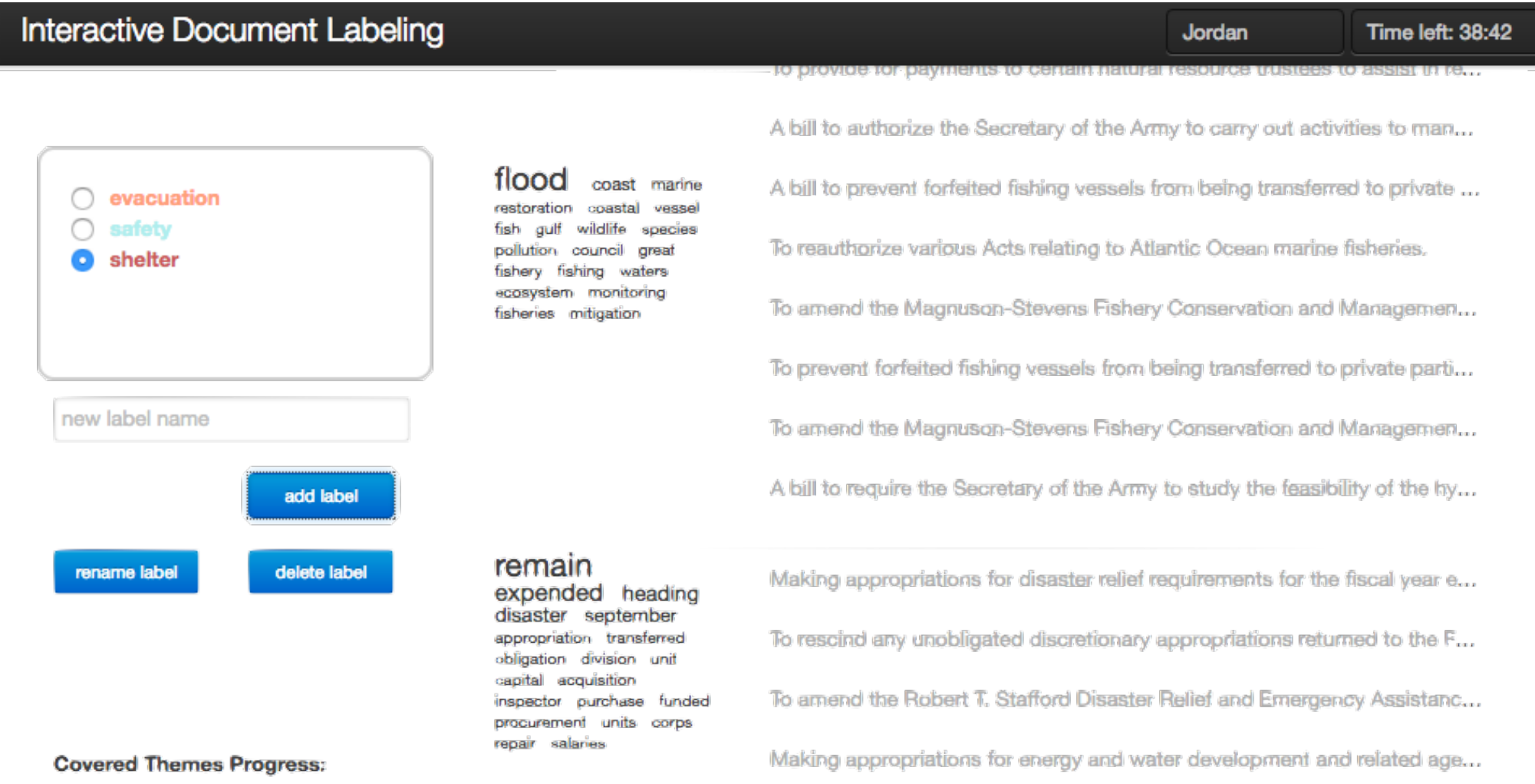
Computational Biology

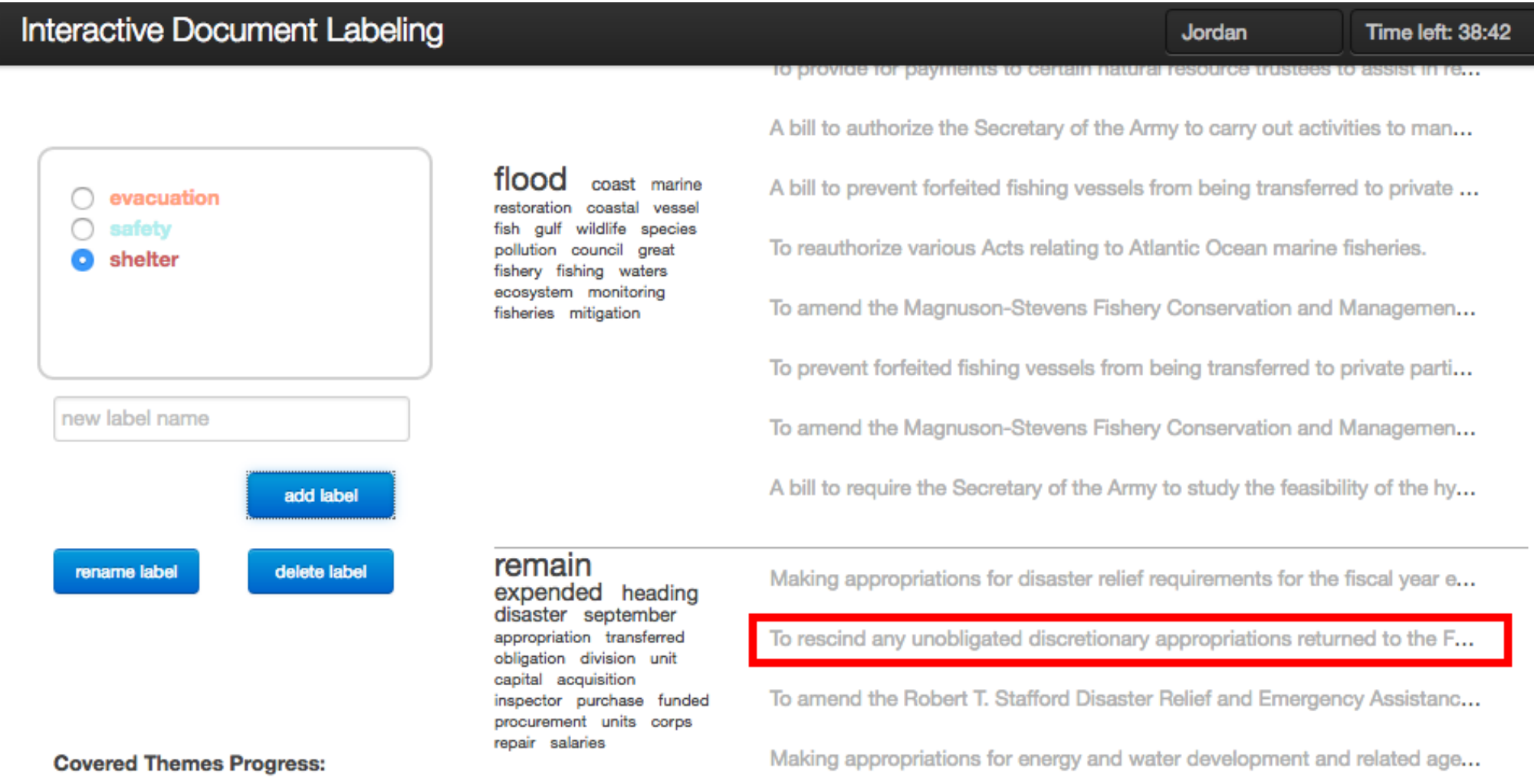


[?, ?]

