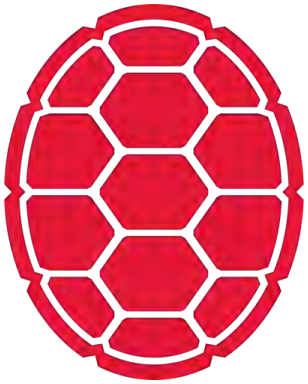


Language Games: Sure, ask AI who wrote *Cogwheels*, but don't use it for relationship advice

Jordan Boyd-Graber
University of Maryland
2025

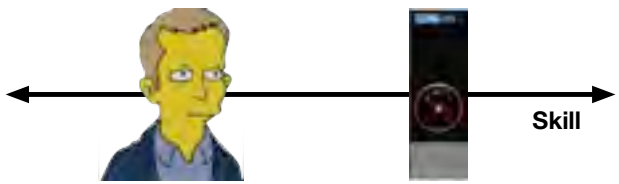
Why Language Games?

- 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

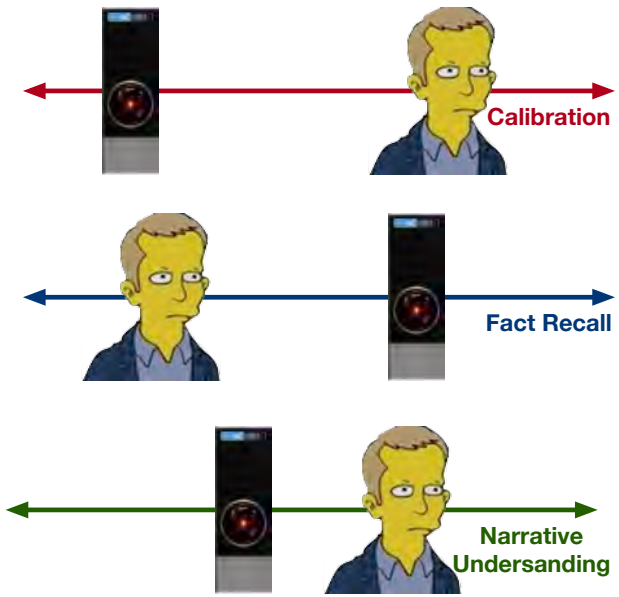
Image: DeepMind

Peter Morgan/Reuters

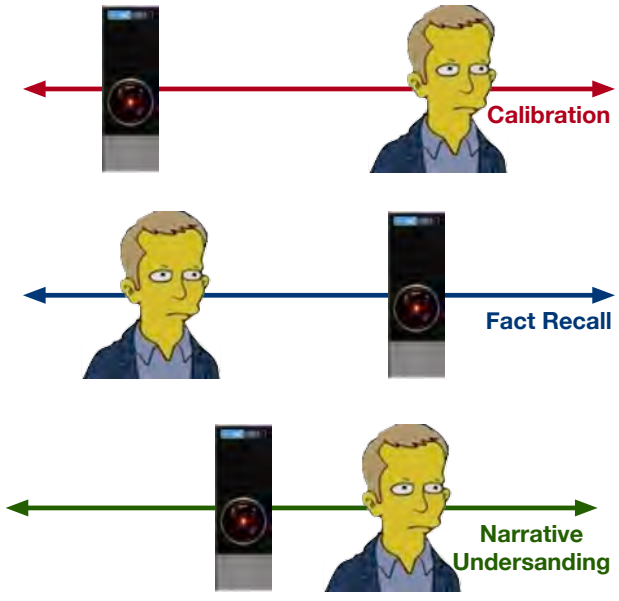
Measuring Skill



Measuring Skill



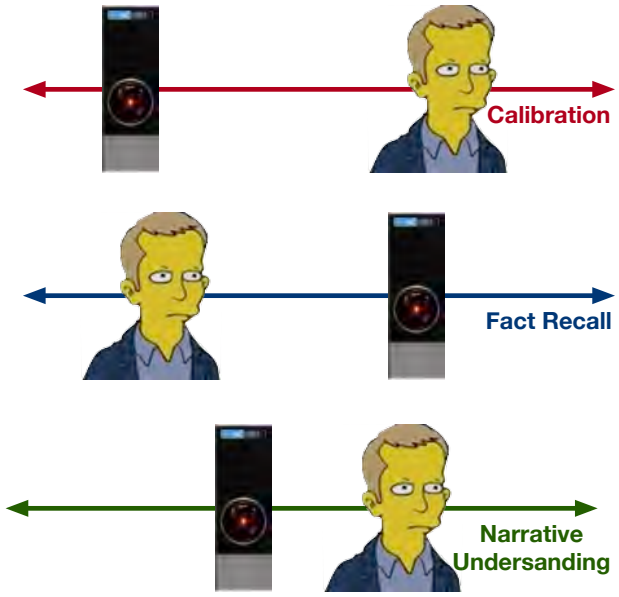
Measuring Skill



Augmentation

How much can we *increase* the skill of a human in a language game with computational (i.e., AI) support?

Measuring Skill



Augmentation

How much can we *increase* the skill of a human in a language game with computational (i.e., AI) support?

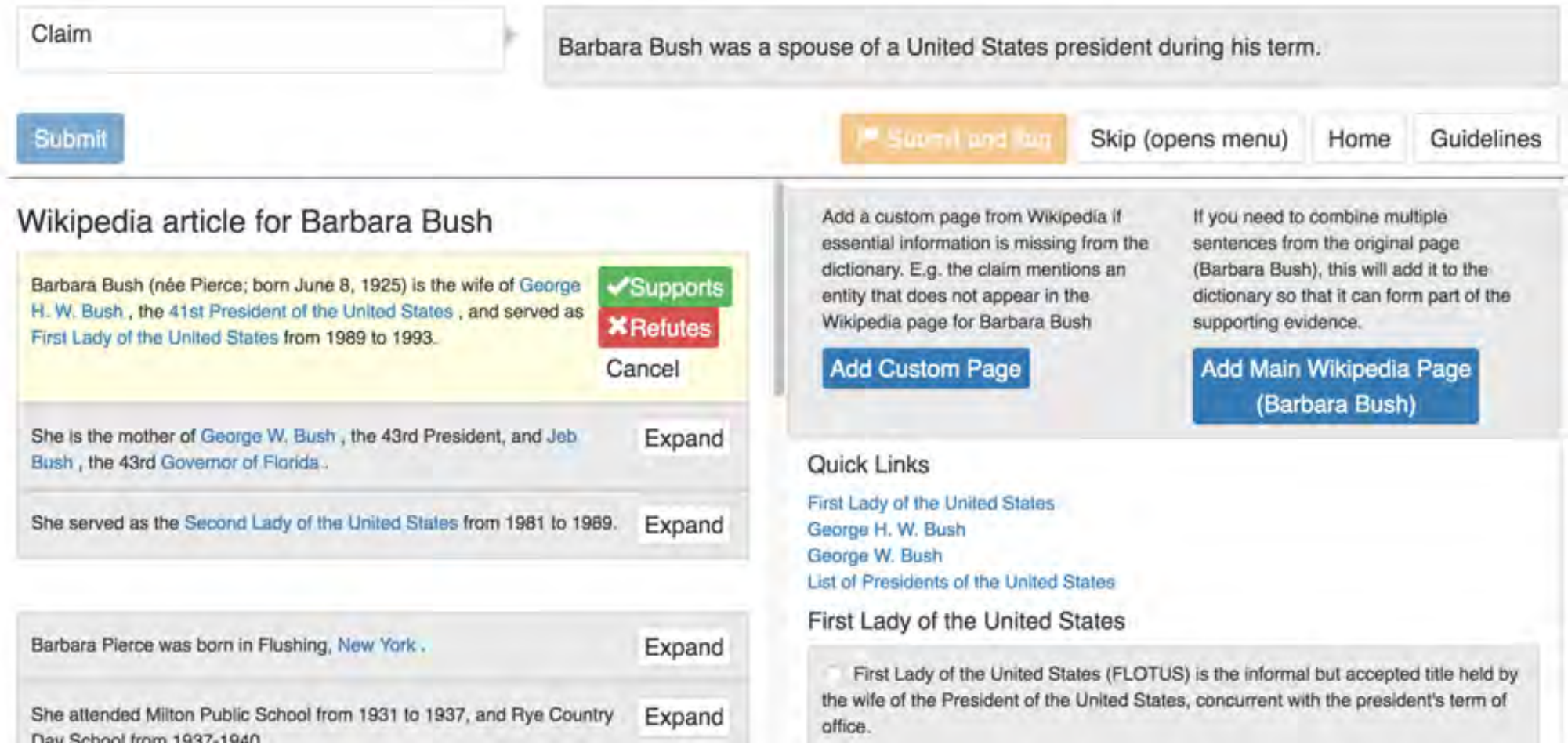
1. Fact Checking
2. Question Answering
3. Negotiation

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

Can we help users detect incorrect statements?



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)

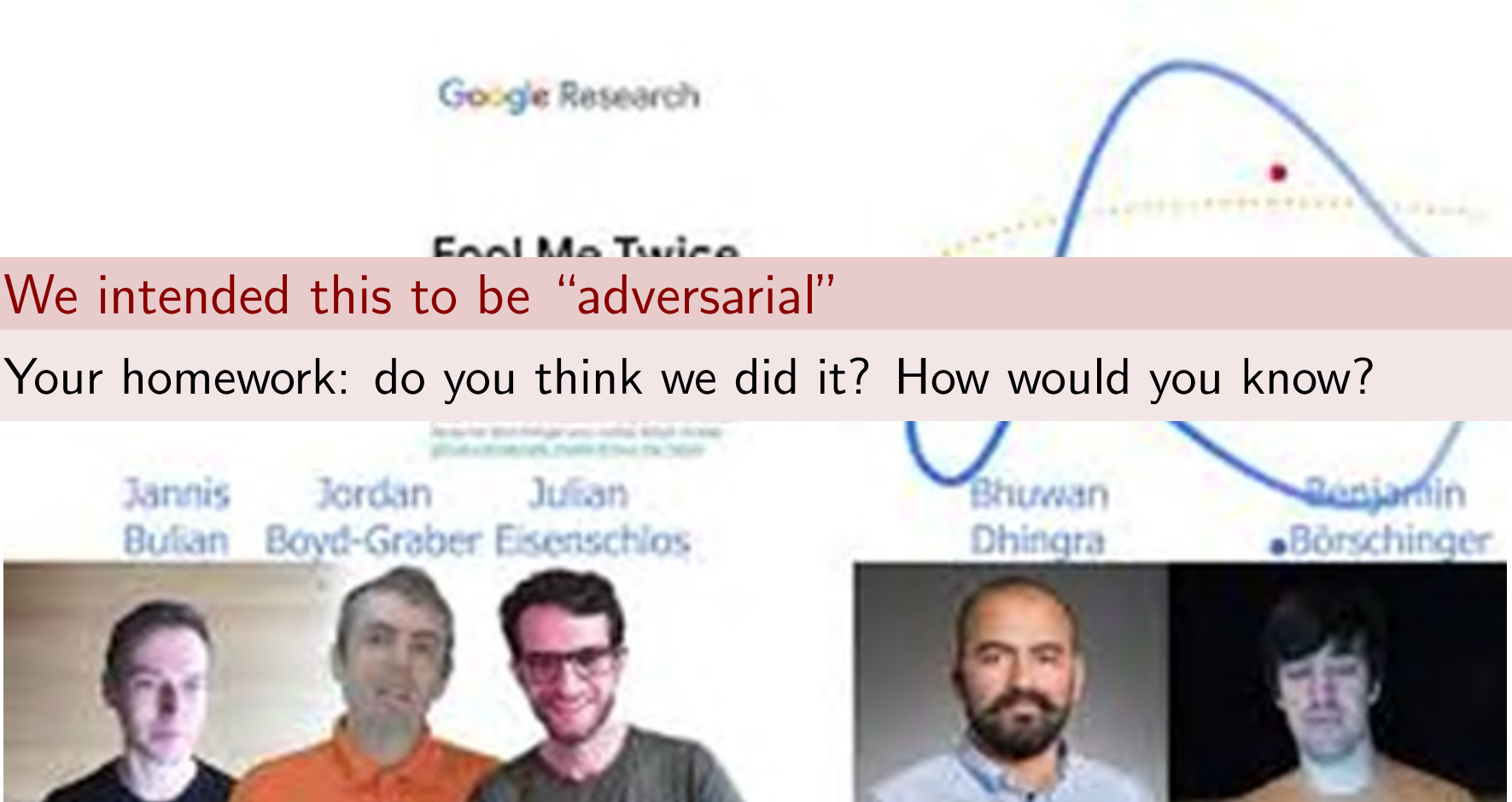
FEVER categories

- 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

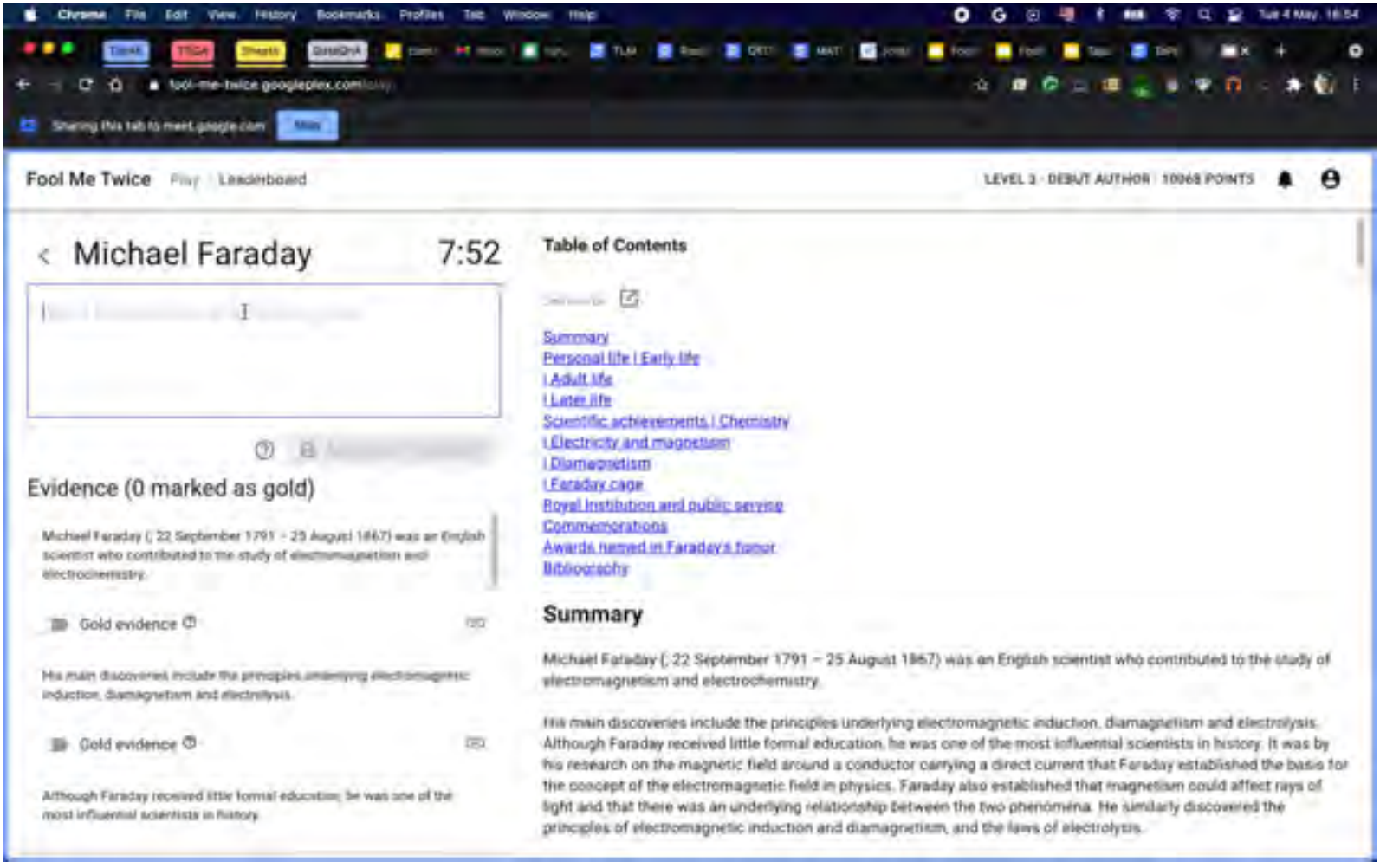
Fool Me Twice



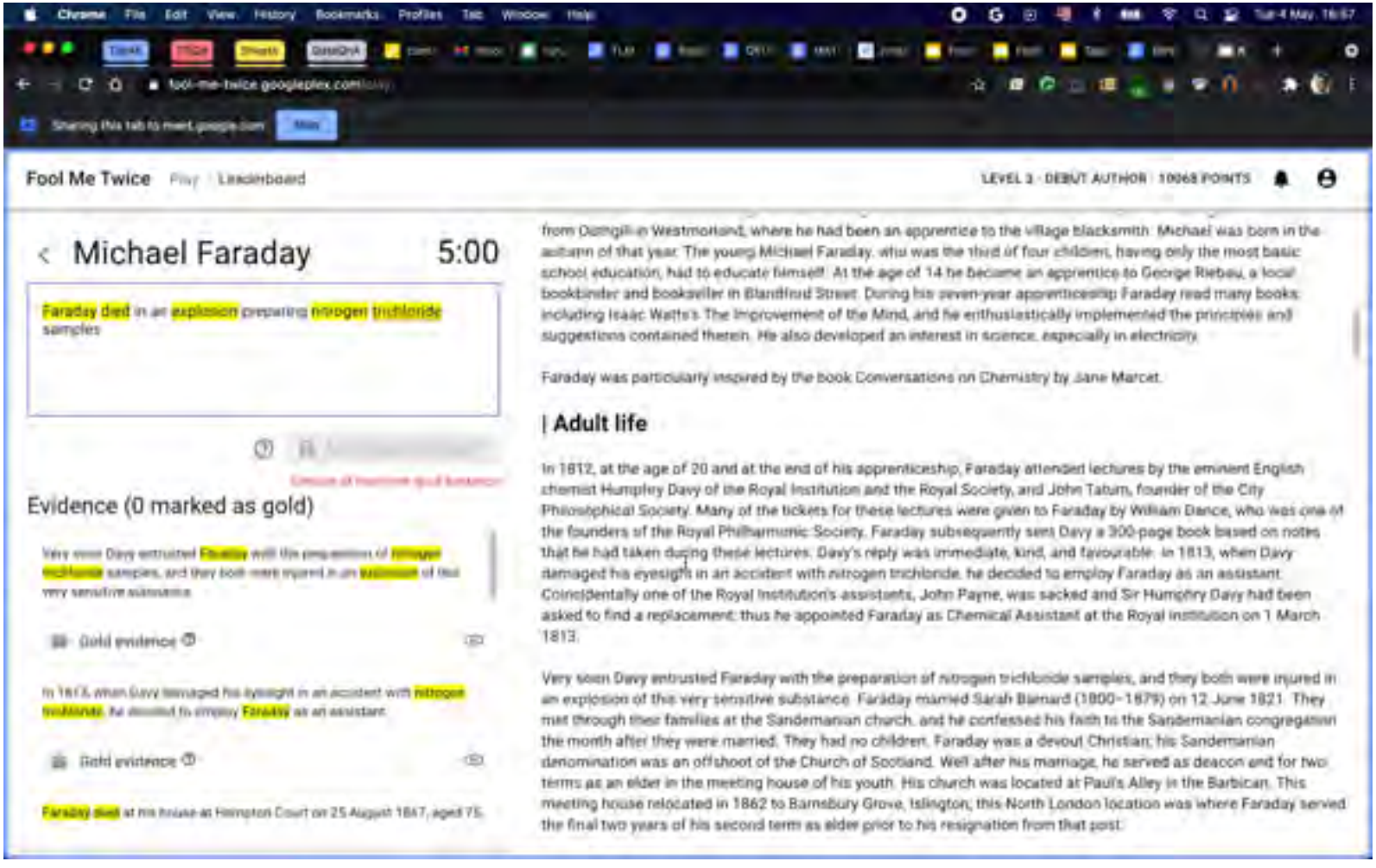
We intended this to be "adversarial"

Your homework: do you think we did it? How would you know?