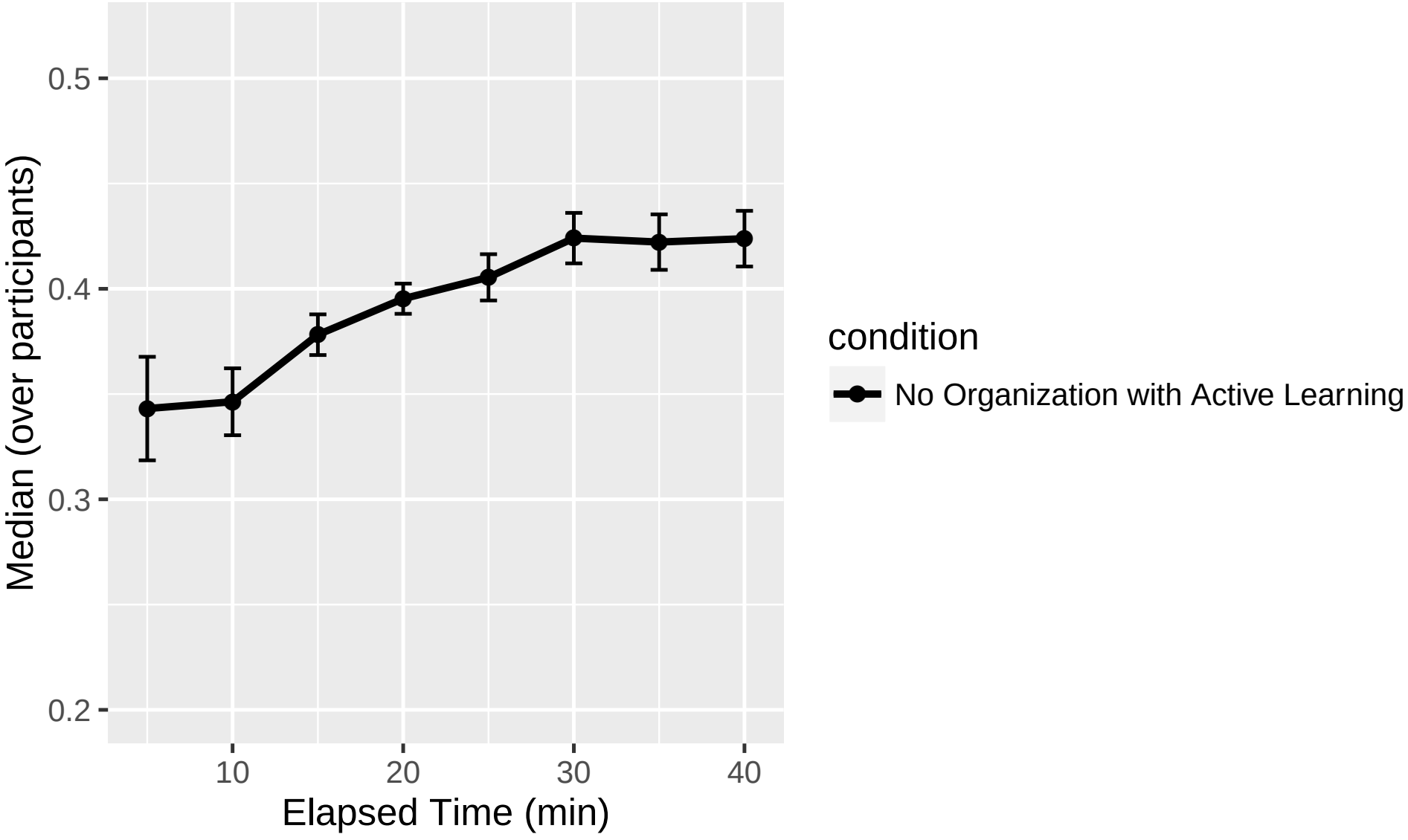
ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi.
Association for Computational Linguistics, 2016.

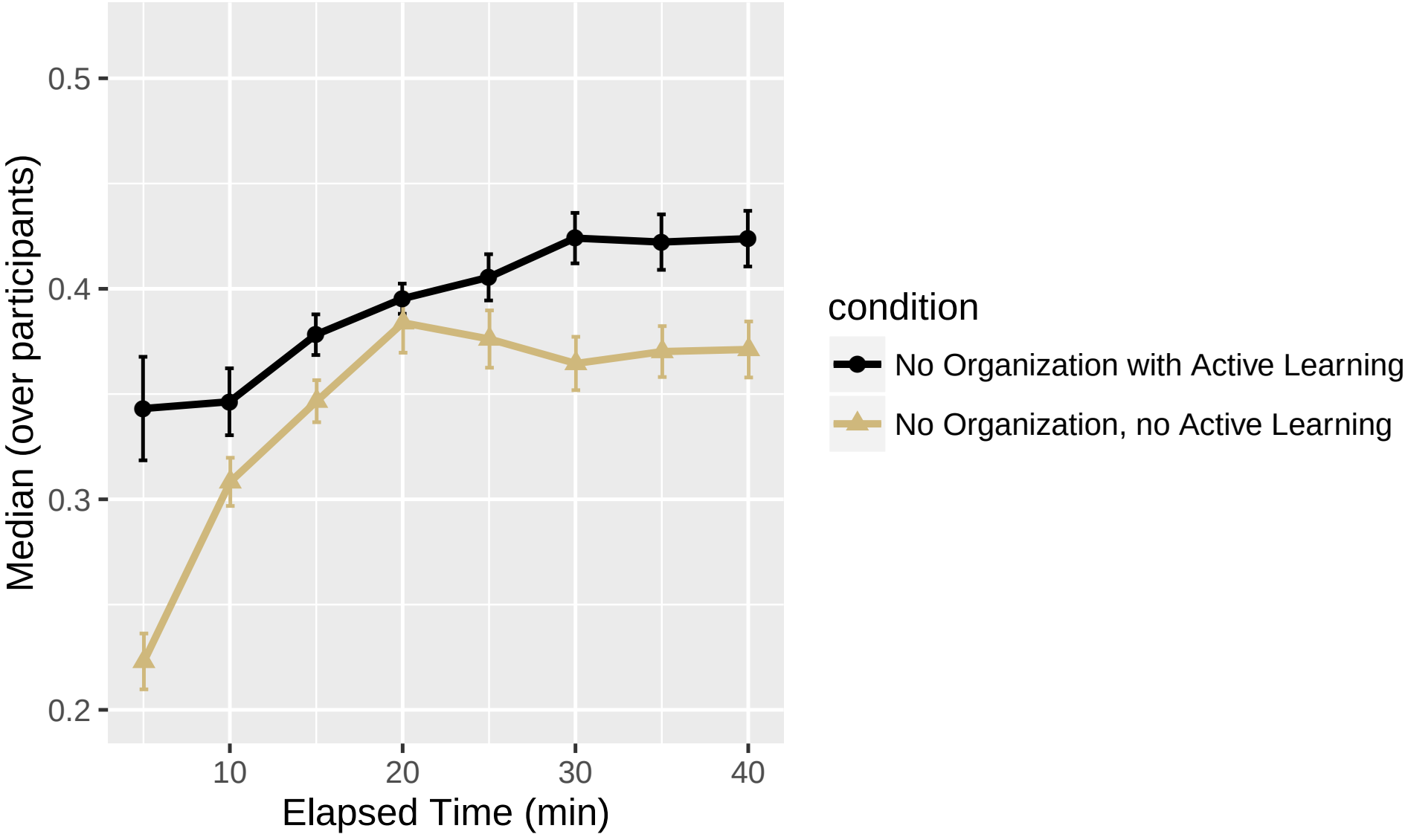

Interactive Document Labeling
Jordan
Time left: 38:42
evacuation
safety
shelter
new label name
add label
rename label
delete label
Covered Themes Progress:
flood coast marine restoration coastal vessel fish gulf wildlife species pollution council great fishery fishing waters ecosystem monitoring fisheries mitigation
remain expended heading disaster september appropriation transferred obligation division unit capital acquisition inspector purchase funded procurement units corps repair salaries
A bill to authorize the Secretary of the Army to carry out activities to man...
A bill to prevent forfeited fishing vessels from being transferred to private ...
To reauthorize various Acts relating to Atlantic Ocean marine fisheries.
To amend the Magnuson-Stevens Fishery Conservation and Managemen...
To prevent forfeited fishing vessels from being transferred to private parti...
To amend the Magnuson-Stevens Fishery Conservation and Managemen...
A bill to require the Secretary of the Army to study the feasibility of the hy...
Making appropriations for disaster relief requirements for the fiscal year e...
To rescind any unobligated discretionary appropriations returned to the F...
To amend the Robert T. Stafford Disaster Relief and Emergency Assistanc...
Making appropriations for energy and water development and related age...