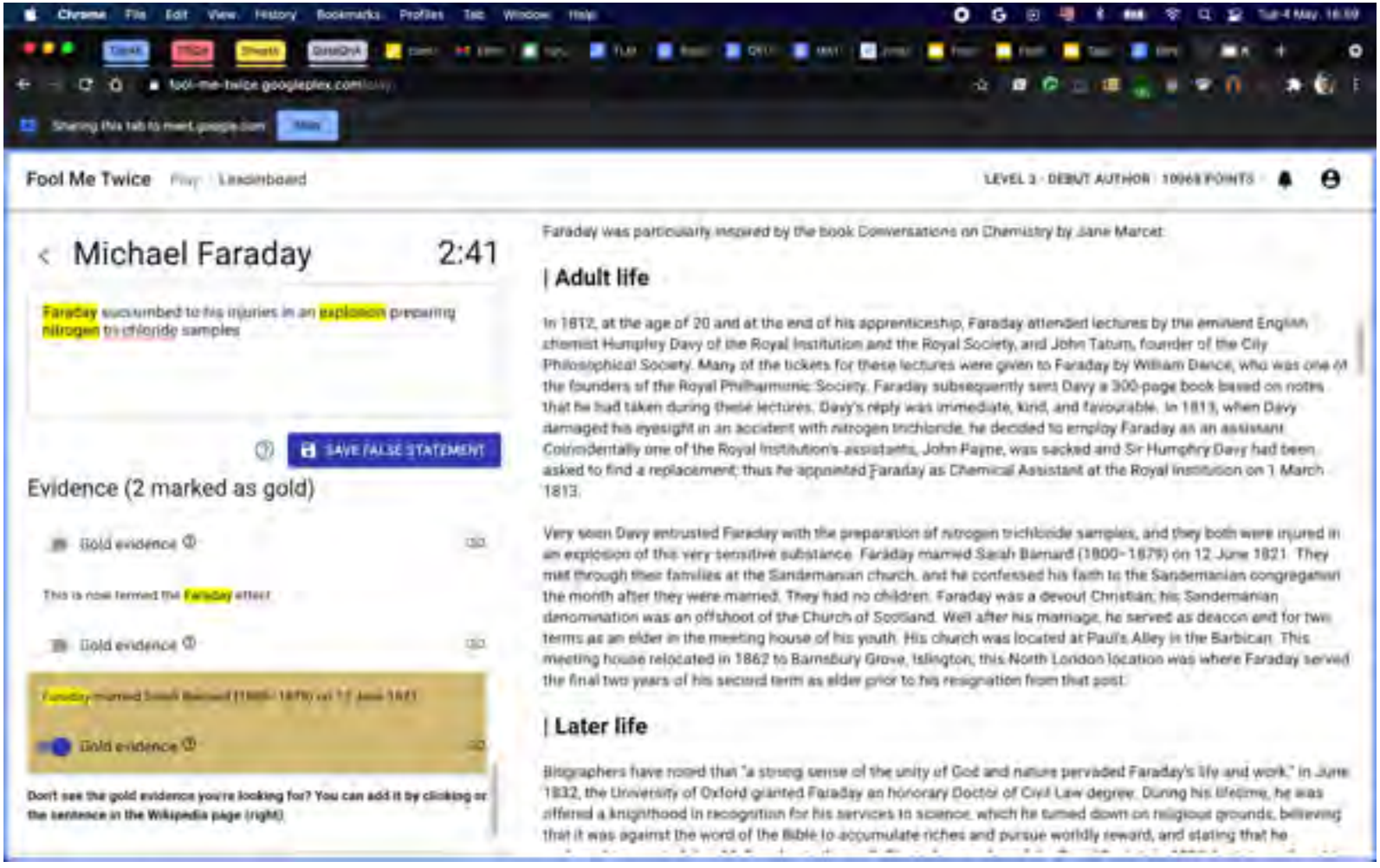
Author Claims where Gold Evidence is not Model's Top Evidence



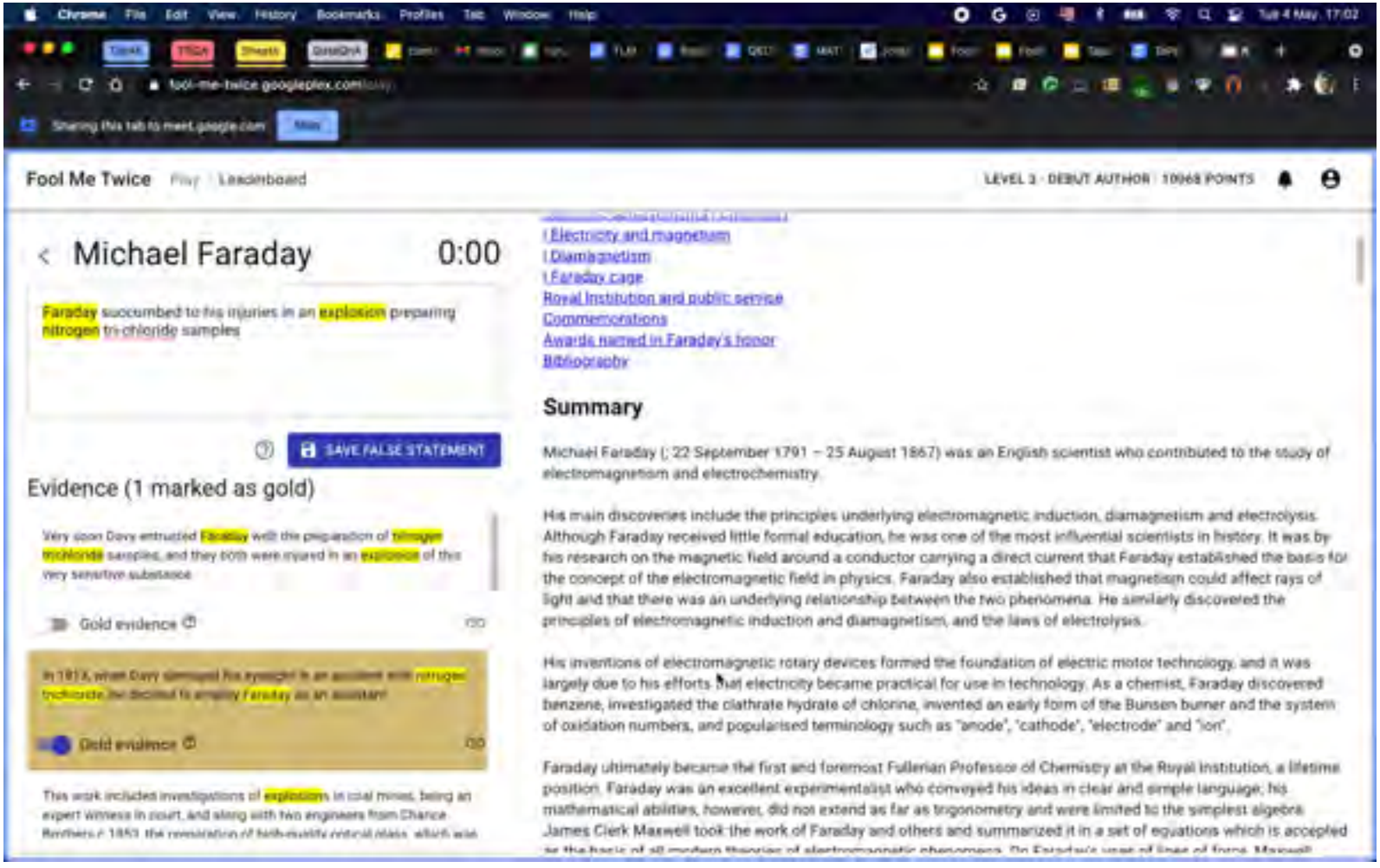
Author Claims where Gold Evidence is not Model's Top Evidence



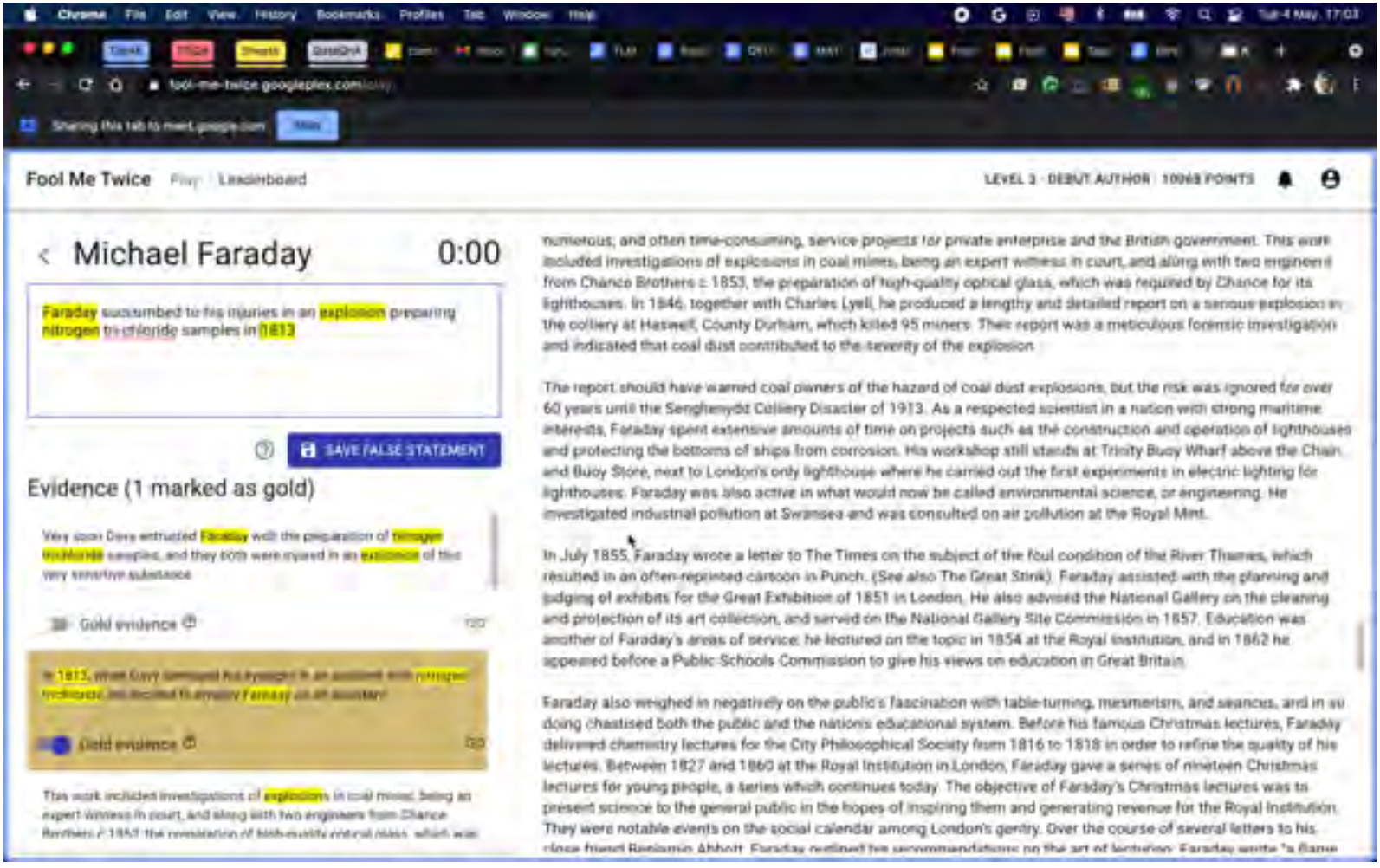
Author Claims where Gold Evidence is not Model's Top Evidence



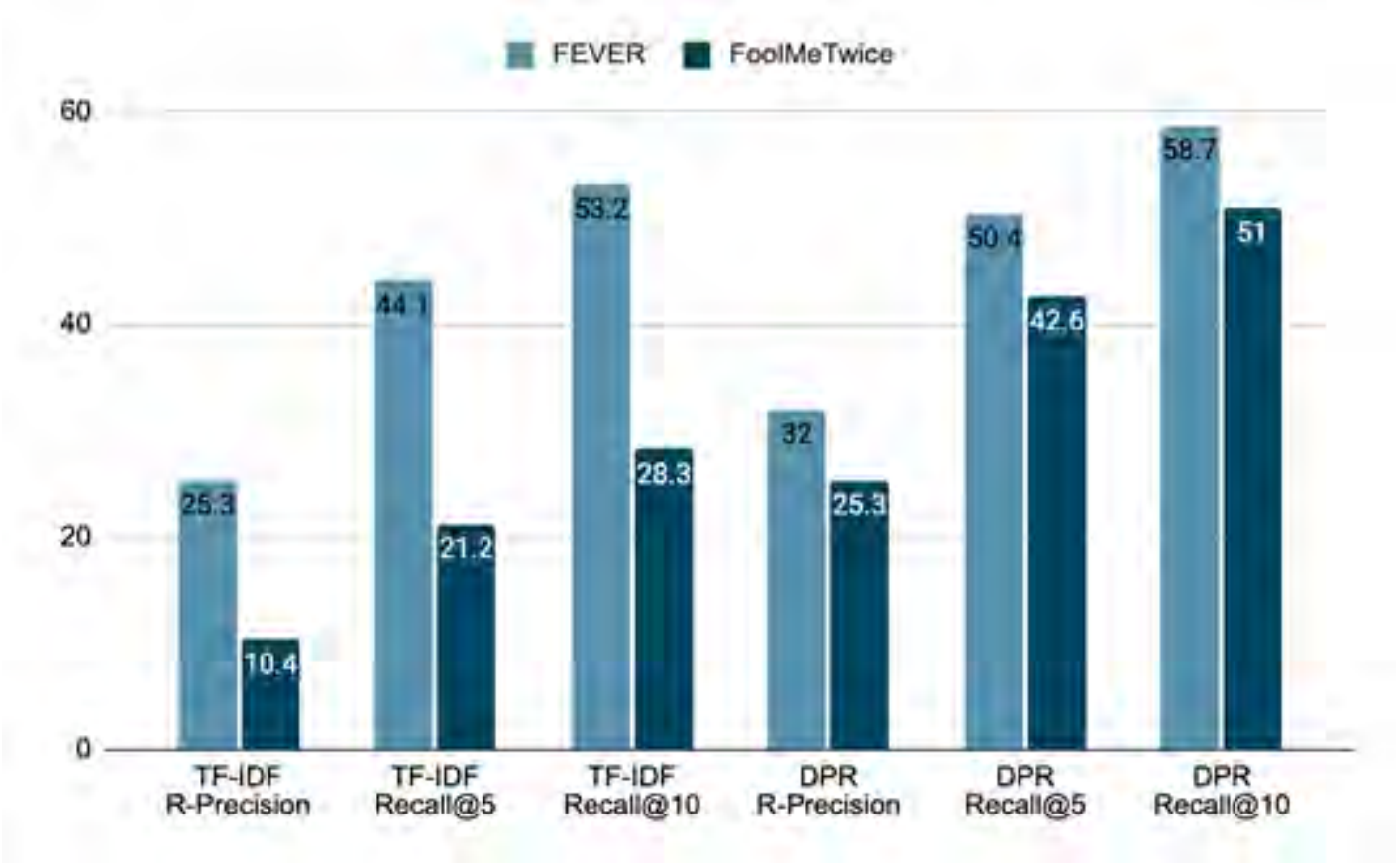
Author Claims where Gold Evidence is not Model's Top Evidence



Author Claims where Gold Evidence is not Model's Top Evidence



It is Harder (a bit)



Claim: Manhattan has the most residents of New York City's boroughs.

Claim: Manhattan has the most residents of New York City's boroughs.

IR

New York City's borough of Manhattan is the highest nominal income county in the United States. In particular, ZIP code 10021 on Manhattan's Upper East Side, with more than 100,000 inhabitants and a per capita income of over $90,000, has one of the largest concentrations of income in the United States. The other boroughs, especially Queens and Staten Island, have large middle-class populations.

Claim: Manhattan has the most residents of New York City's boroughs.

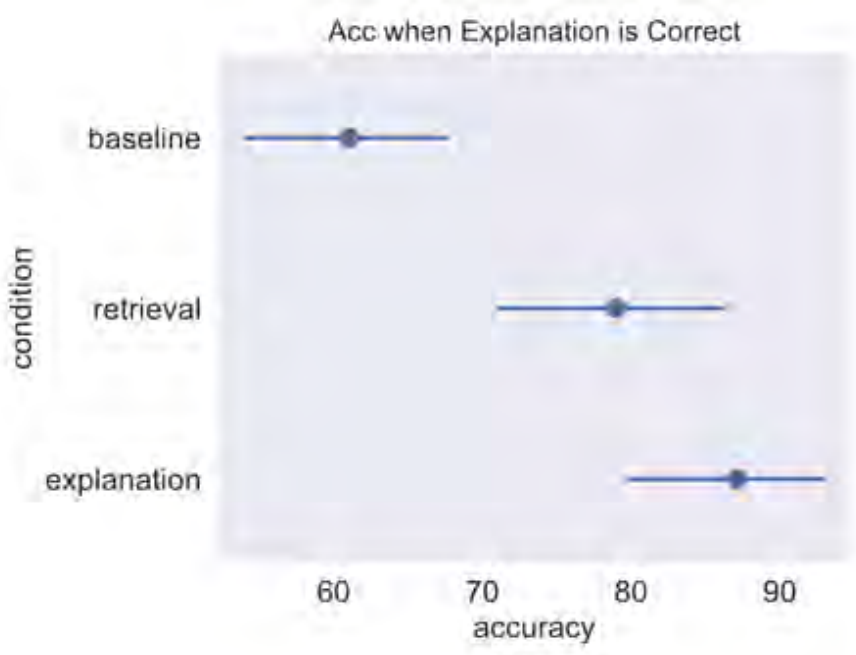
Direct

The claim that Manhattan has the most residents of New York City's boroughs is false. Manhattan has a population of approximately 1.6 million residents. However, New York City as a whole has a population of approximately 8.3 million residents. Therefore, Manhattan does not have the most residents among New York City's boroughs.

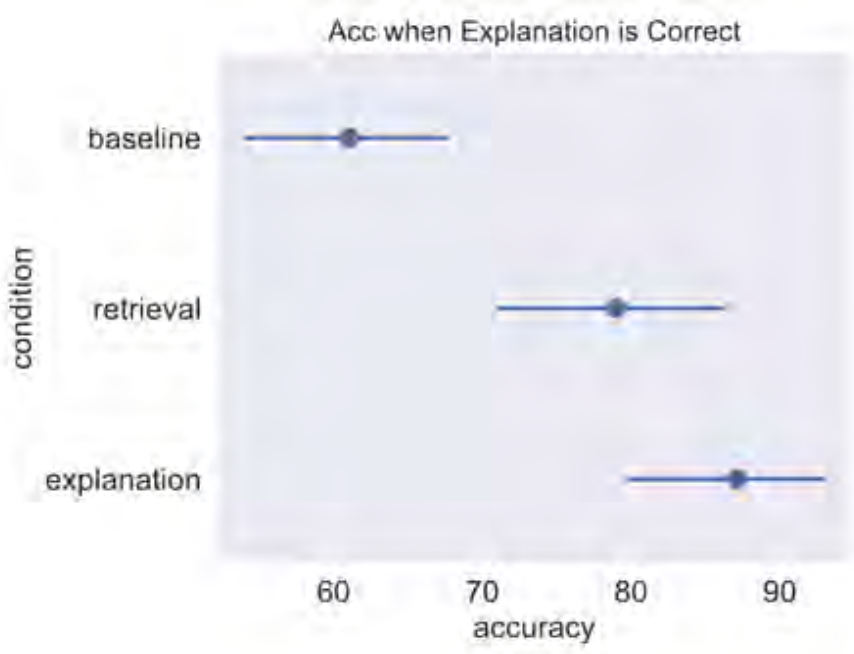
IR

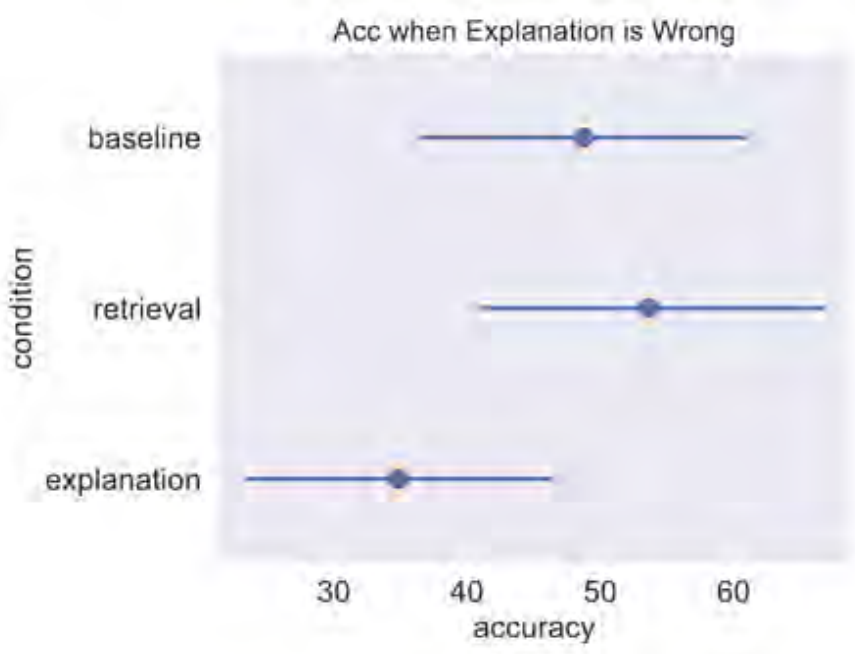
New York City's borough of Manhattan is the highest nominal income county in the United States. In particular, ZIP code 10021 on Manhattan's Upper East Side, with more than 100,000 inhabitants and a per capita income of over $90,000, has one of the largest concentrations of income in the United States. The other boroughs, especially Queens and Staten Island, have large middle-class populations.

Results


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

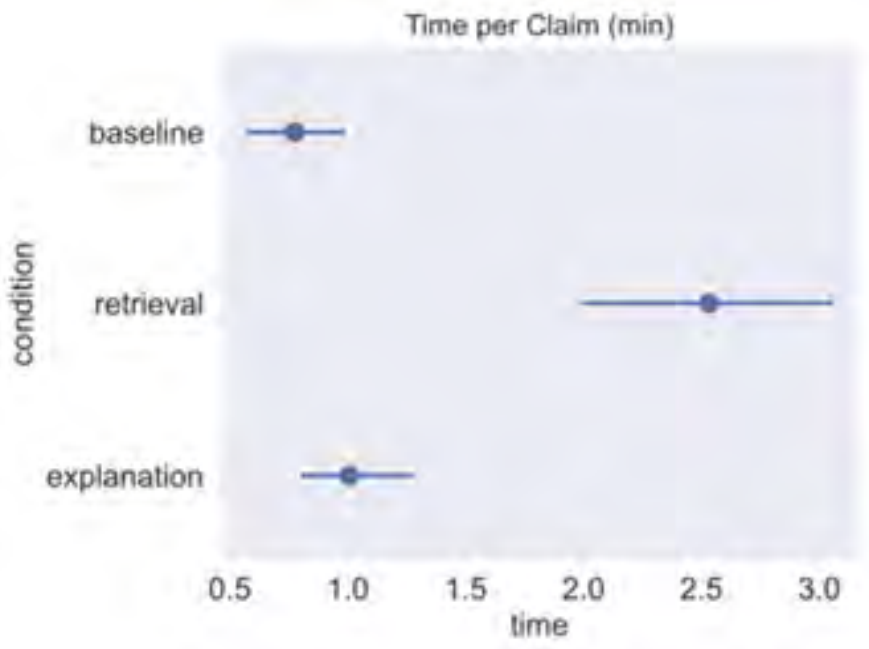
Results


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

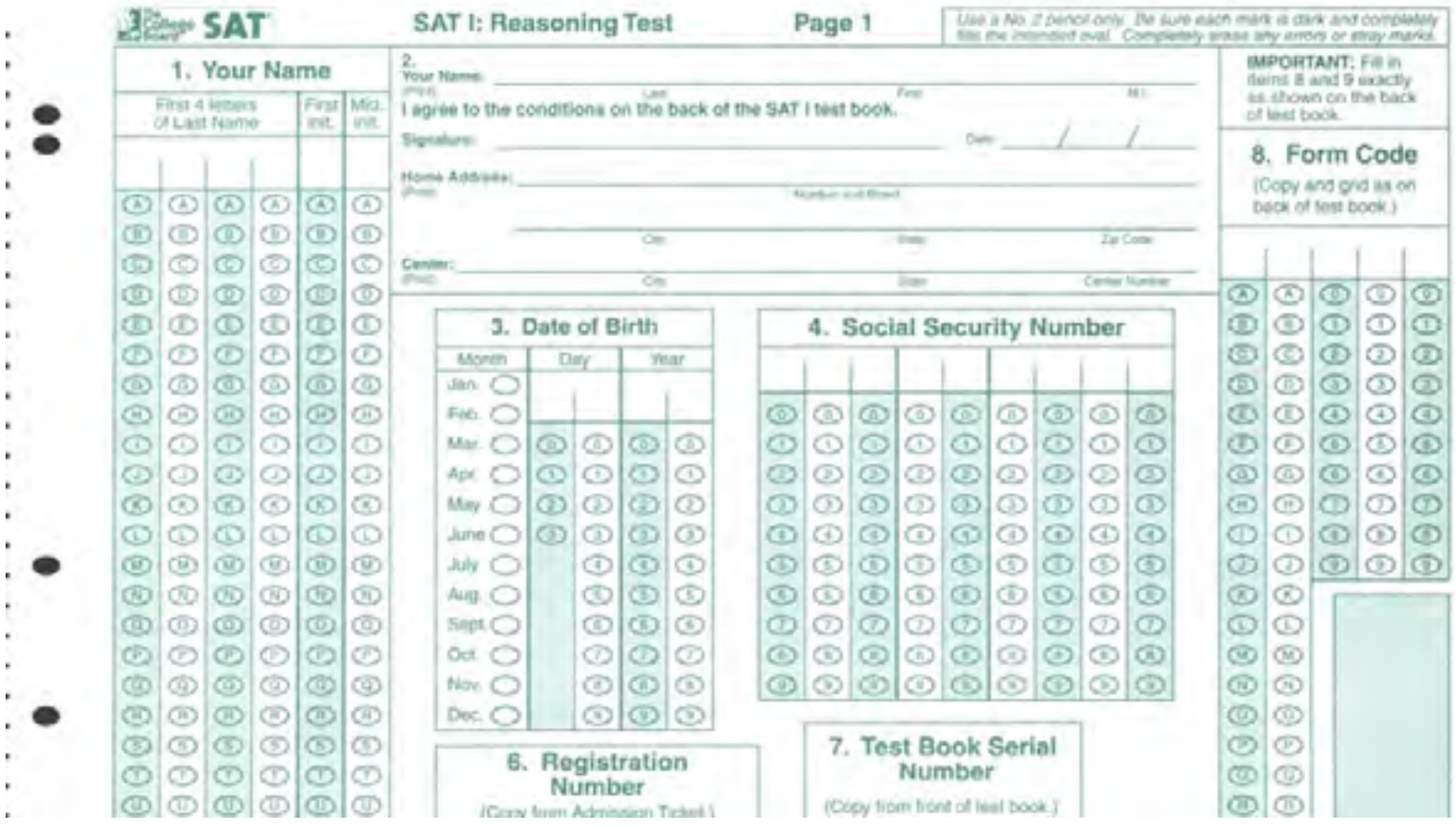
CAIMIRA

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

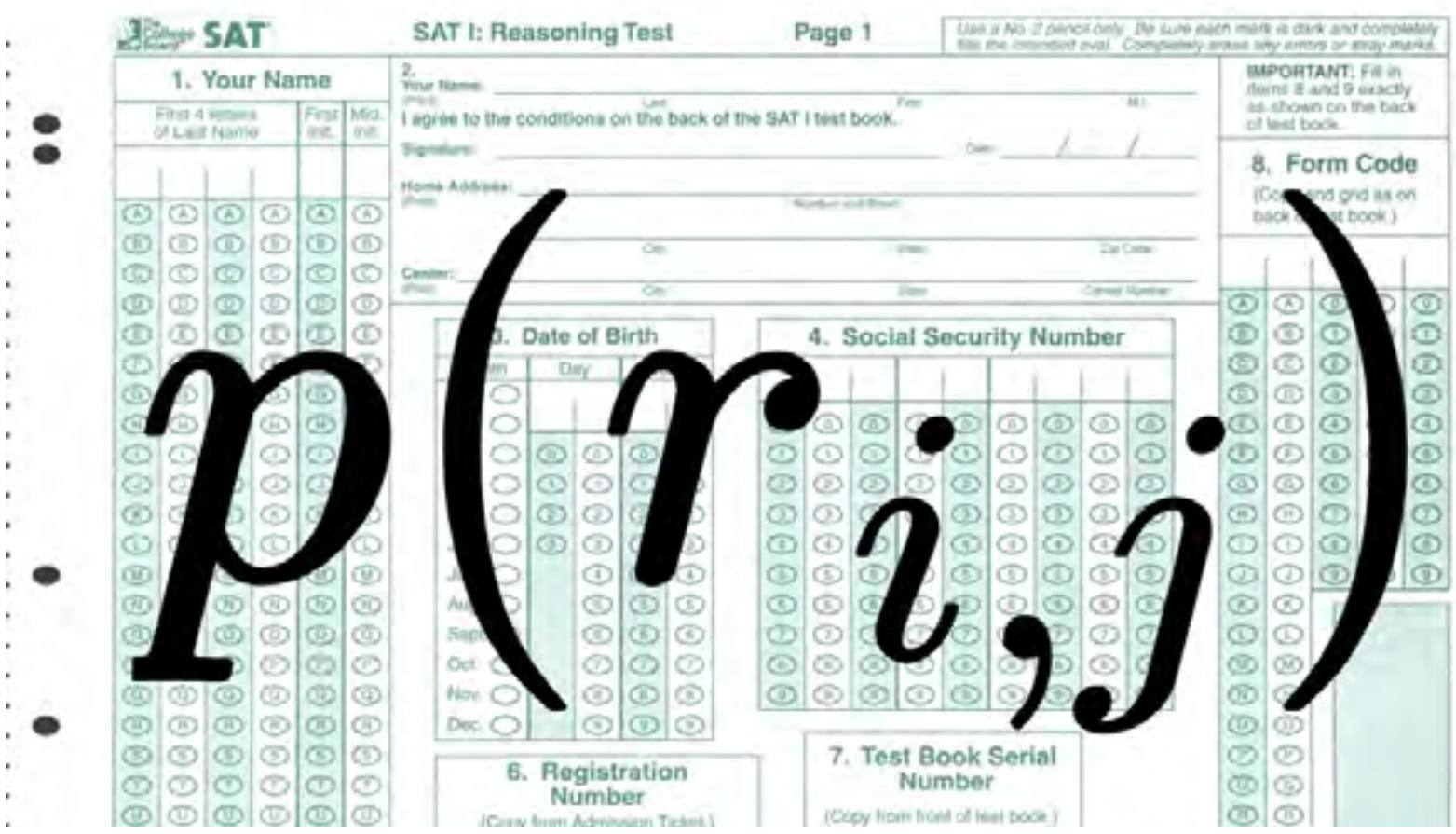
Maharshi Gor[1] **Hal Daumé III**[1,2] **Tianyi Zhou**[1] **Jordan Boyd-Graber**[1]

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

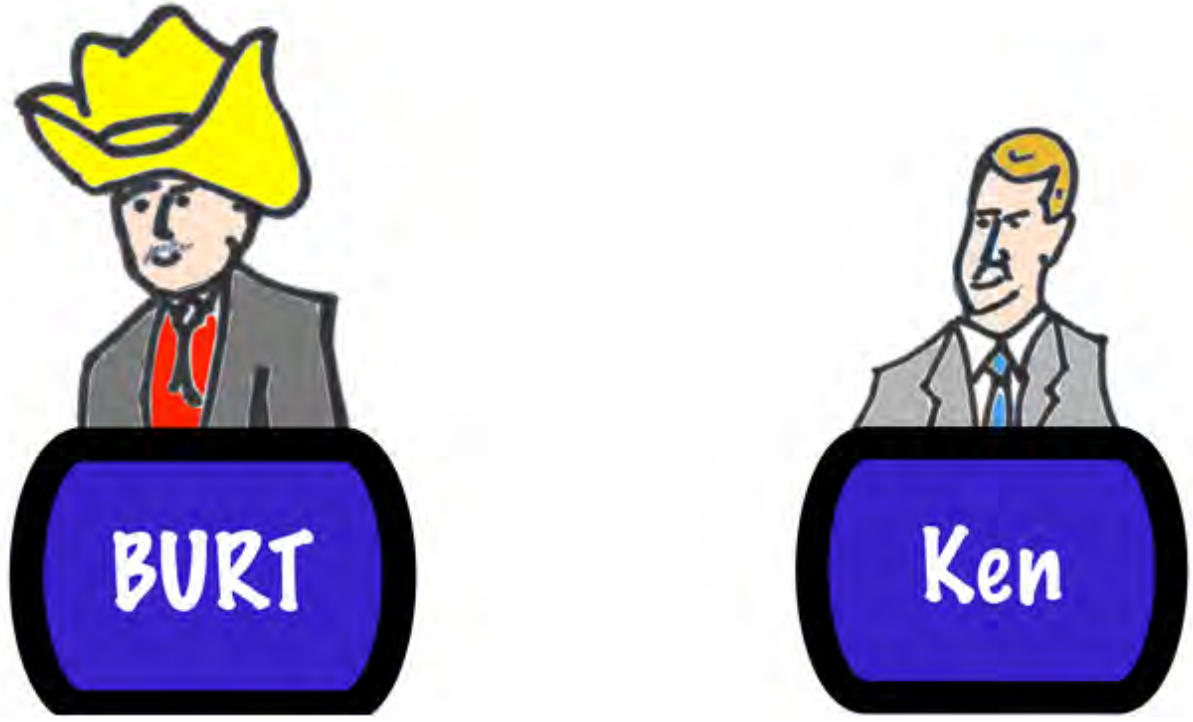
Item Response Theory



Item Response Theory

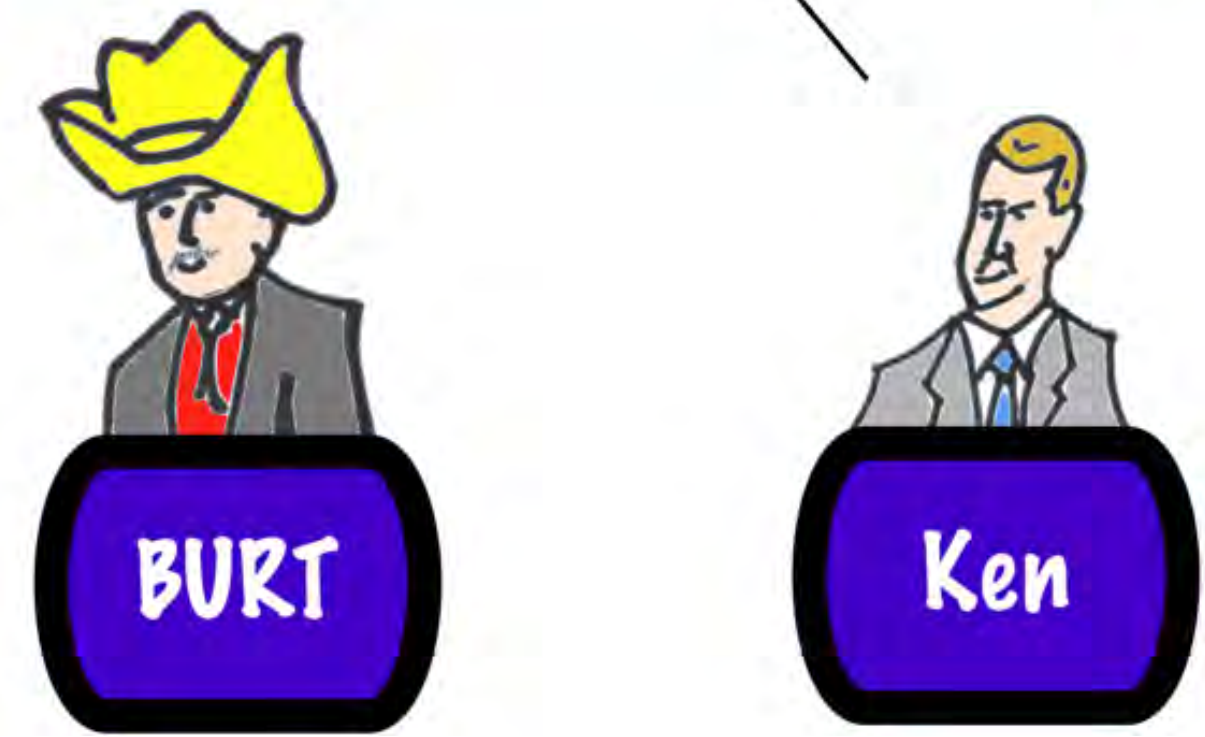


Item Response Theory

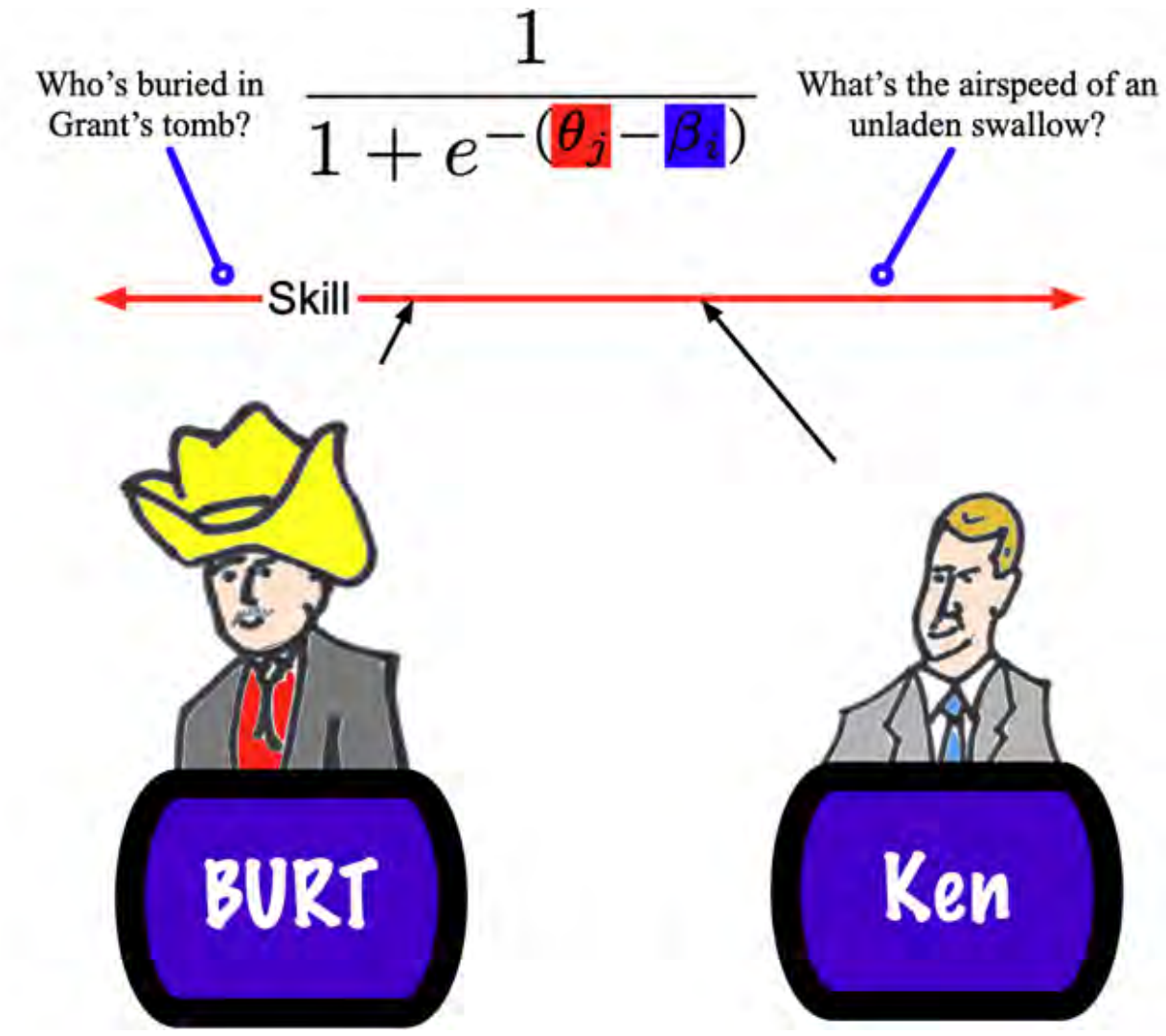


Item Response Theory

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



Item Response Theory



Making Dimensions Interpretable

- Make skills and difficulty vector-valued

$$\gamma_j \sum_k [\theta_{i,k} - \beta_{j,k}] \tag{1}$$

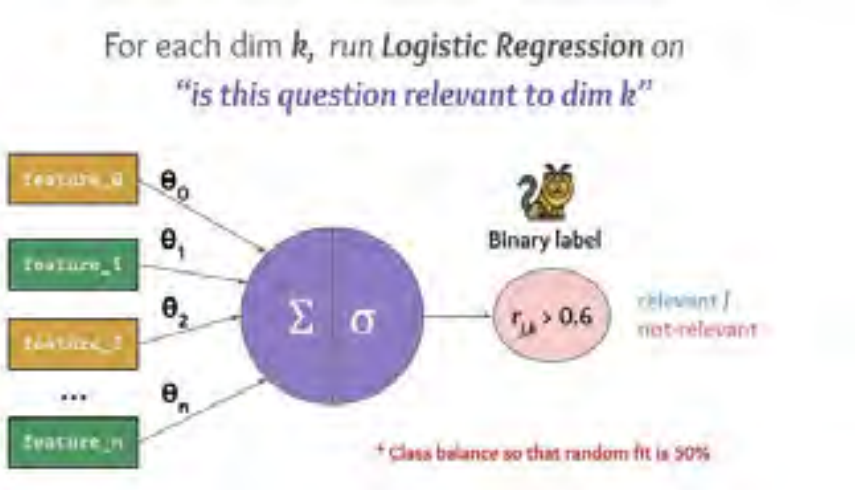
- Where do the dimensions come from?

Making Dimensions Interpretable

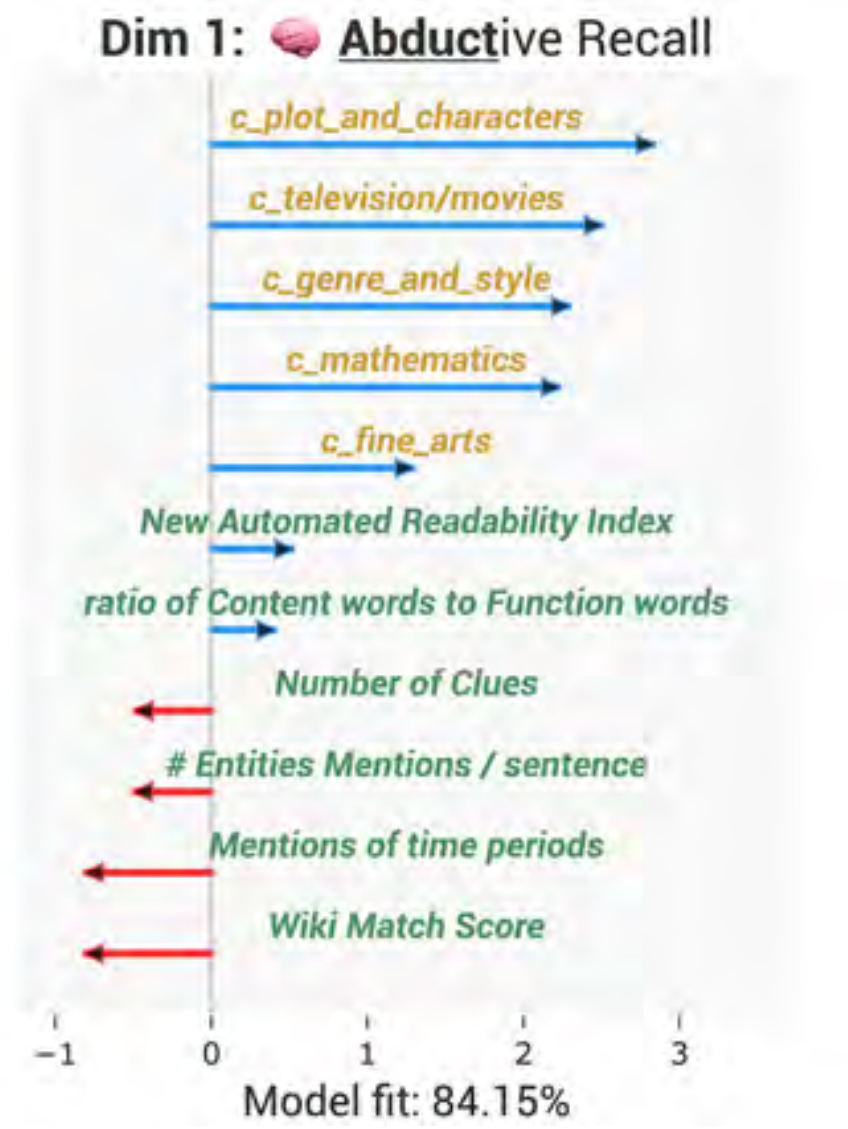
- Make skills and difficulty vector-valued

$$\gamma_j \sum_k [\theta_{i,k} - \beta_{j,k}] \tag{1}$$

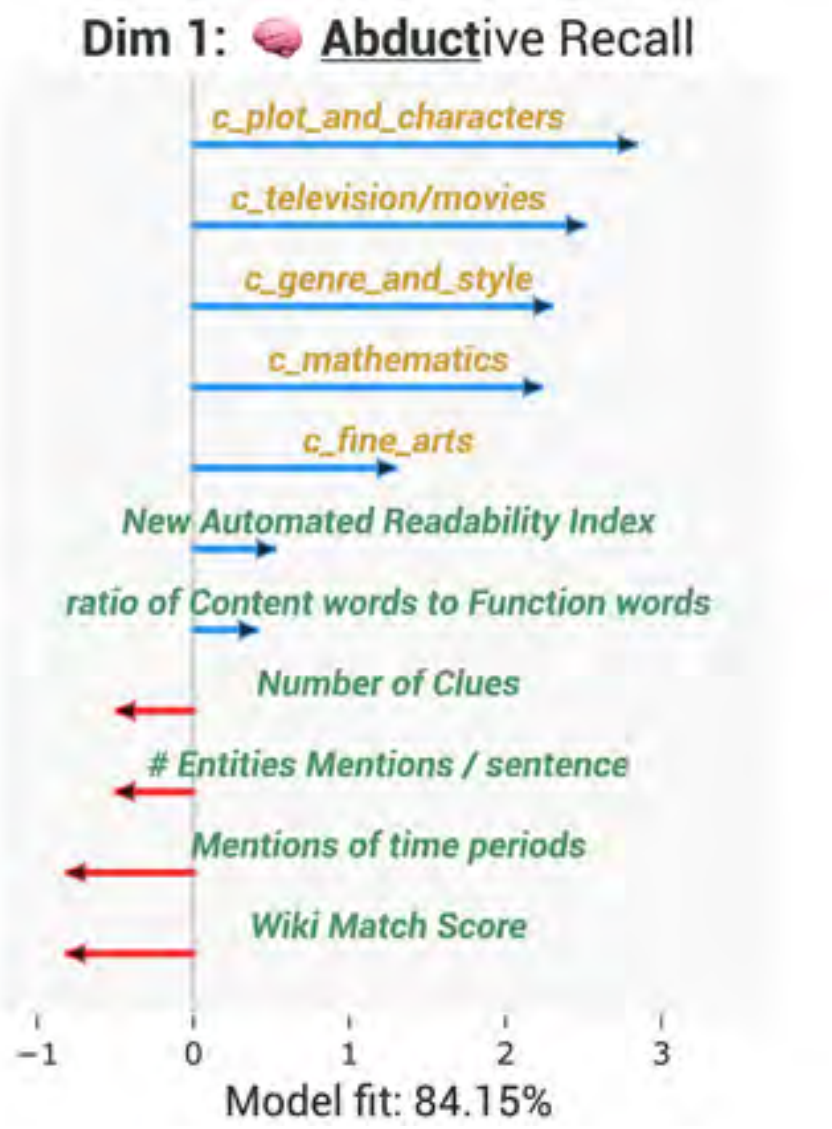
- Where do the dimensions come from?
  - Latent variable: learned to predict correctness
  - Function of question encoding, question features
  - Regularized to be sparse
- Posthoc labeling