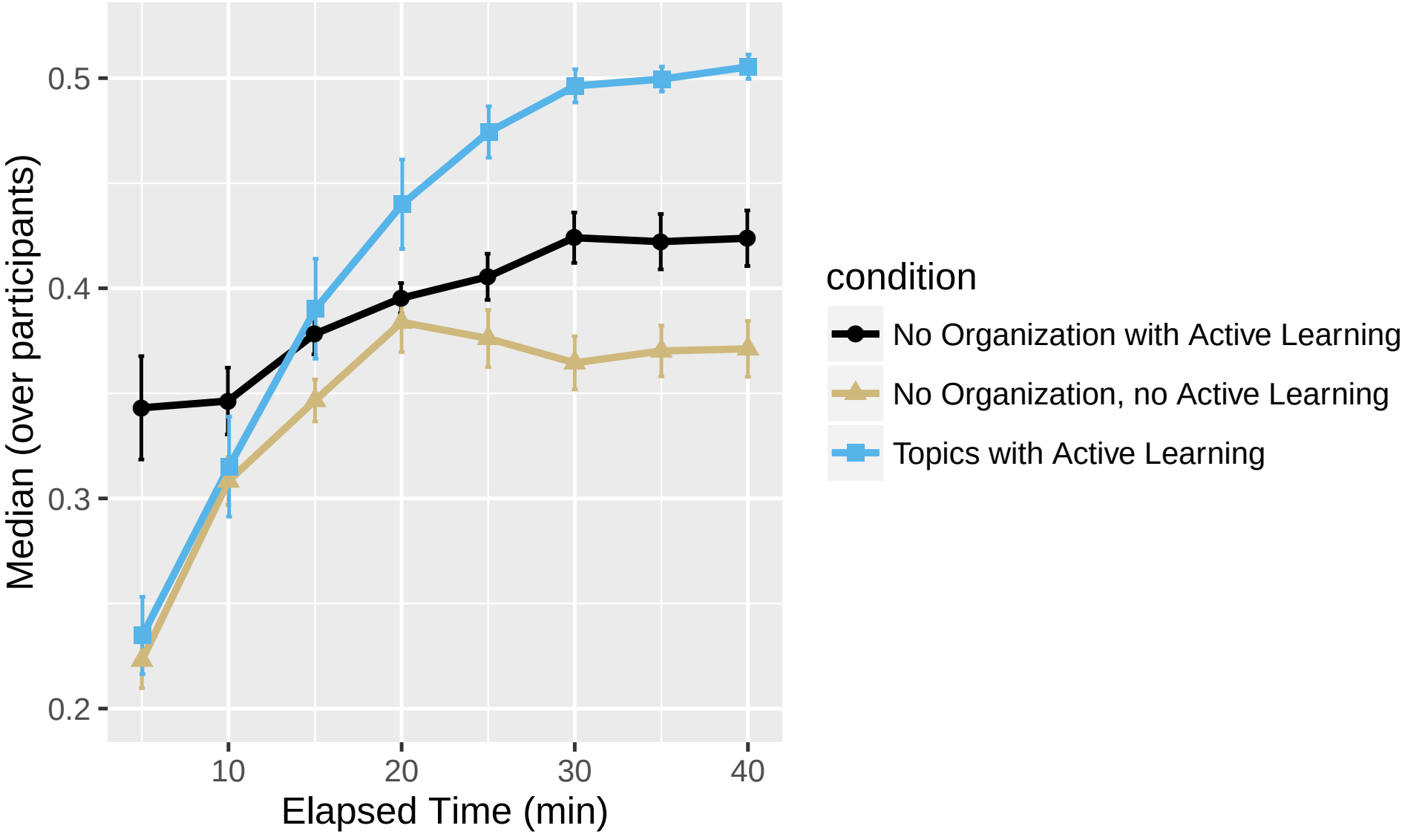
Direct users to document



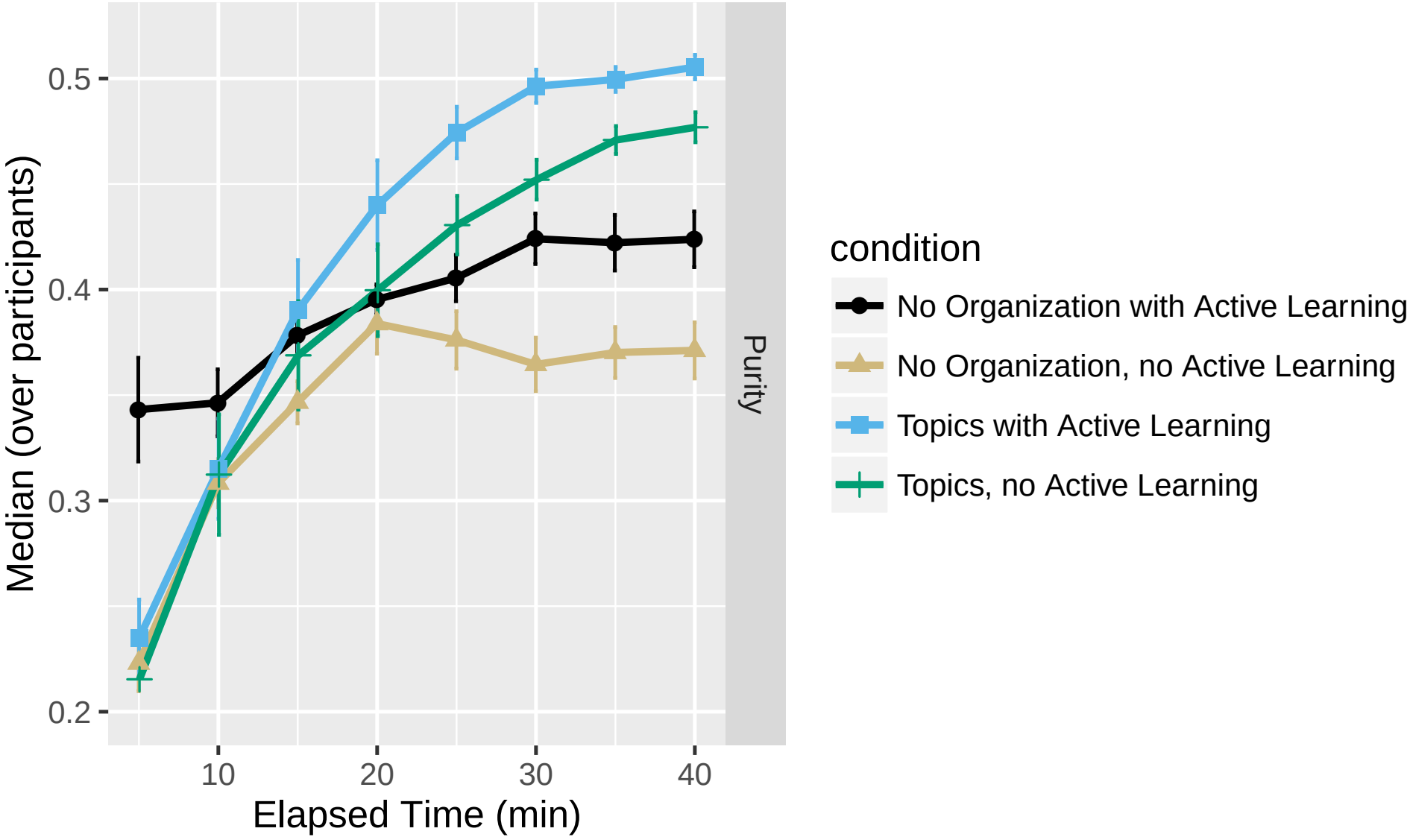
Active learning if time is short



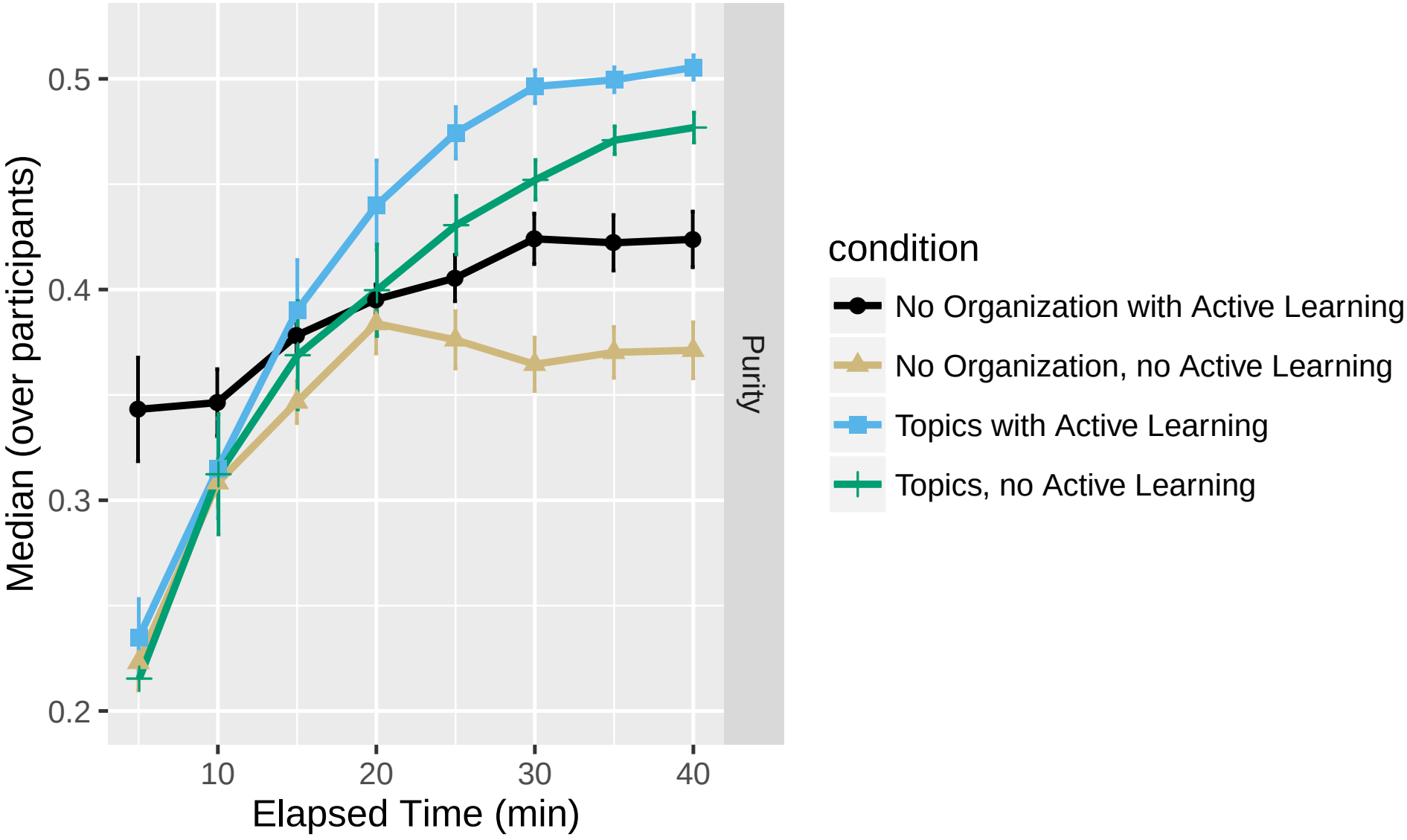
Better than status quo



Active learning can help topic models



Topic models help users understand the collection



Moral: machines and humans together (if you let them)

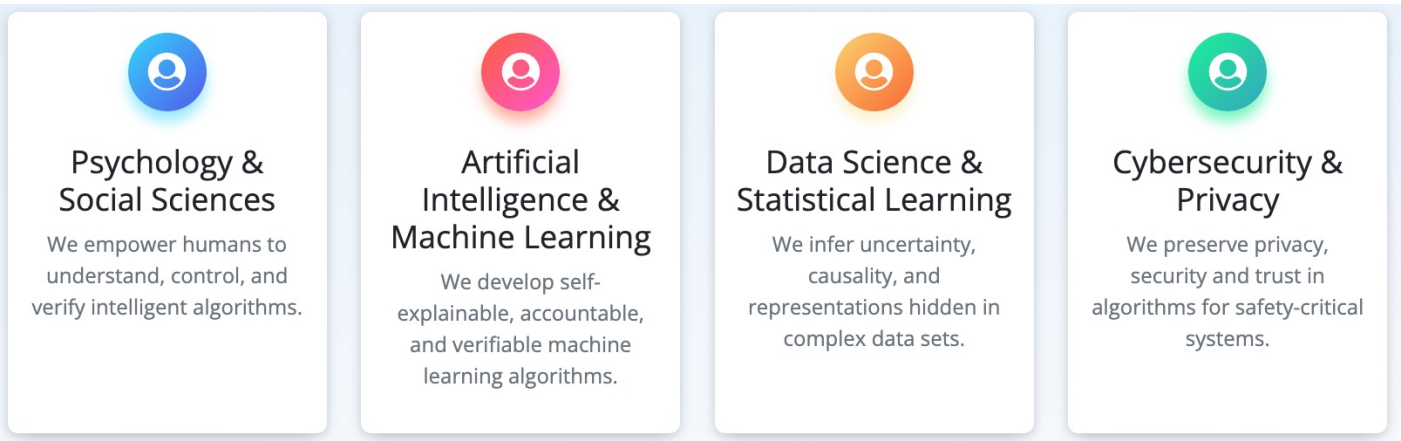
After writing ~The Theory of Moral

Prediction	Evidence
Adam_Smith	2.1049
Samuel_Johnson	1.9459
Hannah_Arendt	1.7826
Erving_Goffman	1.6882
Kenneth_Arrow	1.627
David_Hume	1.6114
John_Dewey	1.5881
Immanuel_Kant	1.5627
Bertrand_Russell	1.5434