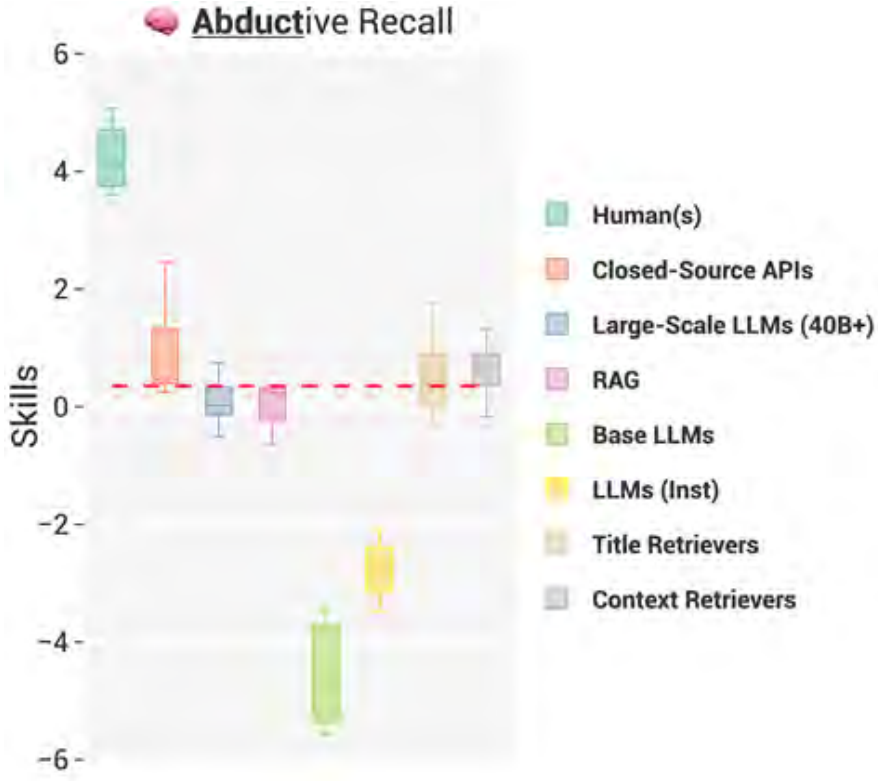


Hard for Computers: Abductive Inference

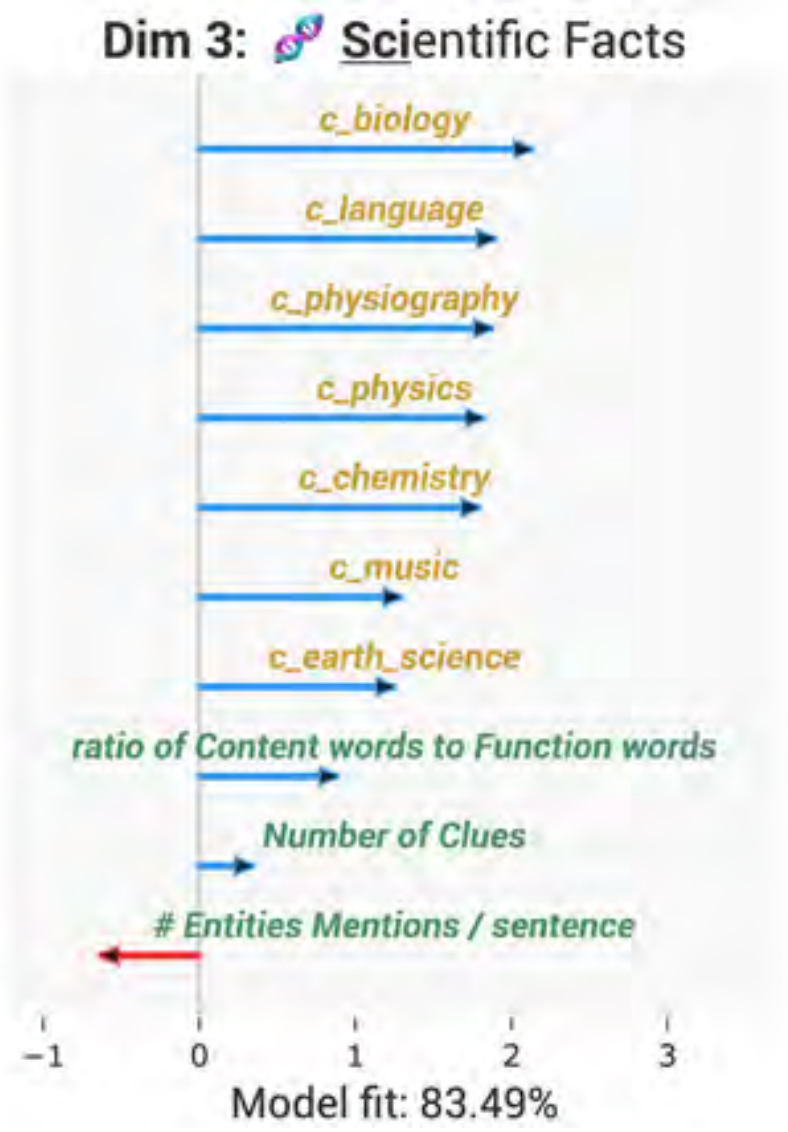


Hard for Computers: Abductive Inference

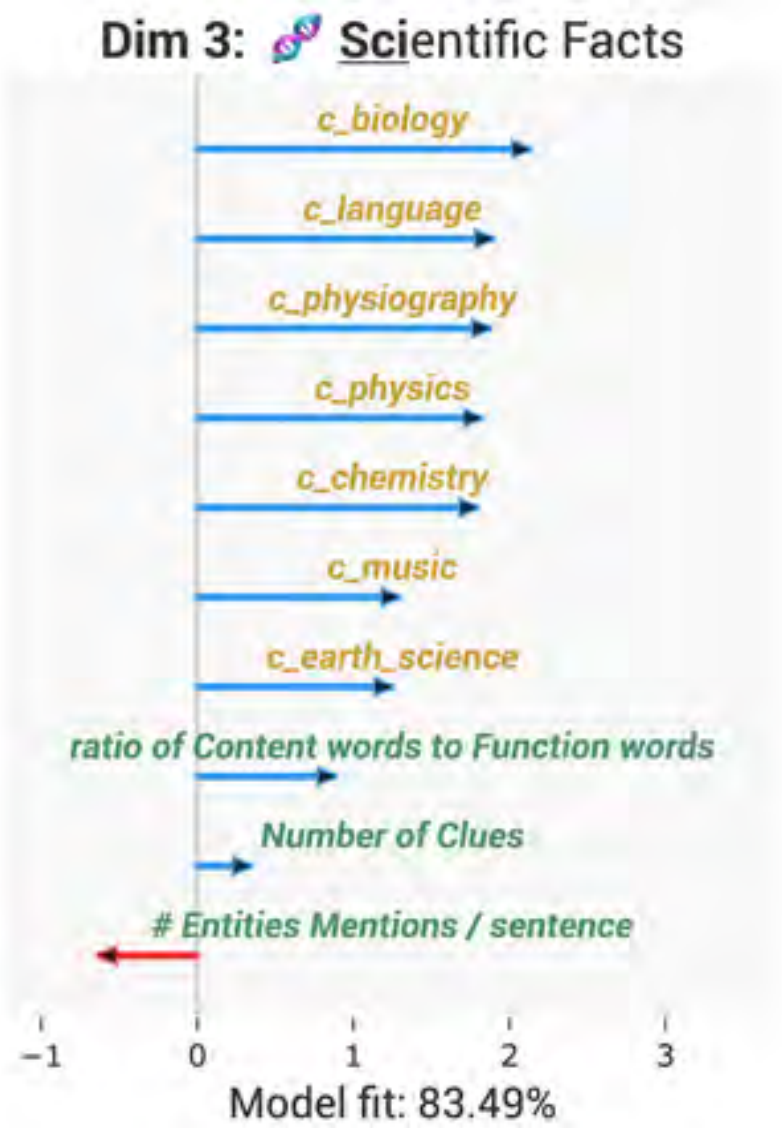


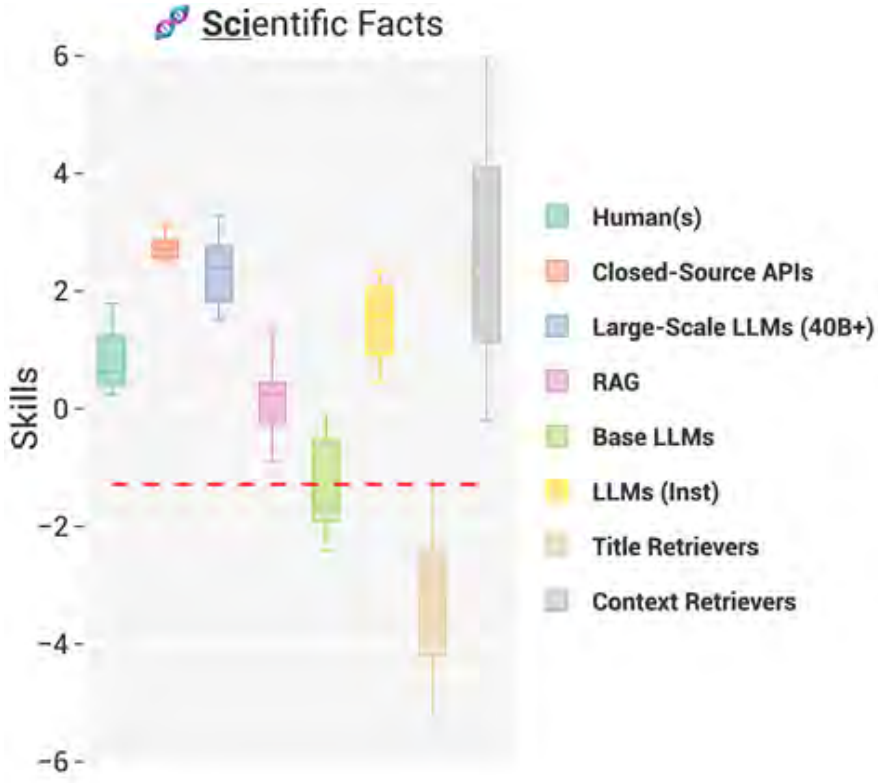


Hard for Humans: Science

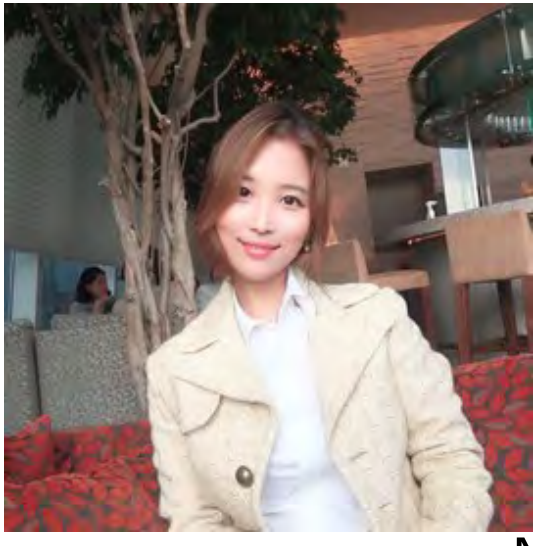


Hard for Humans: Science





AdvScore



Is your benchmark *truly* adversarial?
ADVSCORE: Evaluating Human-Grounded Adversarialness

Yoo Yeon Sung[1], Maharshi Gor[1], Eve Fleisig[2], Ishani Mondal[1], Jordan Boyd-Graber[1]

[1]University of Maryland [2]UC Berkeley

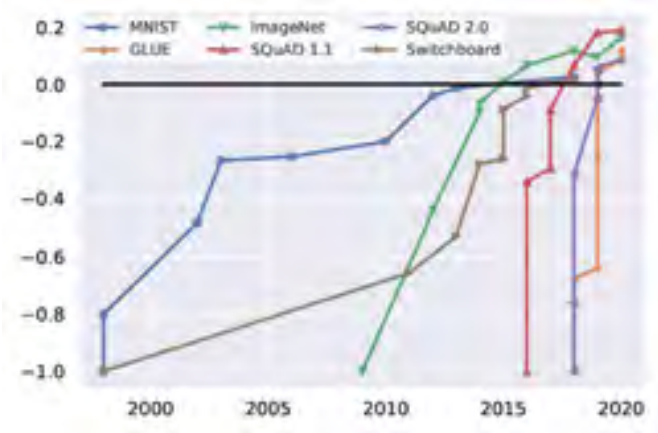
NAACL 2025 Outstanding Paper

Adversarial Datasets

Biggio et al., 2012: Poisoning attacks against Support Vector Machines

- 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



Biggio et al., 2012: Poisoning attacks against Support Vector Machines

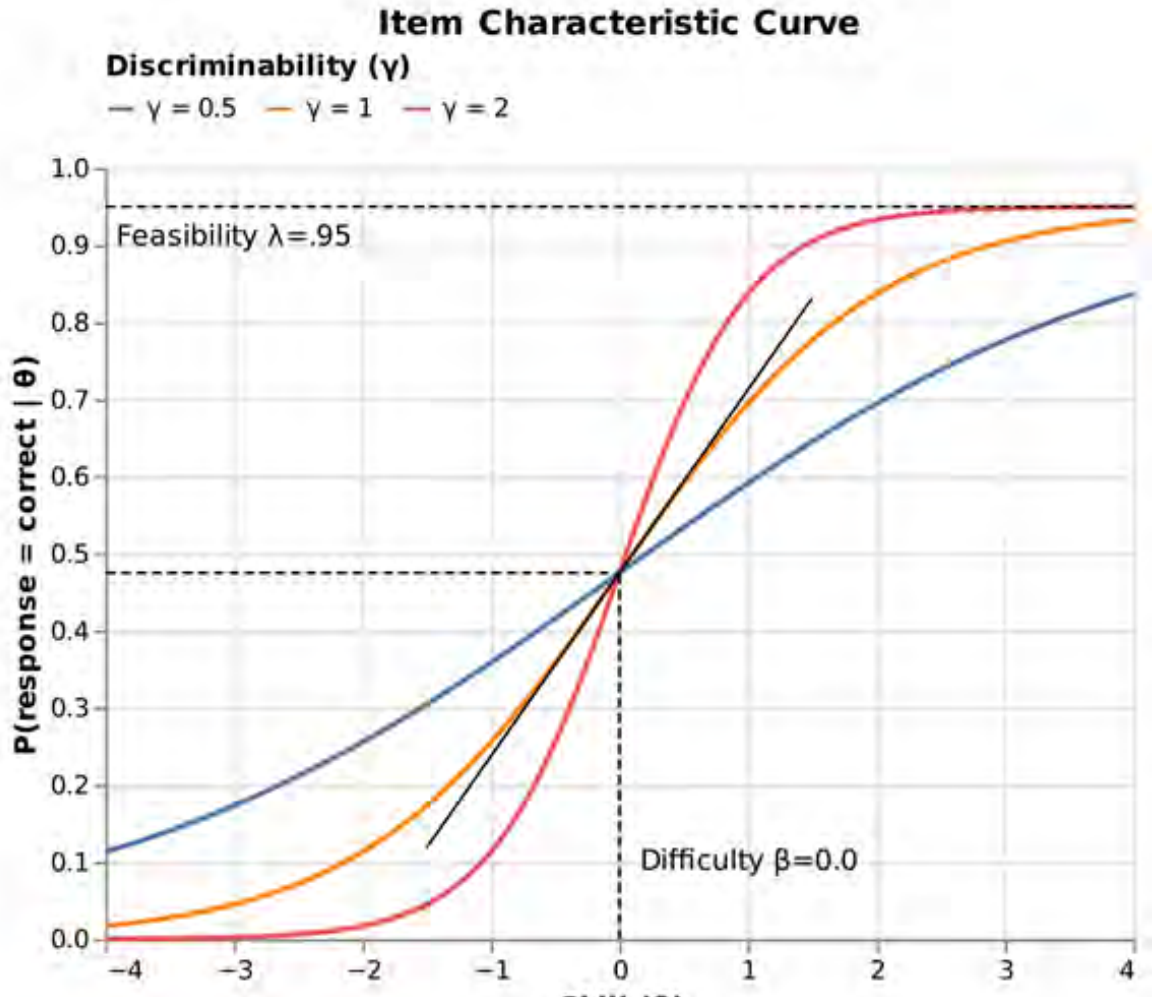
- Many benchmarks are "saturated"
- Newer datasets claim to be "adversarial"
  - Hard for computers, "easy" for humans
  - No real metric / definition
- Can we use the lessons of the previous paper to inform how to write hard examples
- Can we *measure* how well we did?
- Language game: increasing the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

Expanding IRT: Discriminability

$$p_{ij}(r_{ij} = 1) = \frac{\lambda_i}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

- Difficulty: β
- Discriminability: γ
- Feasibility: λ

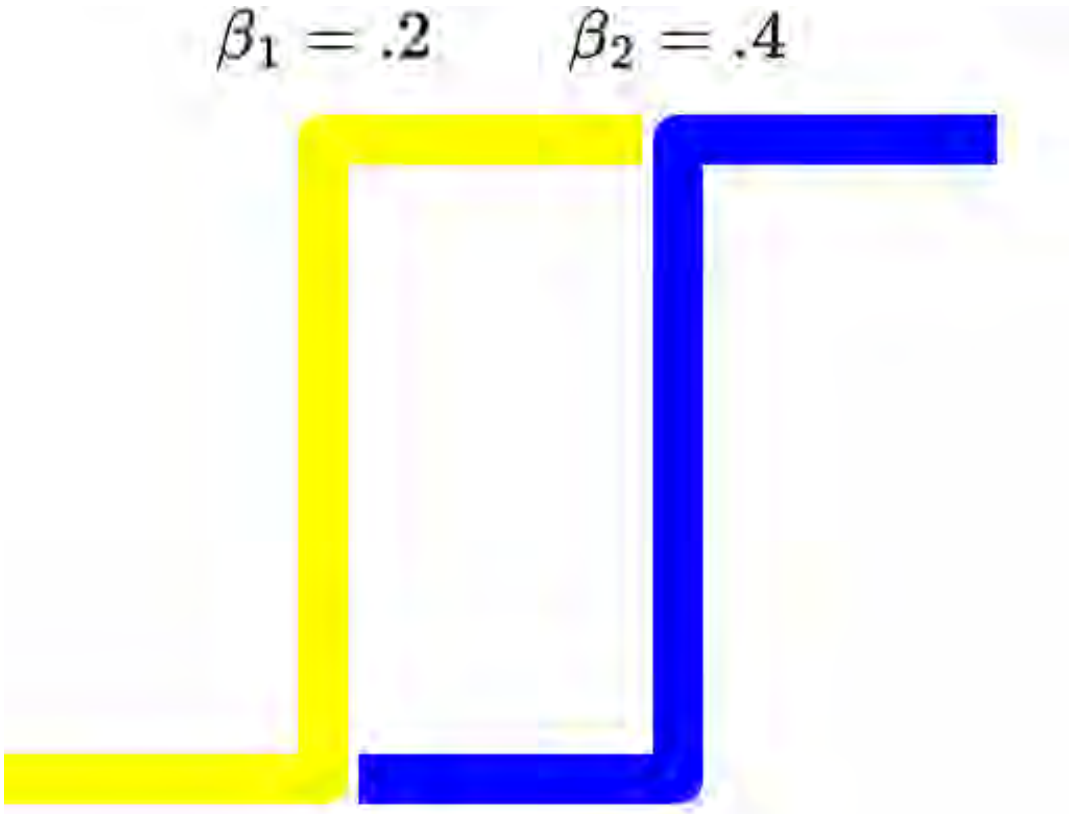
- Skill/ability: θ



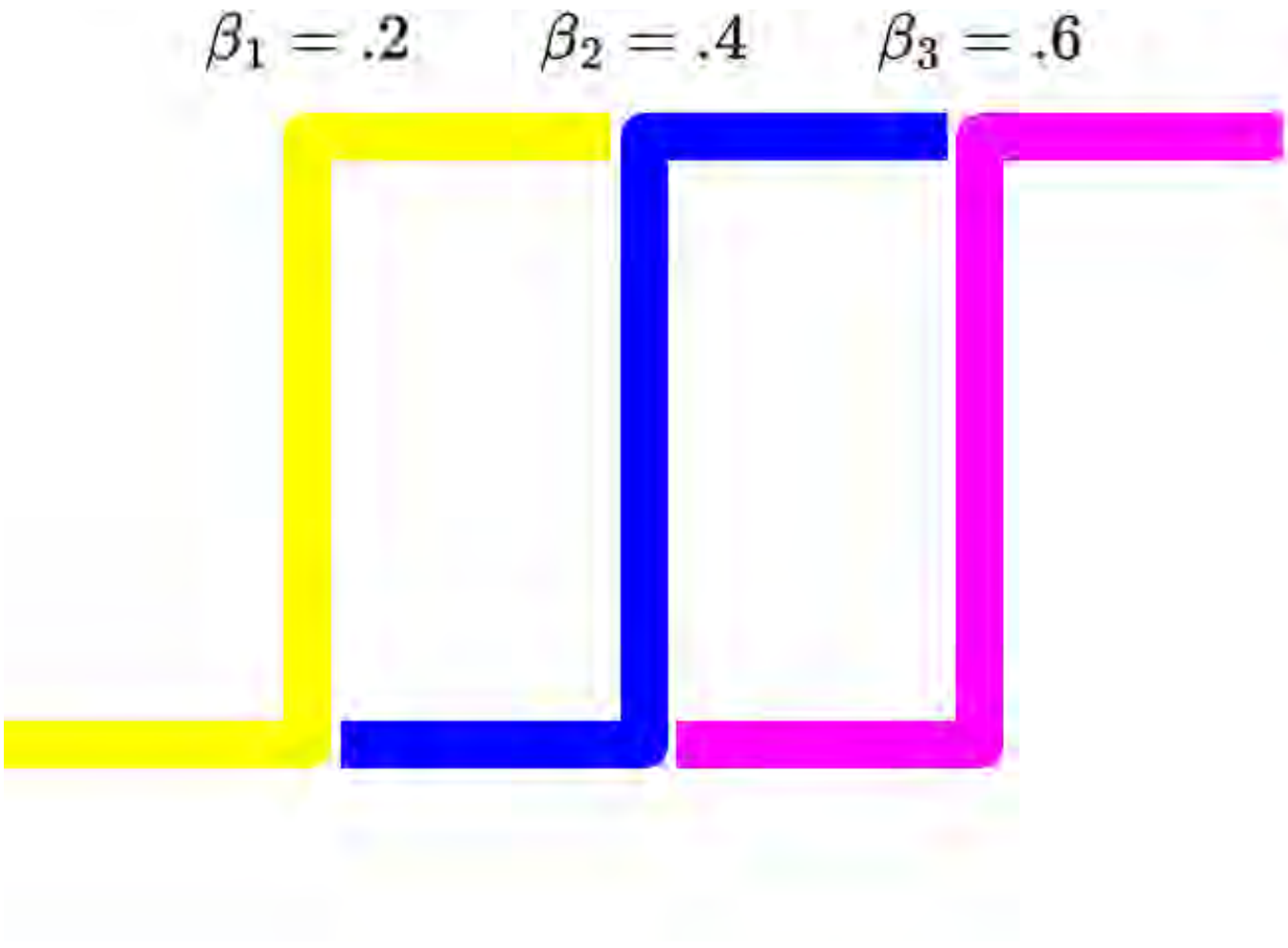
Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2$$

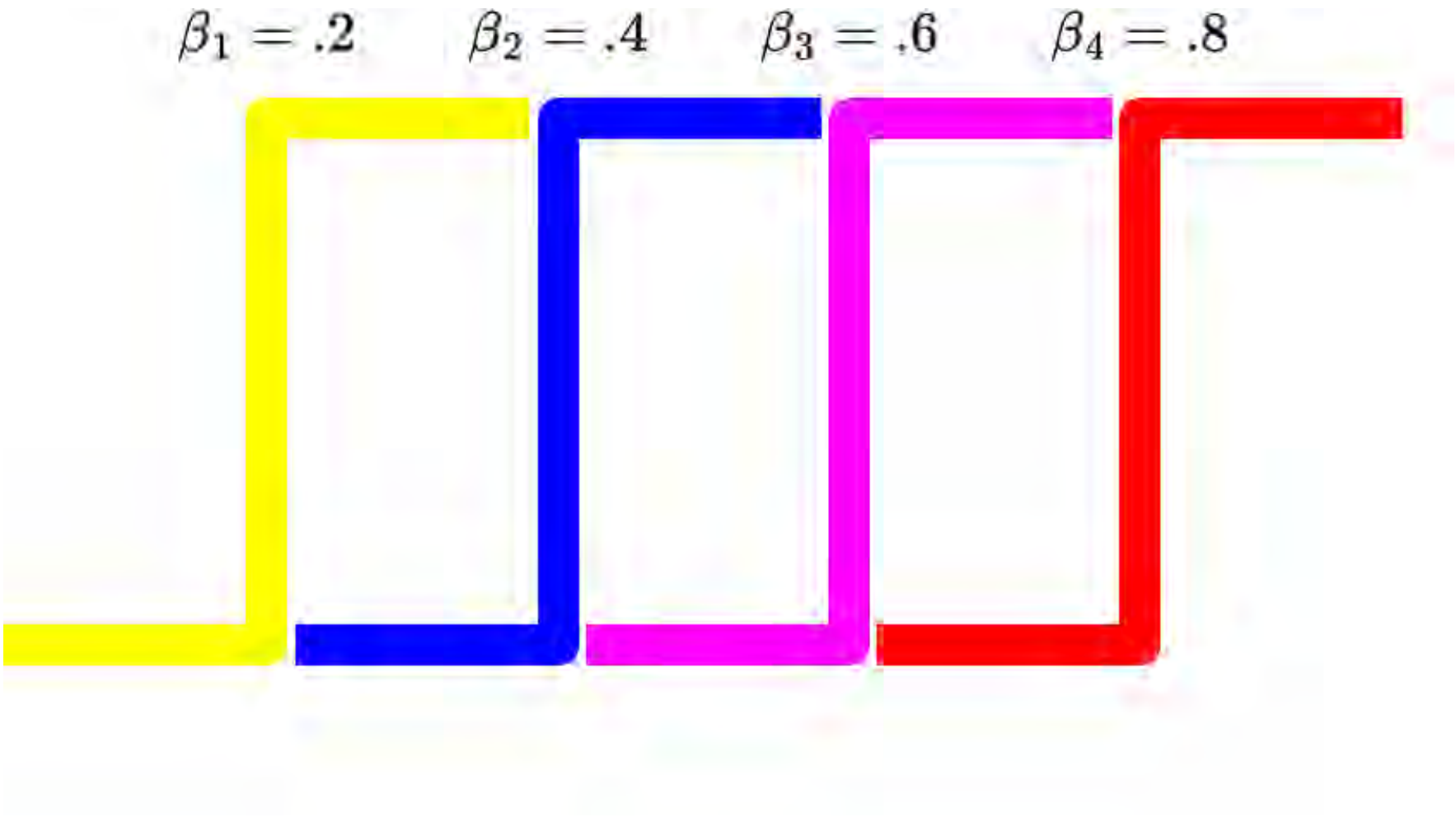
Finding Skill of Subject is Like Binary Search



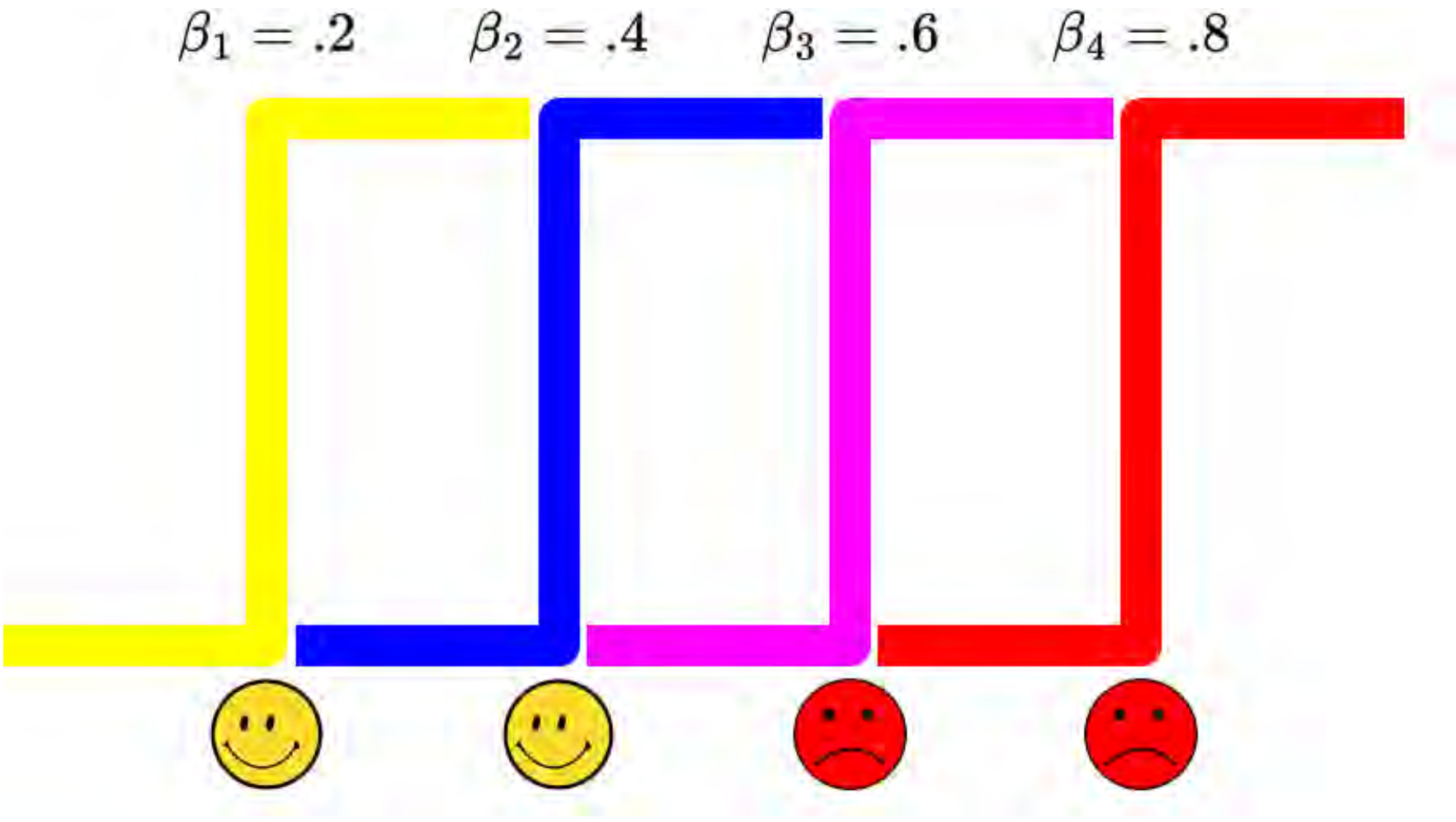
Finding Skill of Subject is Like Binary Search



Finding Skill of Subject is Like Binary Search



Finding Skill of Subject is Like Binary Search



Adversarial Score

- Gap between skilled human getting it right and machine (should be big)

$$\mu_j = \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{H(0)} - \theta_j\right]\right\}}}_{\text{Skilled human rep. prob.}} - \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{M(0)} - \theta_j\right]\right\}}}_{\text{Skilled model rep. prob.}}, \tag{2}$$

Adversarial Score

- Gap between skilled human getting it right and machine (should be big)

$$\mu_j = \underbrace{\frac{1}{1 + \exp\left\{-\gamma_j \left[\beta_*^{H(0)} - \theta_j\right]\right\}}}_{\text{Skilled human rep. prob.}} - \underbrace{\frac{1}{1 + \exp\left\{-\gamma_j \left[\beta_*^{M(0)} - \theta_j\right]\right\}}}_{\text{Skilled model rep. prob.}}, \tag{2}$$

Why not use raw accuracy?

- Want patterns, not luck
- IRT can find (and downweight) bad questions
- What's the capital of Georgia?

Adversarial Score

- Gap between skilled human getting it right and machine (should be big)

$$\mu_j = \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{H_{(0)}} - \theta_j\right]\right\}}}_{\text{Skilled human rep. prob.}} - \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{M_{(0)}} - \theta_j\right]\right\}}}_{\text{Skilled model rep. prob.}}, \tag{2}$$

- Skilled humans should agree on the answer

$$\delta_j = \sum_{i \sim H_{(1)}} \left[\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_i^{H_{(1)}} - \theta_j\right]\right\}} - \overline{p_{H_{(1)}}}(r_{i,j})\right] / |H_{(1)}| \tag{3}$$

Adversarial Score

$$\textsc{AdvScore}_{\text{j}} = \frac{\mu_j}{1+\delta_j} \tag{2}$$

- Gap between skilled human getting it right and machine (should be big)

$$\mu_j = \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{H_{(0)}} - \theta_j\right]\right\}}}_{\text{Skilled human rep. prob.}} - \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{M_{(0)}} - \theta_j\right]\right\}}}_{\text{Skilled model rep. prob.}}, \tag{3}$$

- Skilled humans should agree on the answer

$$\delta_j = \sum_{i \sim H_{(1)}} \left[\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_i^{H_{(1)}} - \theta_j\right]\right\}} - \overline{p_{H_{(1)}}}(r_{i,j})\right] / |H_{(1)}| \tag{4}$$

Adversarial Score

$$\textsc{AdvScore}_{\text{j}} = \frac{\mu_j}{1+\delta_j} \tag{2}$$

- Gap between skilled human getting it right and machine (should be big)

$$\mu_j = \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{H_{(0)}} - \theta_j\right]\right\}}}_{\text{Skilled human rep. prob.}} - \underbrace{\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_*^{M_{(0)}} - \theta_j\right]\right\}}}_{\text{Skilled model rep. prob.}}, \tag{3}$$

- Skilled humans should agree on the answer

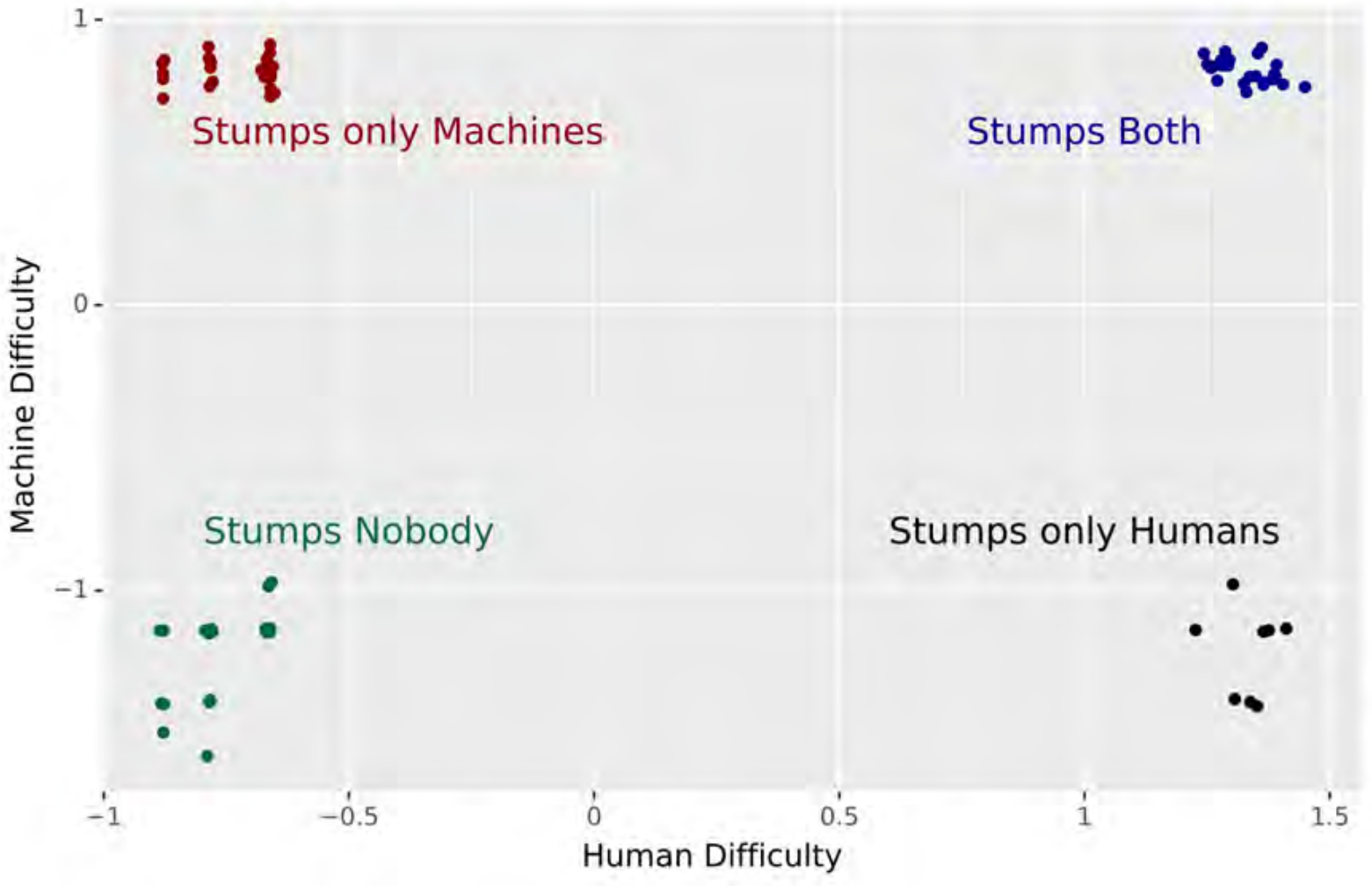
$$\delta_j = \sum_{i \sim H_{(1)}} \left[\frac{1}{1+\exp\left\{-\gamma_j\left[\beta_i^{H_{(1)}} - \theta_j\right]\right\}} - \overline{p_{H_{(1)}}}(r_{i,j})\right] \Big/ |H_{(1)}| \tag{4}$$

- In paper, also have a bonus for Fischer information wrt to difficulty θ, but not a huge factor

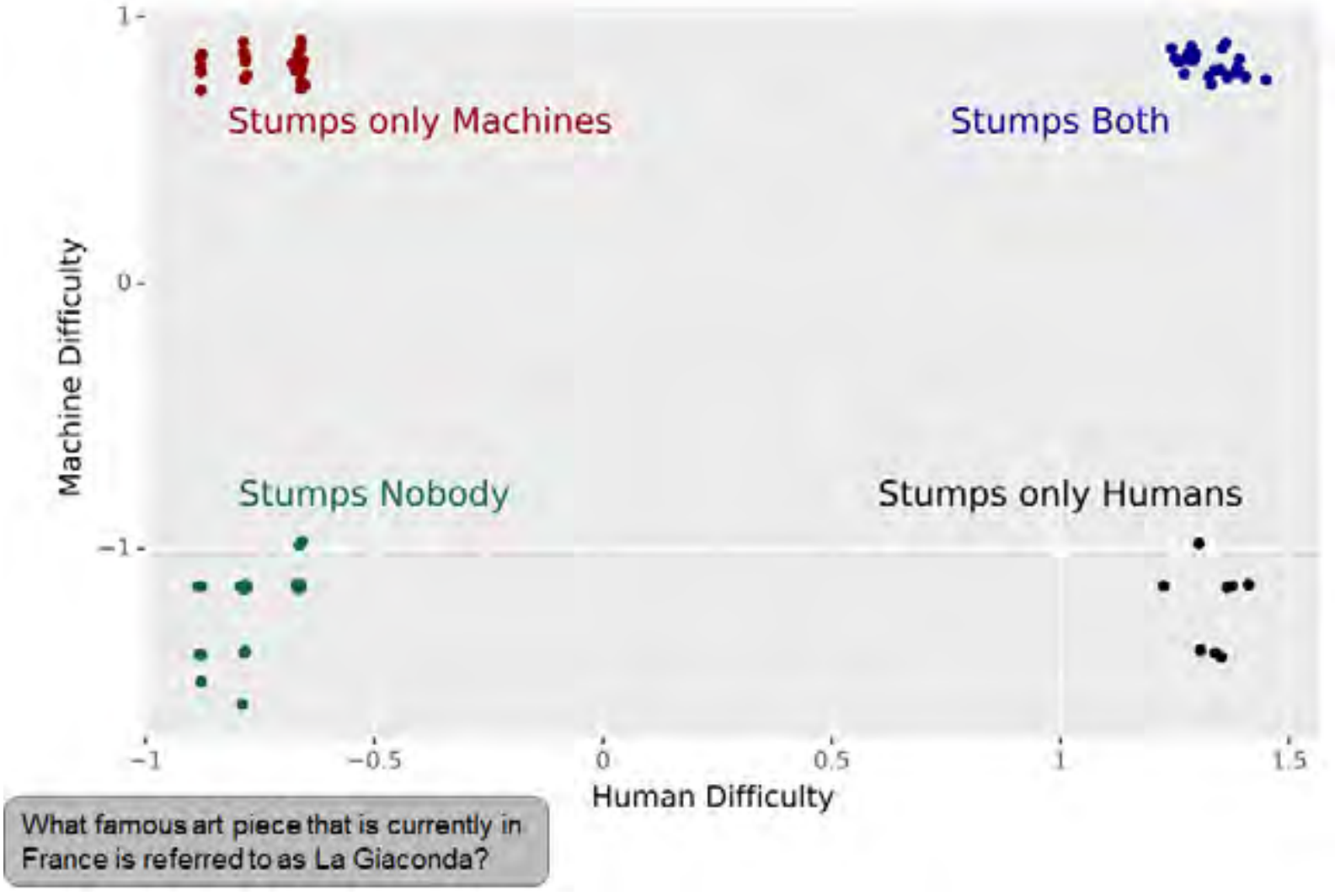
AdvQA: Is this a viable incentive structure?