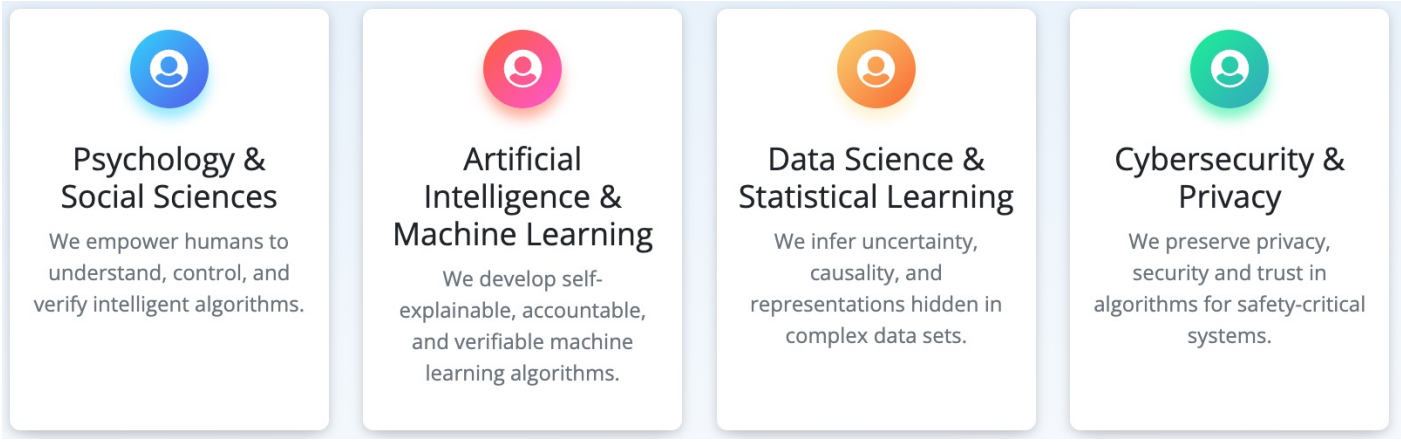
Andrea Lin

References

RC TRUST


Psychology & Social Sciences
We empower humans to understand, control, and verify intelligent algorithms.
Artificial Intelligence & Machine Learning
We develop self-explainable, accountable, and verifiable machine learning algorithms.
Data Science & Statistical Learning
We infer uncertainty, causality, and representations hidden in complex data sets.
Cybersecurity & Privacy
We preserve privacy, security and trust in algorithms for safety-critical systems.

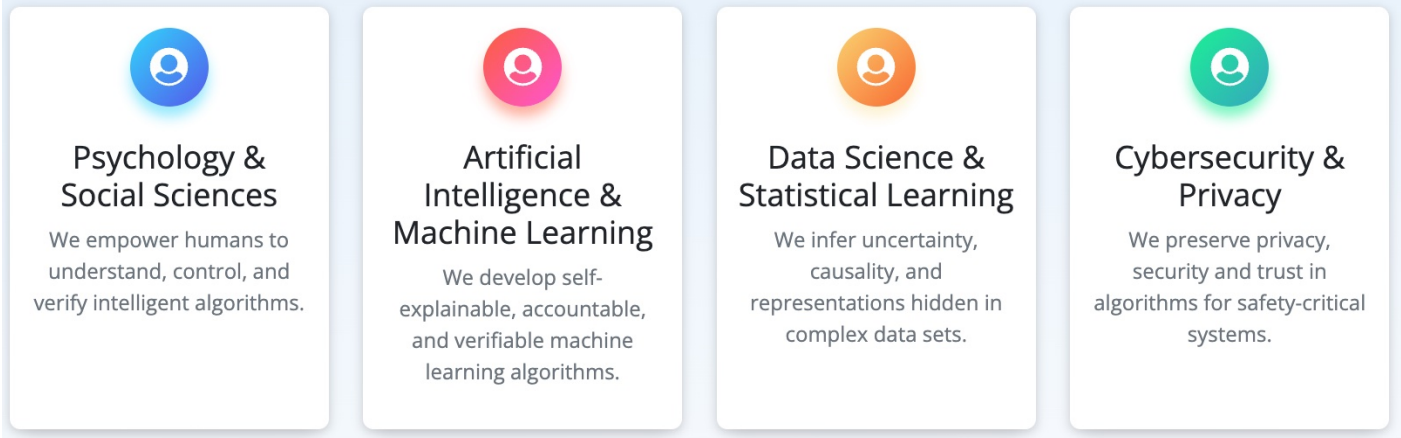
RC TRUST



Psychology & Social Sciences

- Method: Item Response Theory
- Method: Ideal Point Models
- Application: Multilingual and Multicultural Models of Persuasion and Alliance Building
- Collaboration: Bundesverfassungsgericht Entscheidungen