- 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

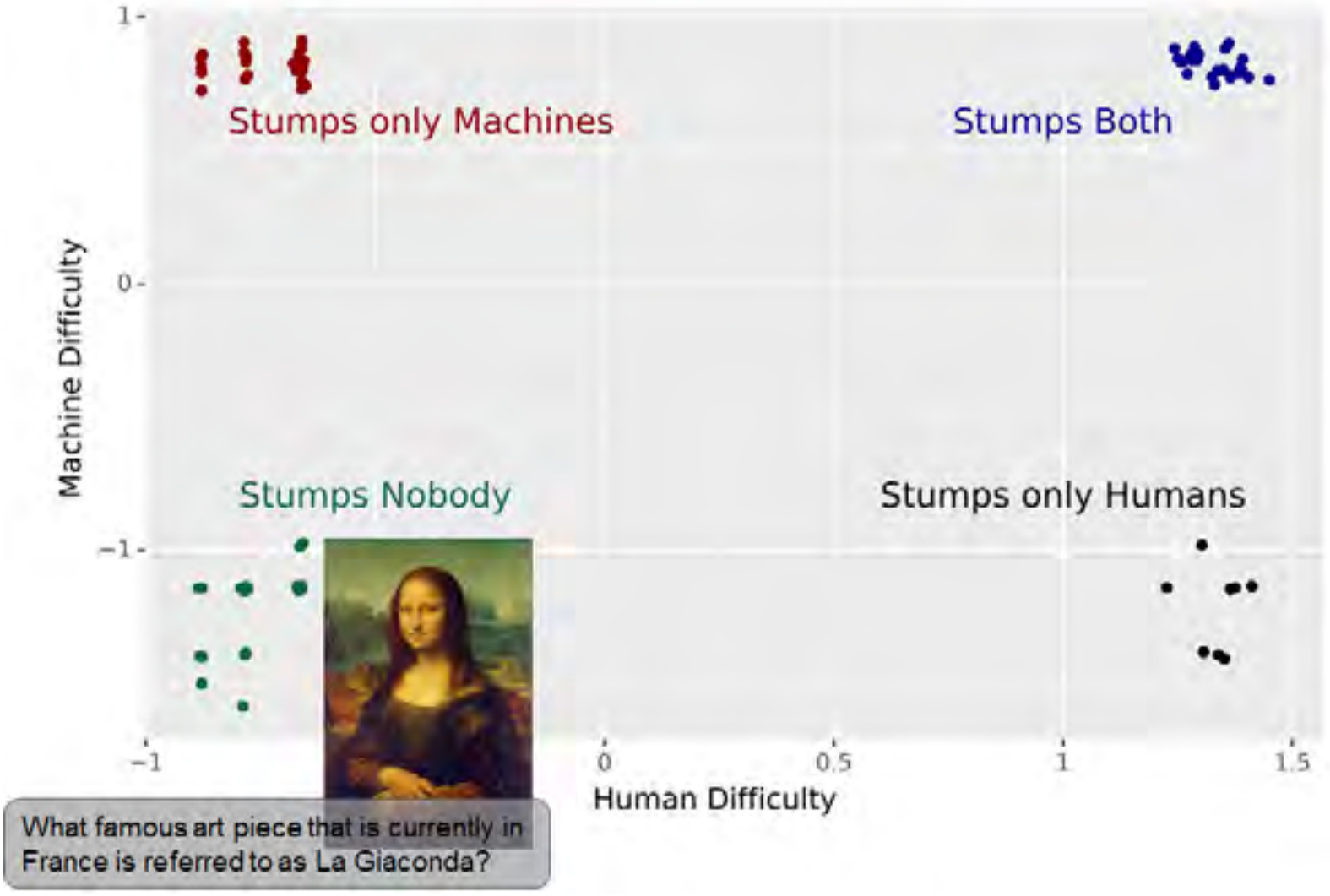
What makes for Adversarial Example



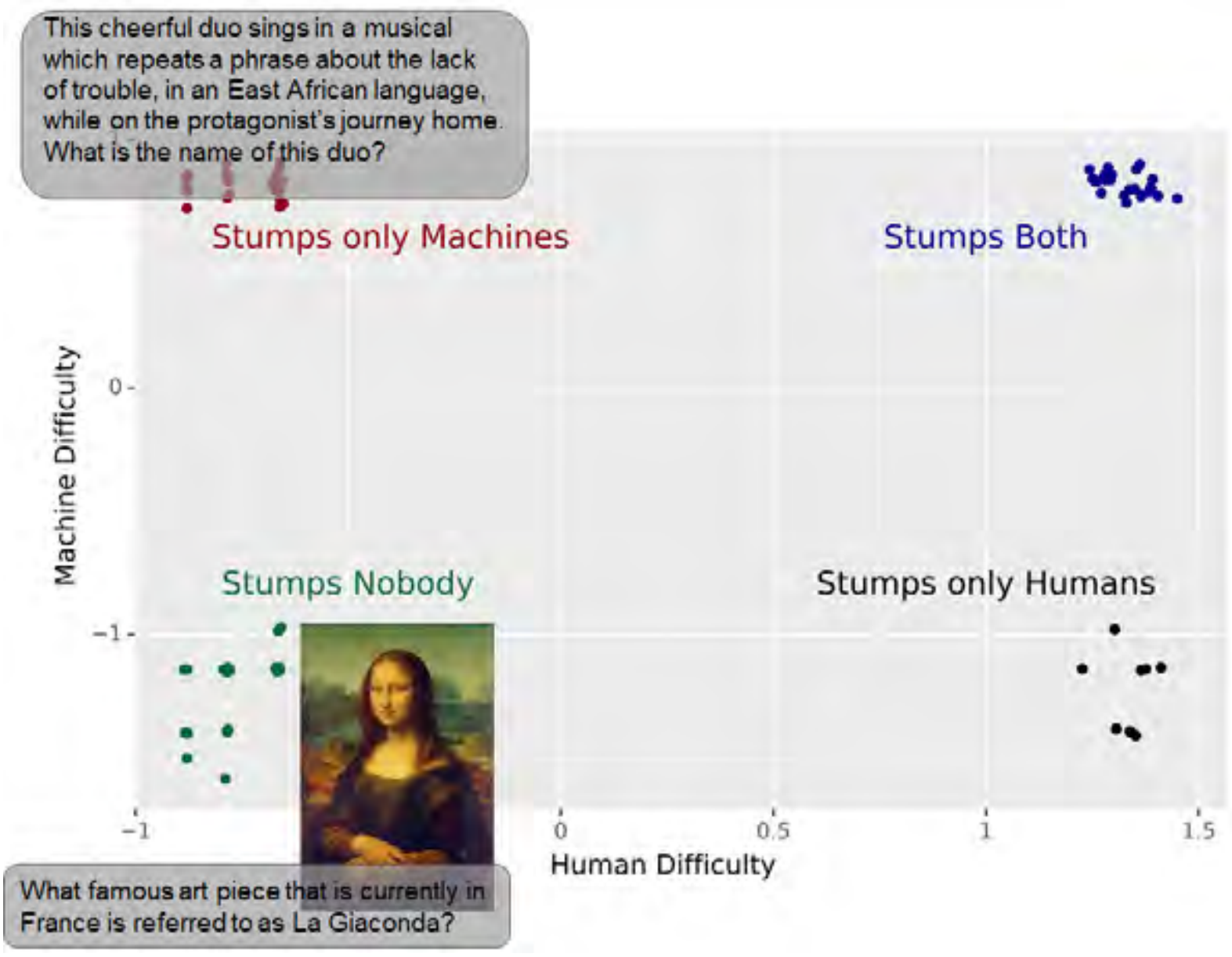
What makes for Adversarial Example



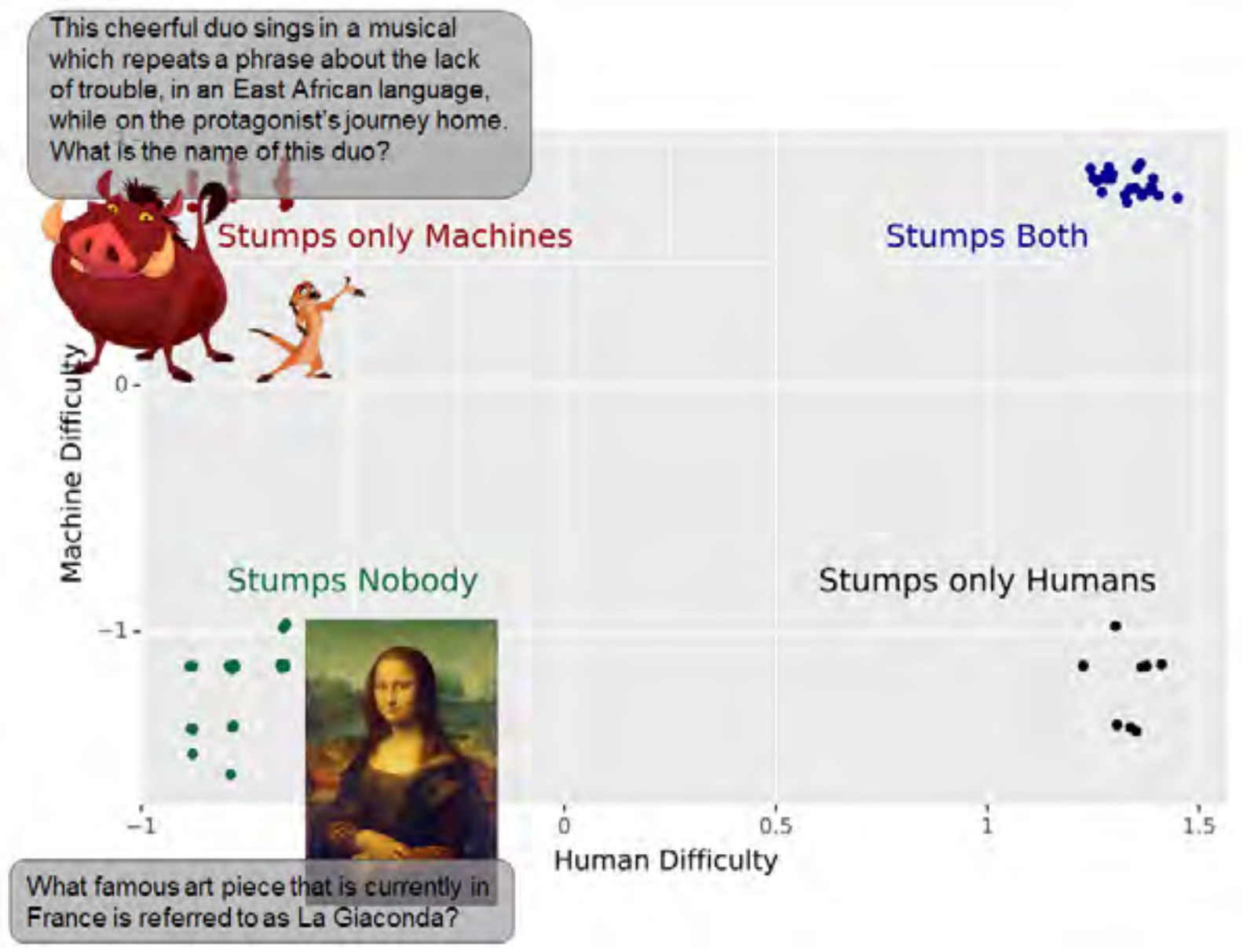
What makes for Adversarial Example



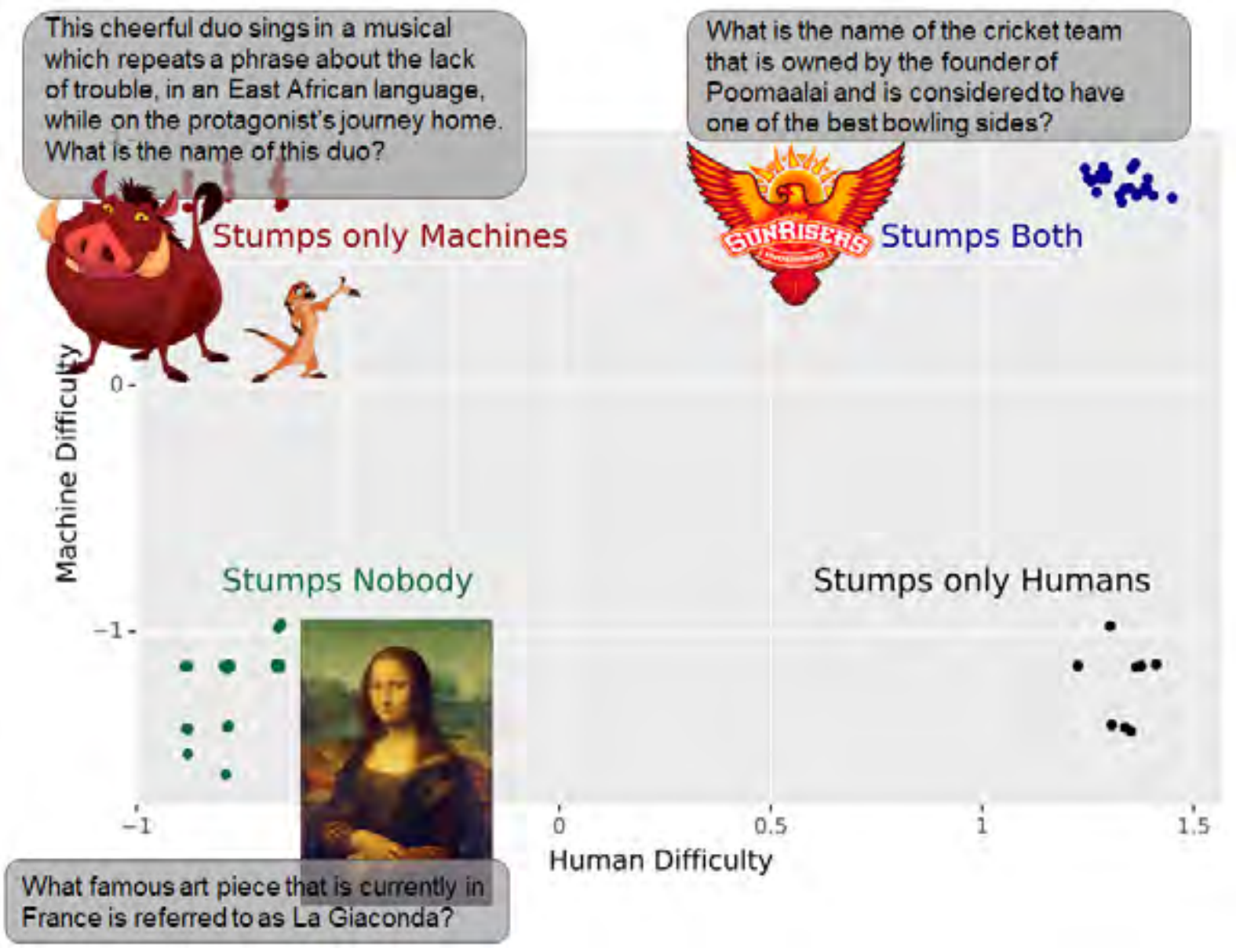
What makes for Adversarial Example



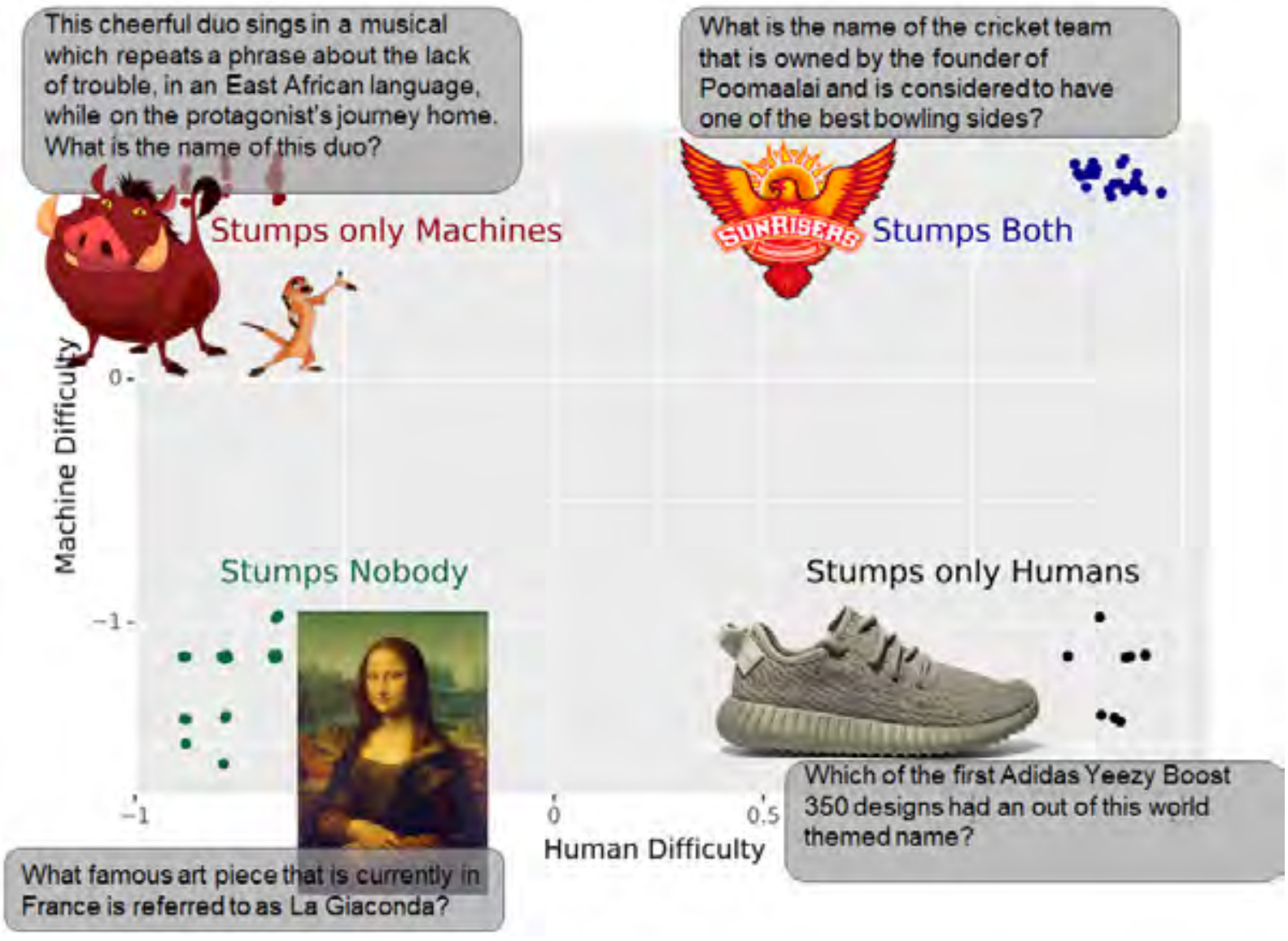
What makes for Adversarial Example



What makes for Adversarial Example



What makes for Adversarial Example



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?
Brad Pitt / Will Smith

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?

Snowpiercer / Train to Busan

Adversarial Strategies

It's not headquartered in Biel, Switzerland but this activewear company has a logo that resembles the last letter of the Greek alphabet.

Adversarial Strategies

It's not headquartered in Biel, Switzerland but this activewear company has a logo that resembles the last letter of the Greek alphabet.

Omega / Lululemon

Adversarial Strategies

A character in one story by this author opens Crime and Punishment to discover that it has turned into The Brothers Karamazov

Adversarial Strategies

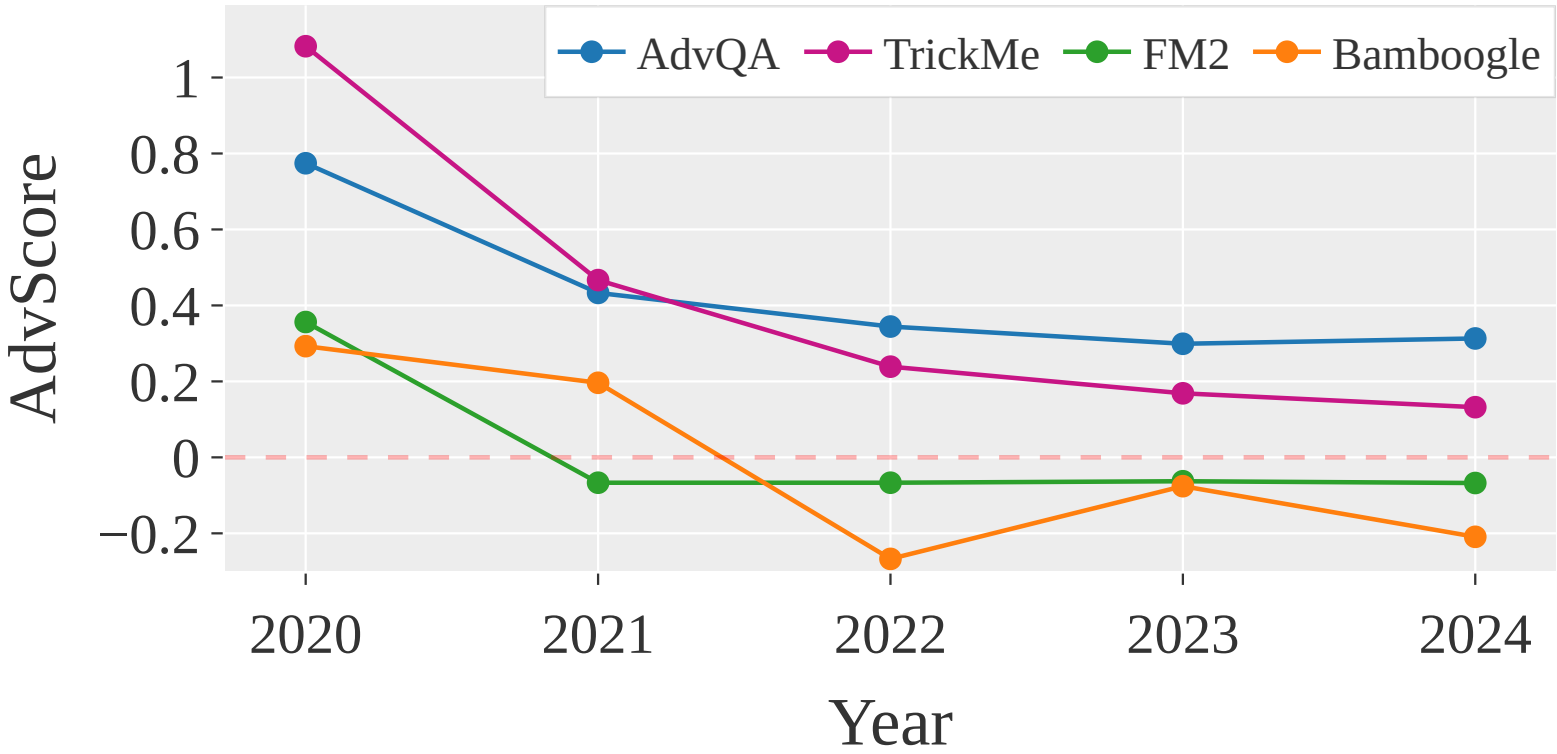
A character in one story by this author opens Crime and Punishment to discover that it has turned into The Brothers Karamazov

Adversarial Strategies

A character in one story by this author opens Crime and Punishment to discover that it has turned into The Brothers Karamazov

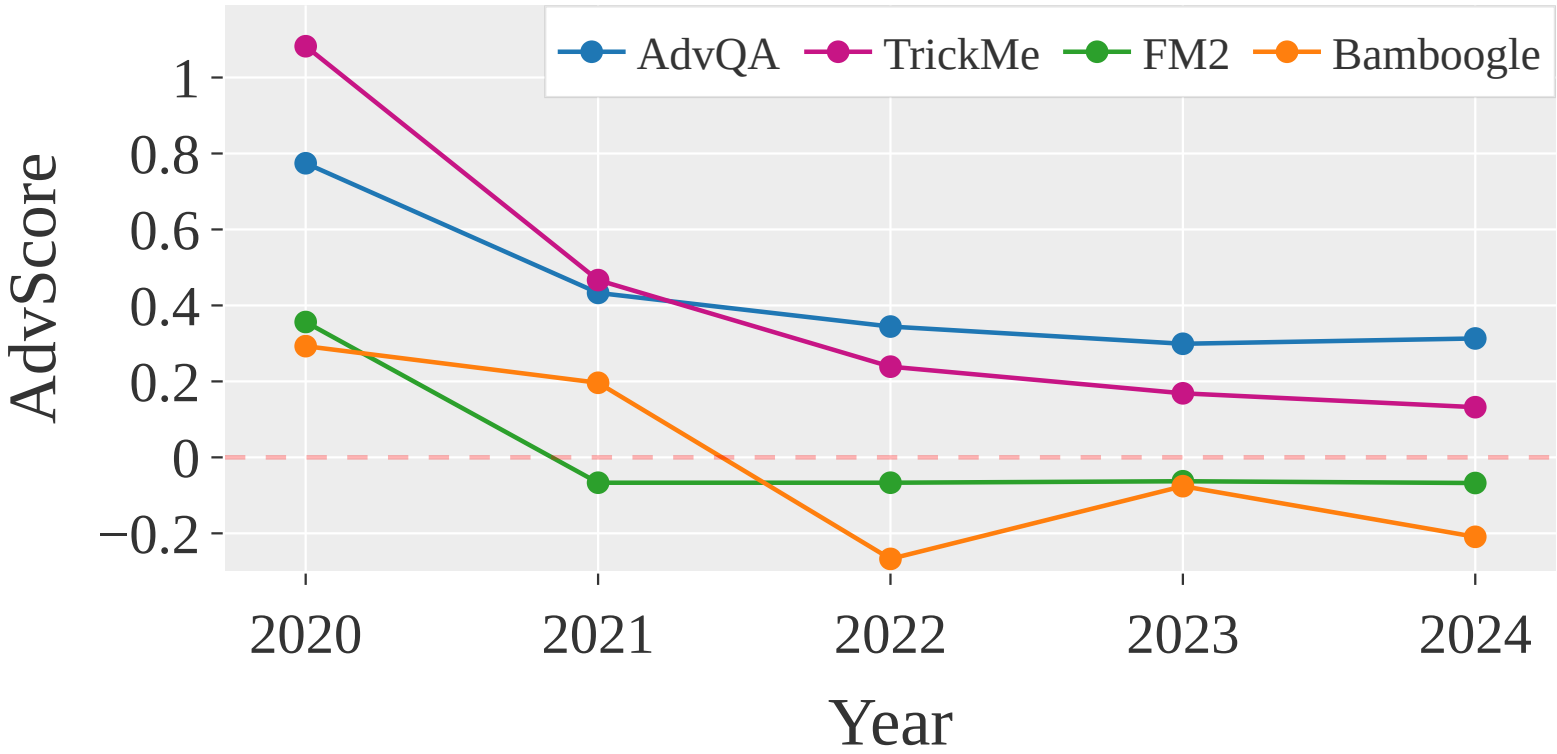
Dostoyevski / Akutagawa

Which Datasets are Adversarial?



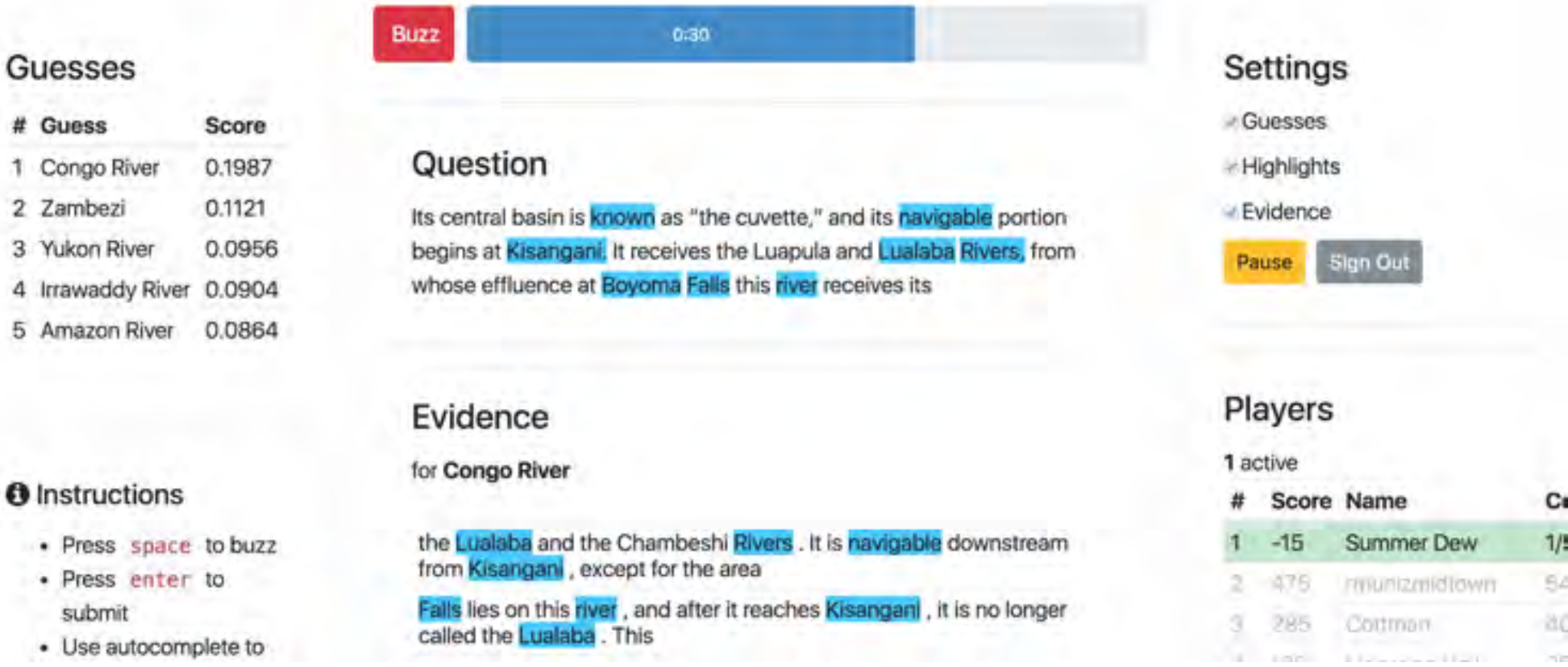
- Not all datasets remain adversarial forever
- What helps make datasets adversarial?
  - Bamboogle: Automatically generated
  - TrickMe: Human in the loop interface (expert), IR models
  - FM2: Human in the loop interface (crowdworker), IR models
  - AdvQA: Huaman in the loop (expert), LLM model + category

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
  - AdvQA: Huaman in the loop (expert), LLM model + category

Centaur Chess


Buzz
0:30
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
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
Settings
Guesses
Highlights
Evidence
Pause
Sign Out
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
Players
1 active
Score Name
1 -15 Summer Dew
Instructions
Press space to buzz
Press enter to submit
Use autocomplete to

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

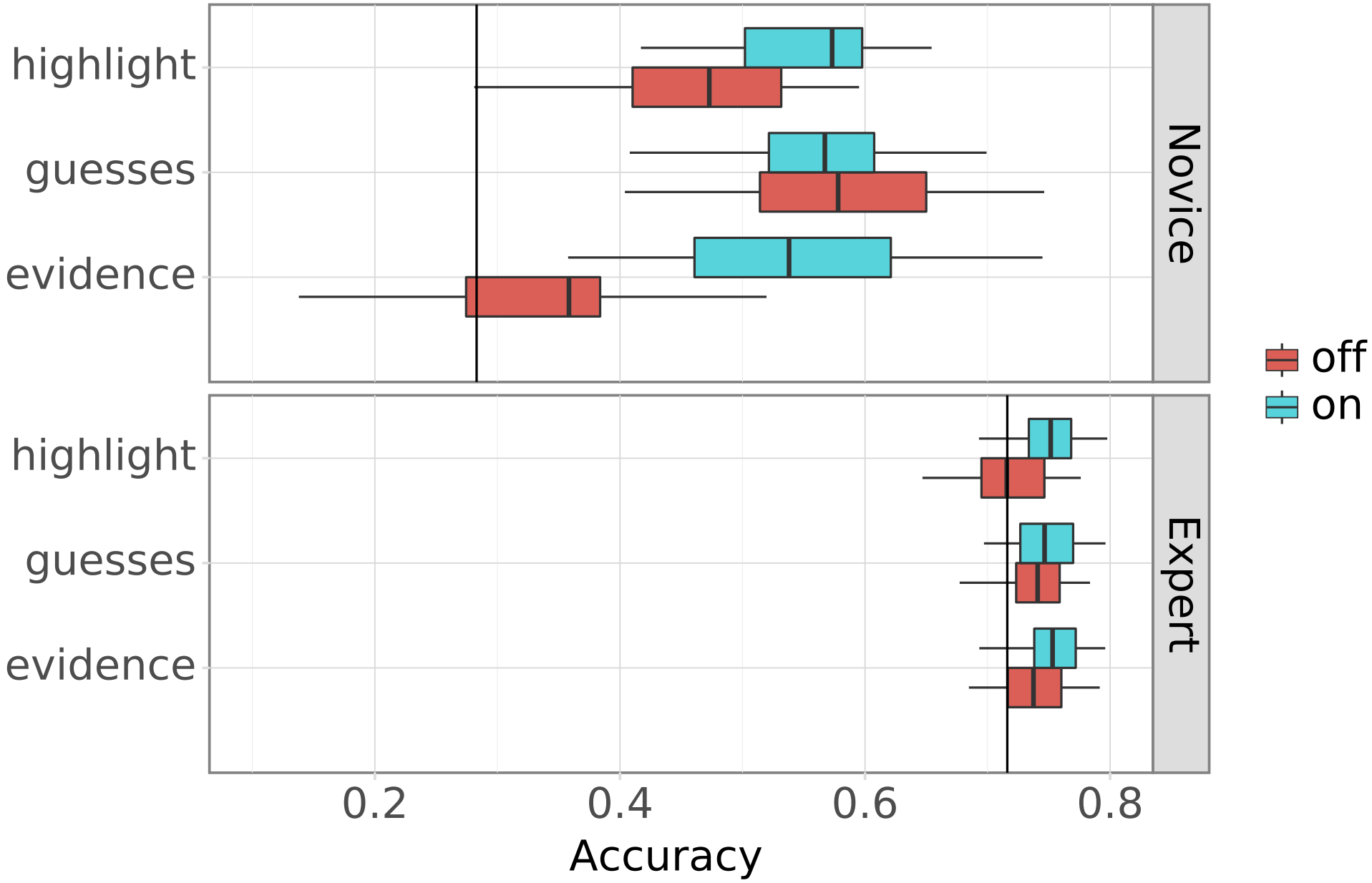
Experts vs. Novices

Experts

Trivia experts, familiar with task, enjoy the task

Mechanical Turkers

Mechanical Turkers: easily overwhelmed, need the help



Evidence helps novices, experts are expert

Skill Boost

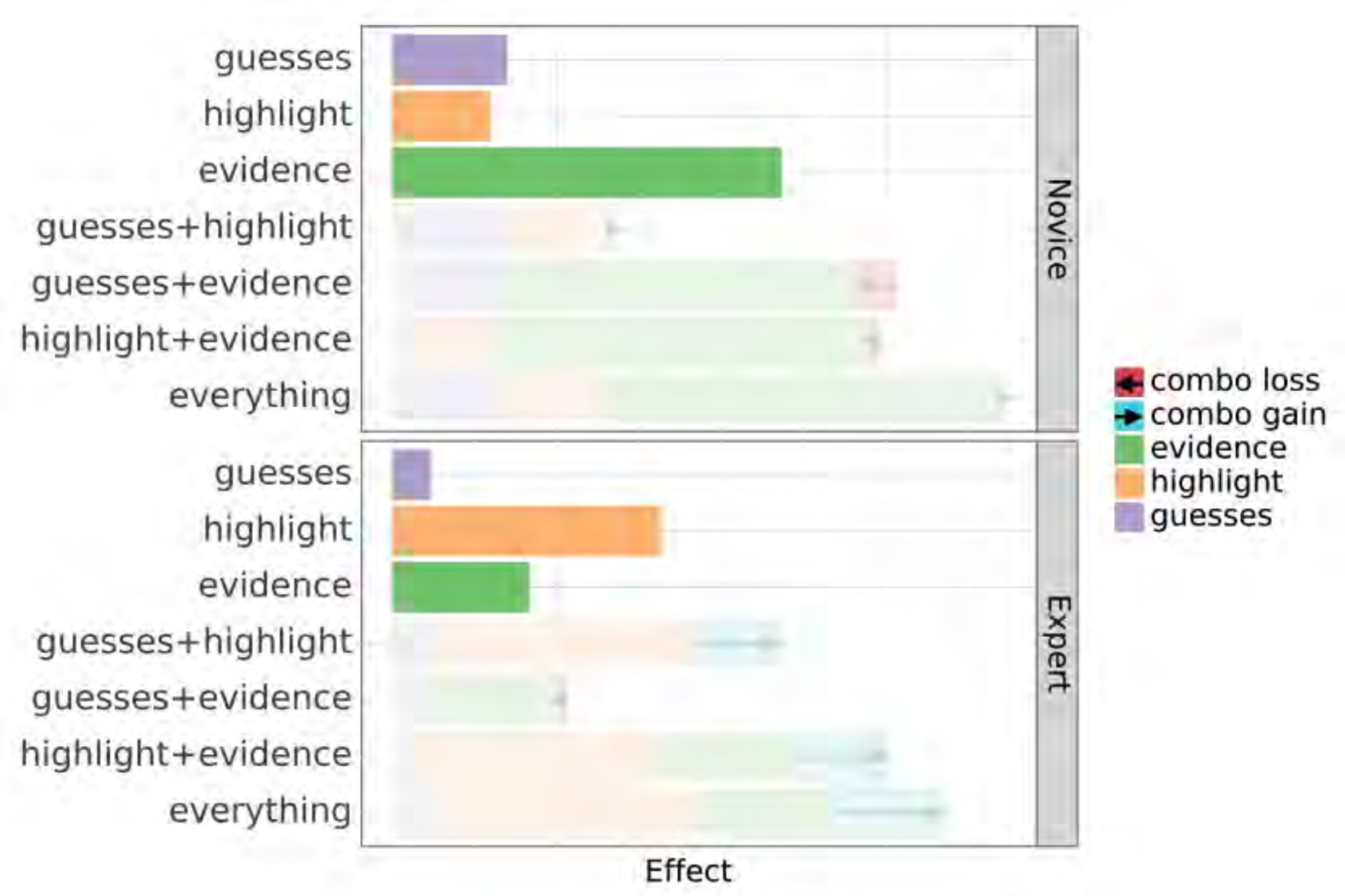
Rerun IRT analysis: rather than contrast human vs. computer skill, we see how much interface boosts (or diminishes) user skill.

Skill Boost

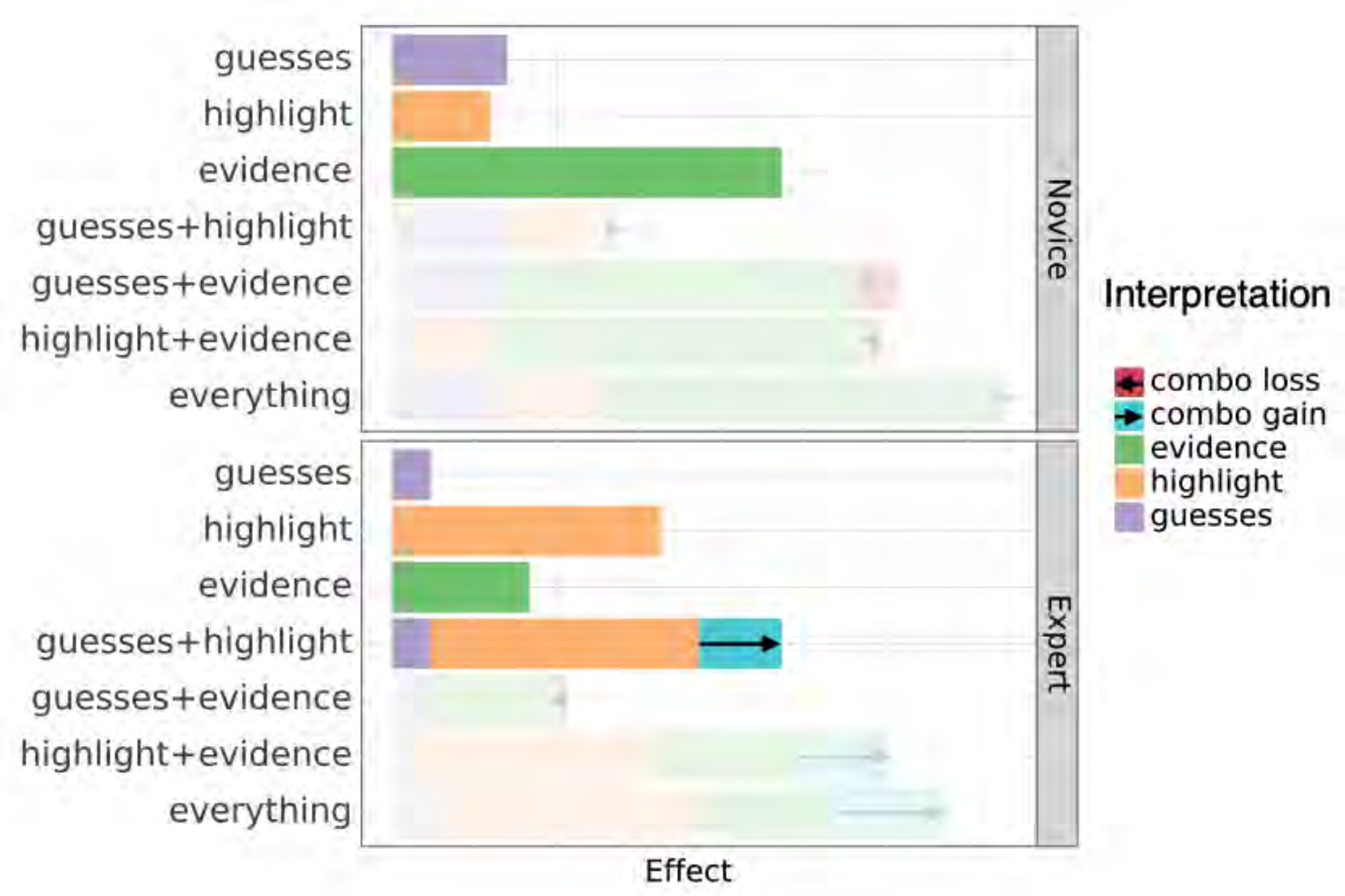
Rerun IRT analysis: rather than contrast human vs. computer skill, we see how much interface boosts (or diminishes) user skill.

Coefficients tell story!

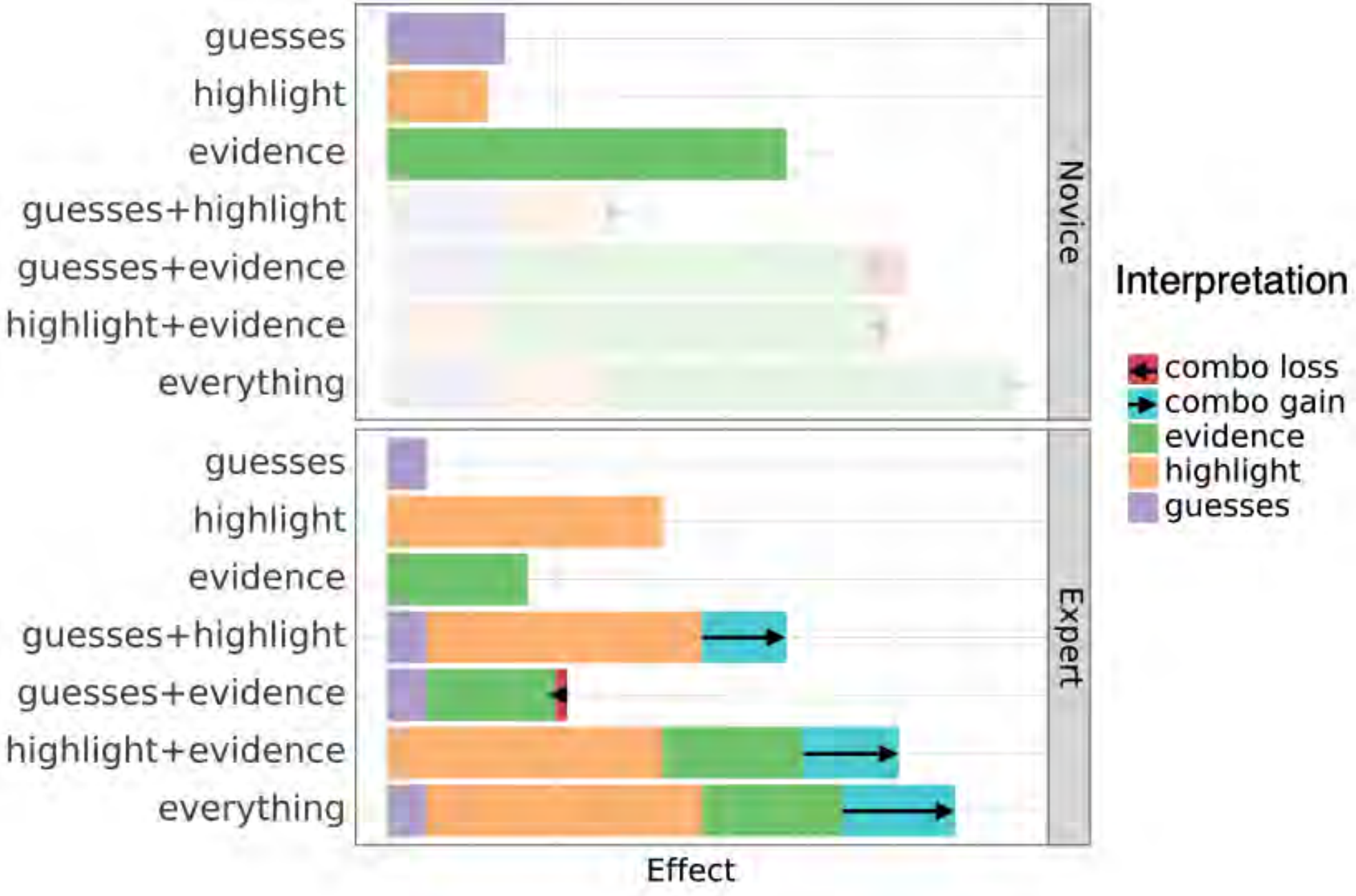
- **Big, Positive**: Help
- **Big, Negative**: Hurt
- **Small**: Neutral



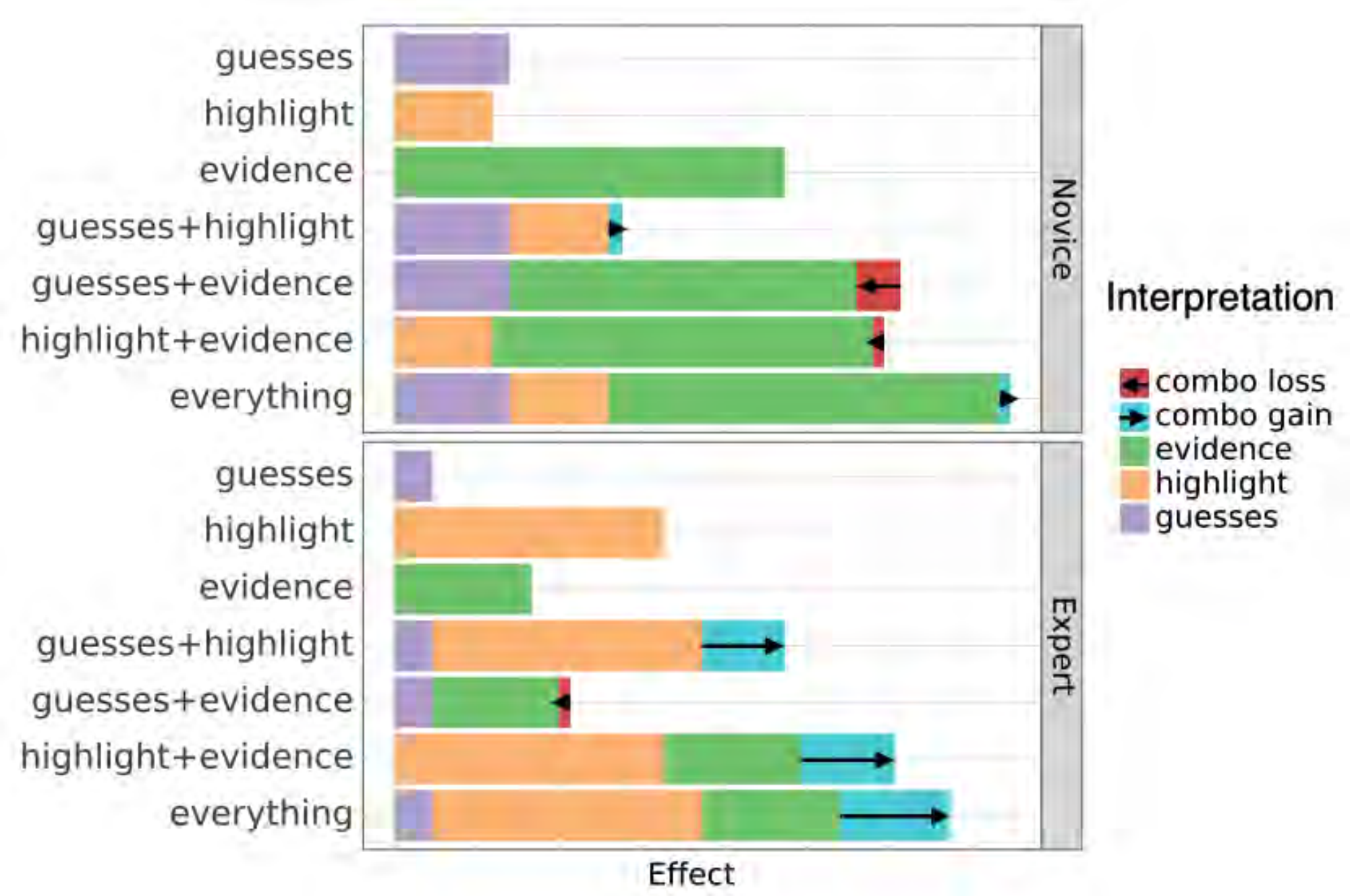
Everything helps: Evidence for novies, Highlight for experts



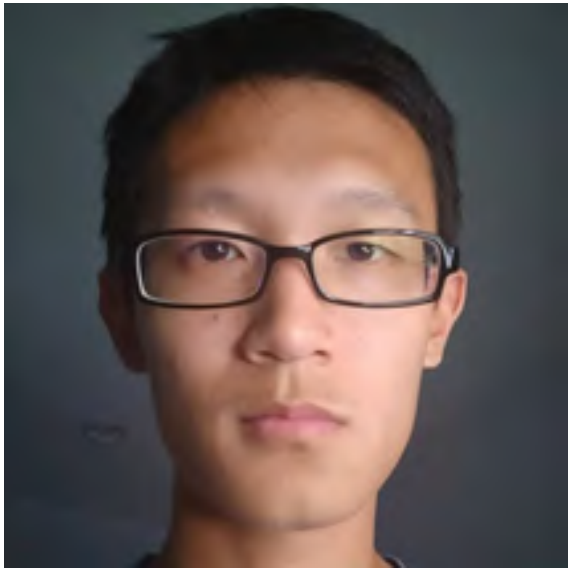
Synergistic effects



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

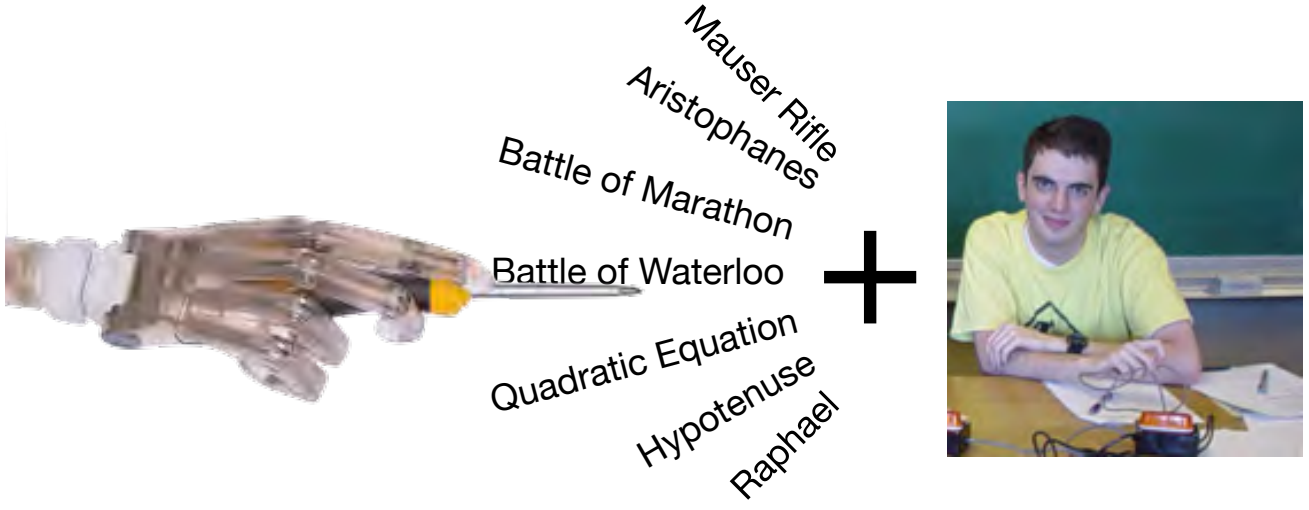
Mauser Rifle
Aristophanes
Battle of Marathon
Battle of Waterloo
Quadratic Equation
Hypotenuse
Raphael