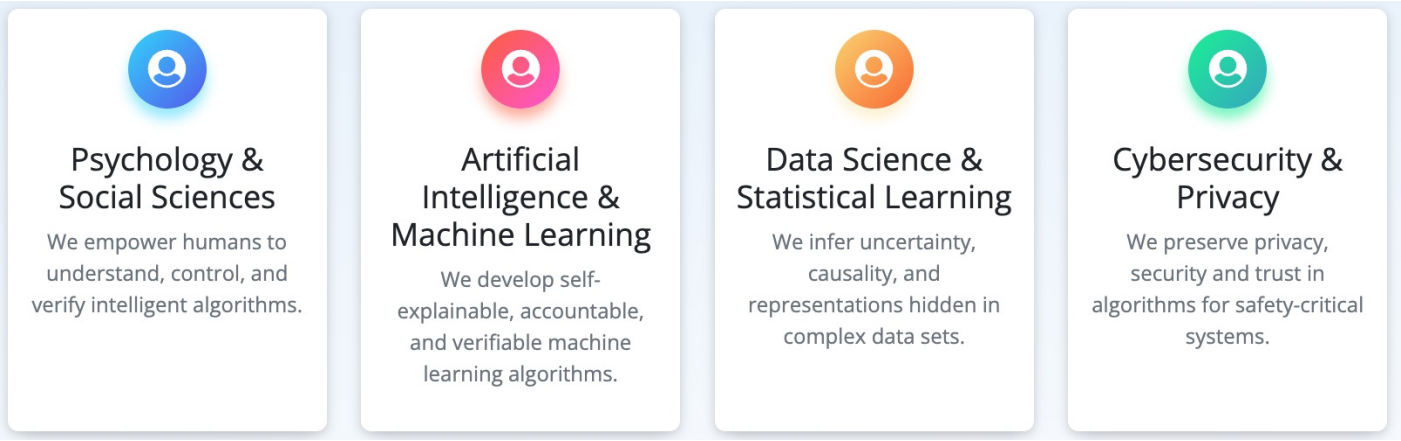
RC TRUST



AI & ML

- Method: Reinforcement Learning
- Method: Neural Text Similarity
- Application: Training QA Retrieval Mechanisms
- Collaboration: Improving Google QA

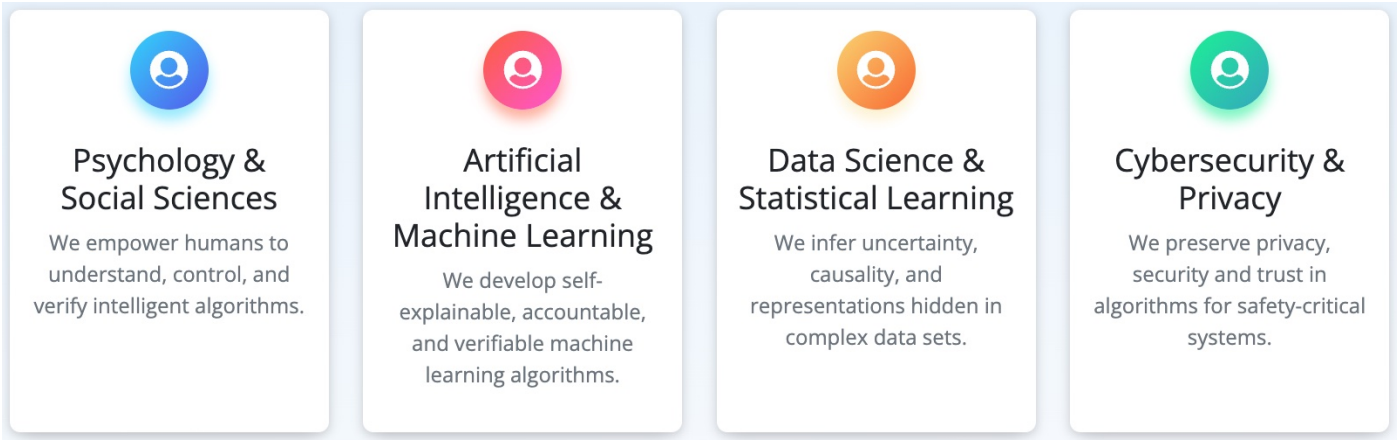
RC TRUST



Data Science & Statistical Learning

- Method: Bayesian Nonparametrics
- Method: Interactive Dirichlet Forest Priors
- Application: Sensemaking
- Collaboration: Monitoring Local Reseliency

RC TRUST



Cybersecurity and Privacy

- Method: Adversarial Example Construction
- Method: Disinformation Gamification
- Application: Fake News Detection
- Collaboration: Climate Fact Checking