+

Solo

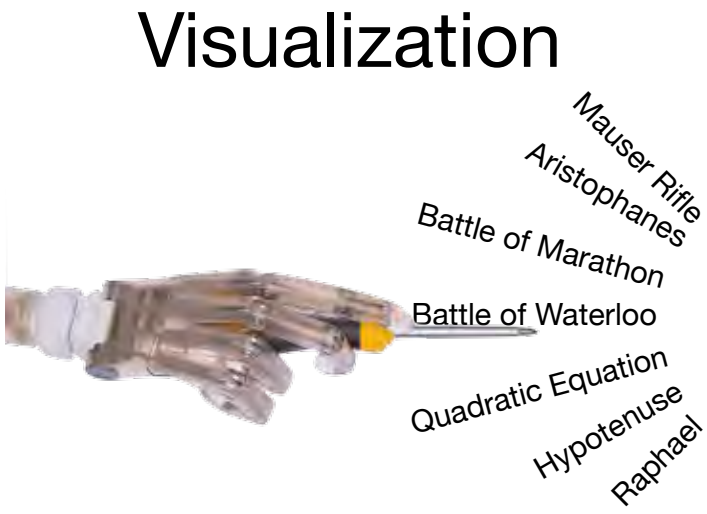
6

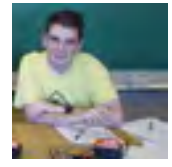
Measuring Interpretability



Exp A	Exp B	Exp C	Solo
5.5	7.3	6.1	6

Improvement through Bandit Algorithms

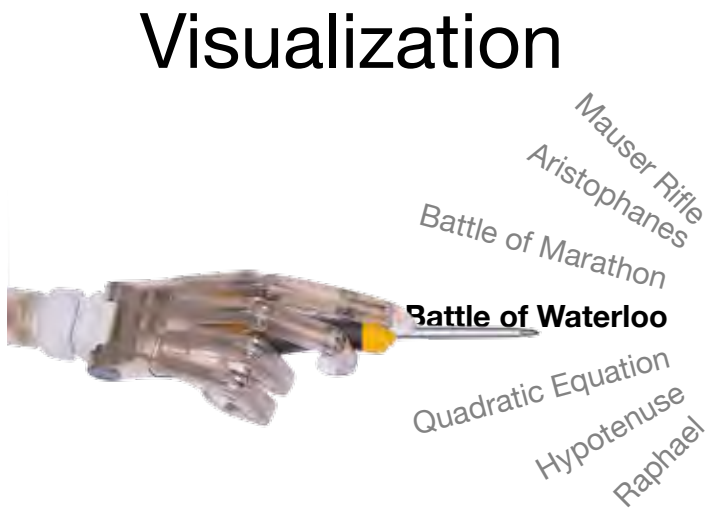


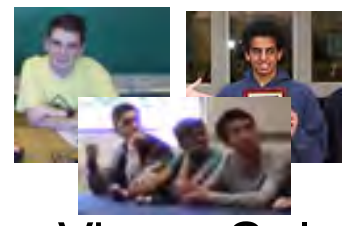


Viz	Solo
7.3	6

Improvement through Bandit Algorithms

Improvement through Bandit Algorithms

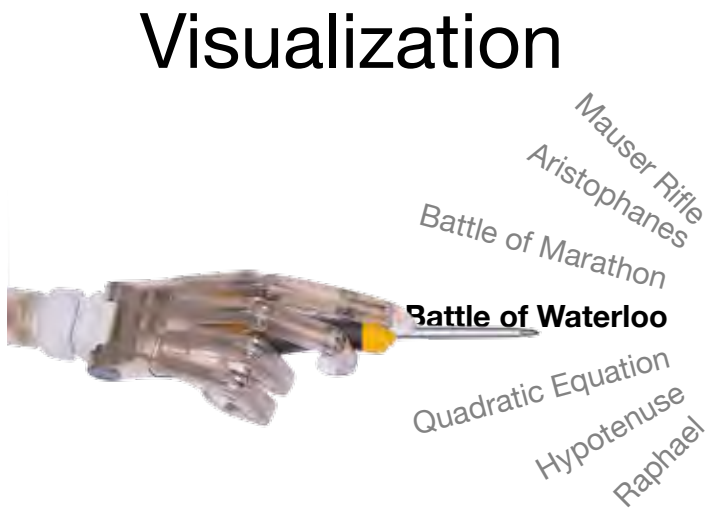


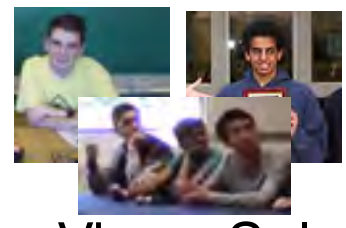


Viz Solo

7.3 6

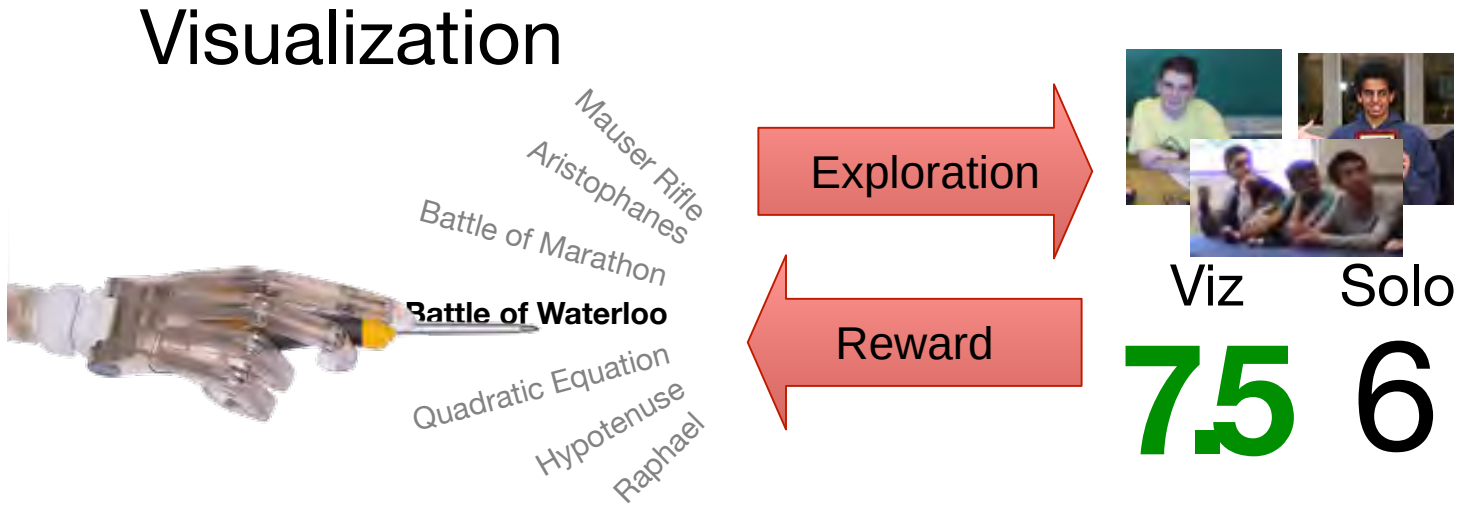
Improvement through Bandit Algorithms



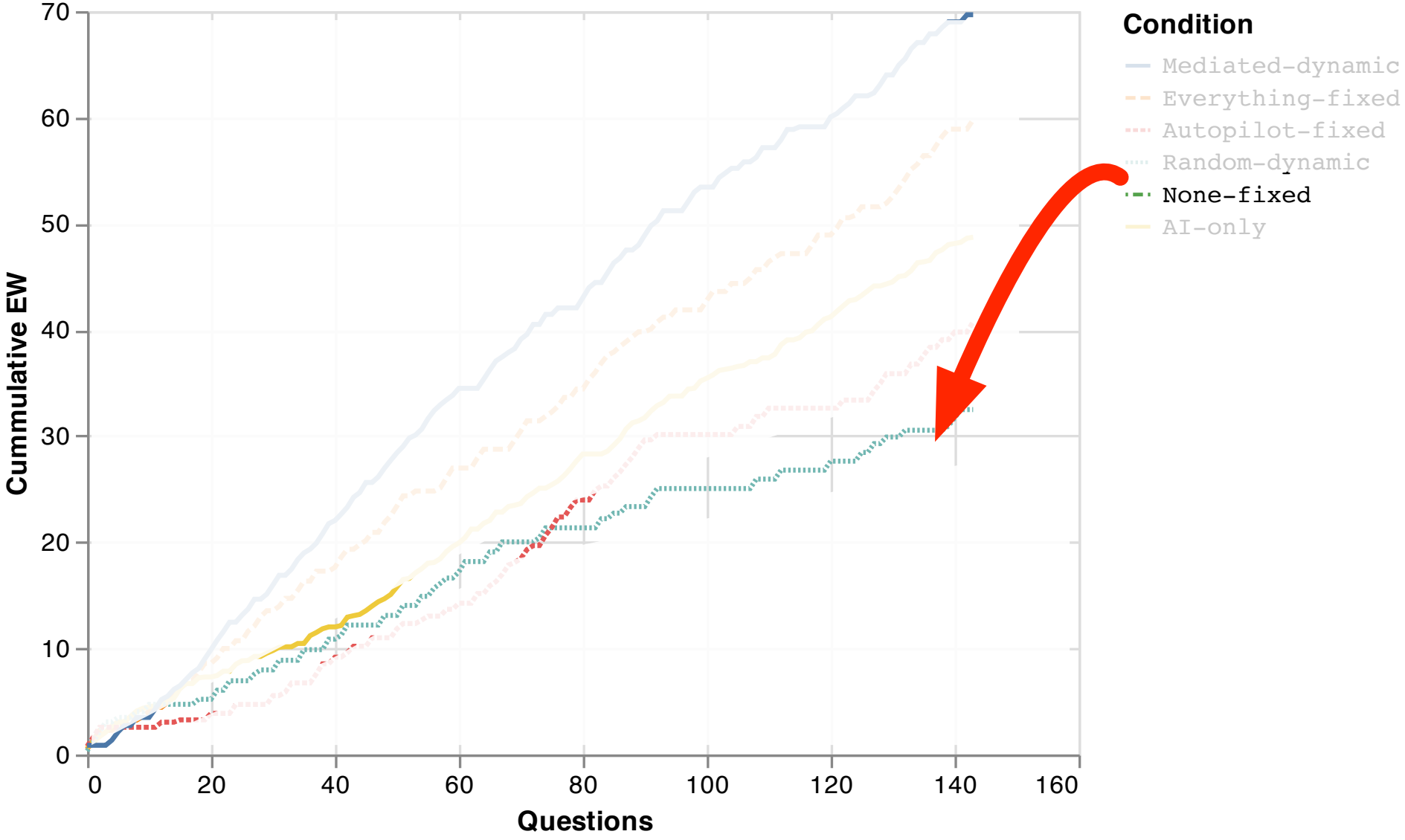


Viz	Solo
7.5	6

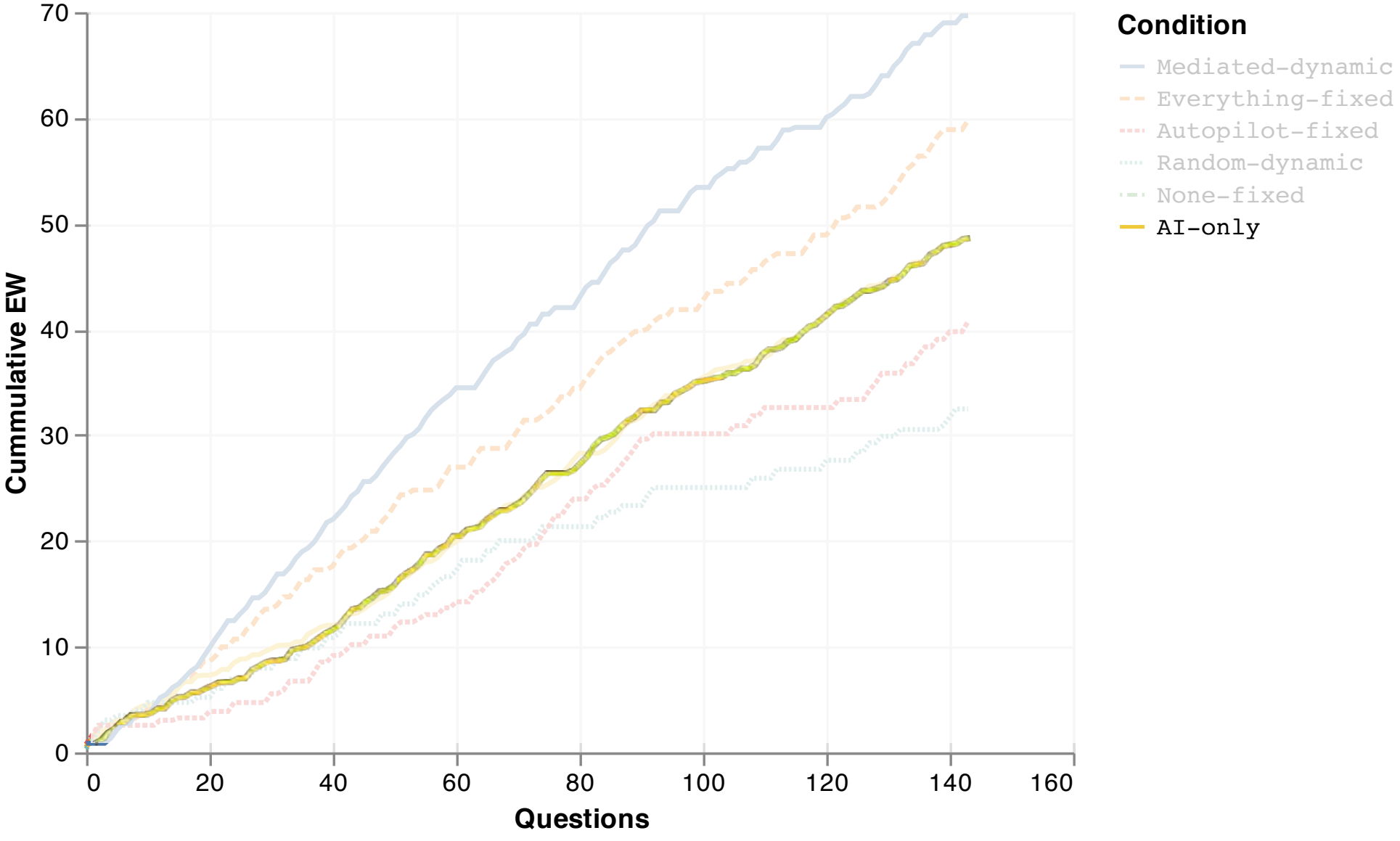
Improvement through Bandit Algorithms



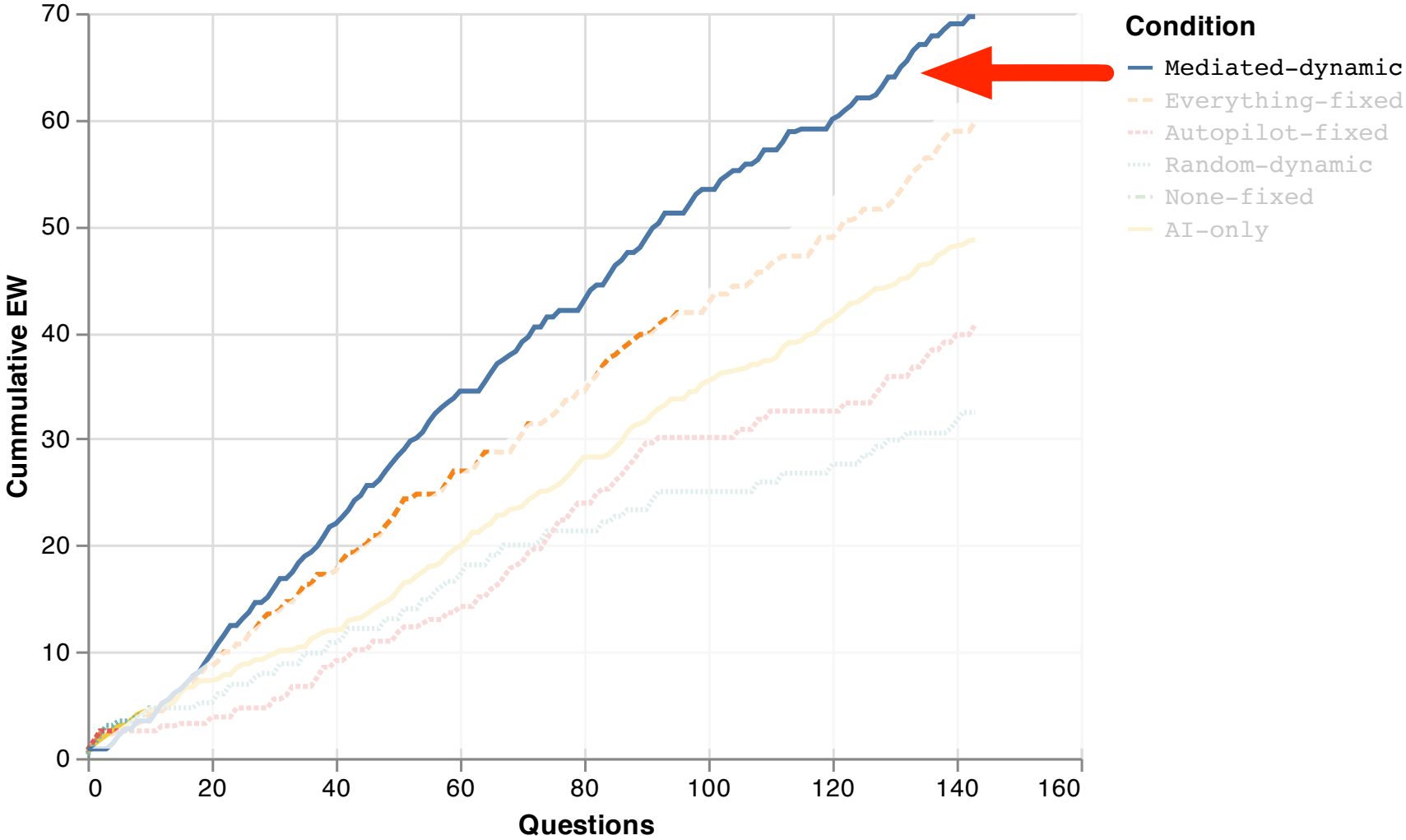
Bandit actions [**?**]: turn each of the explanations (Guess, Highlight, Evidence) on or off.



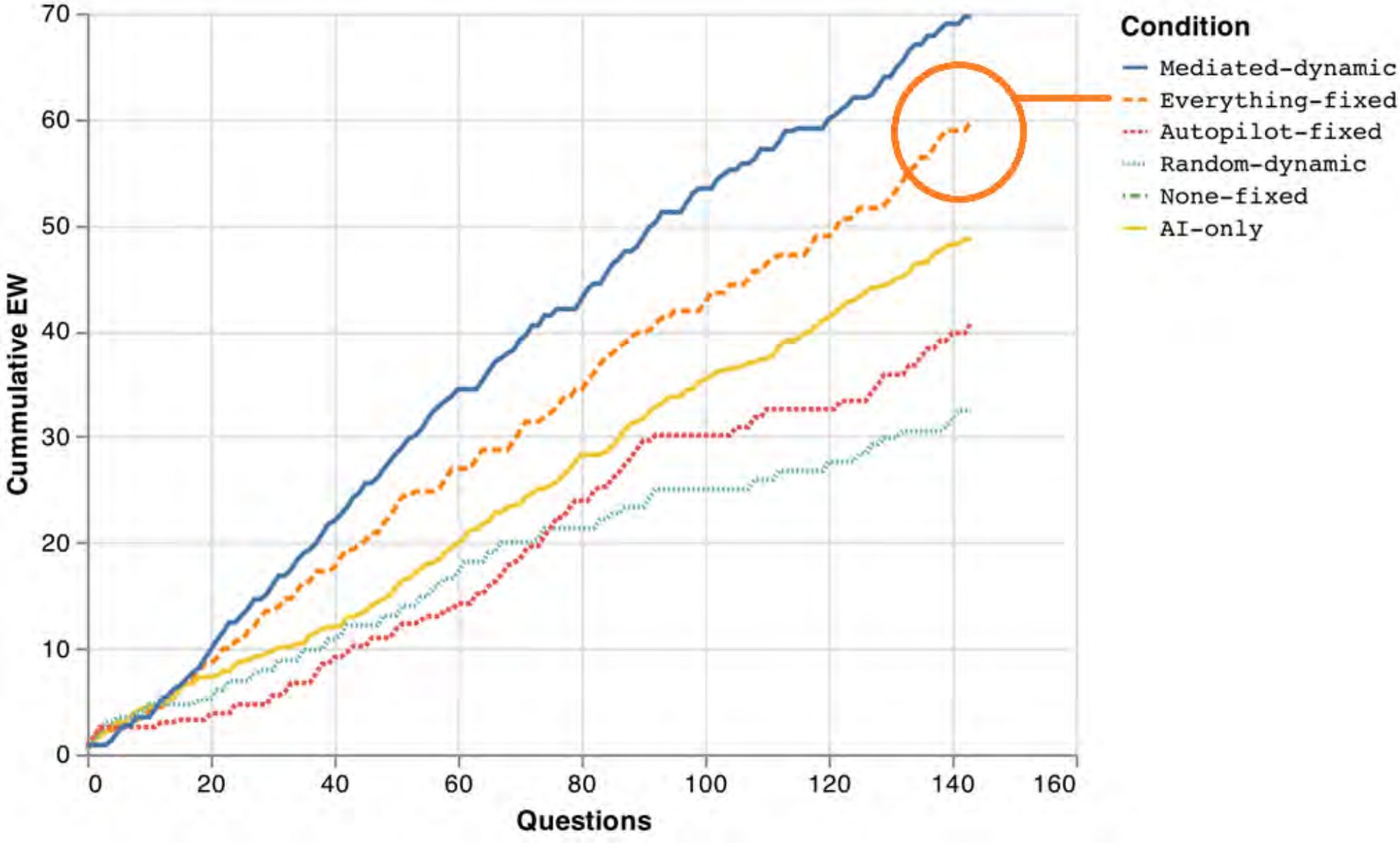
Human alone without an AI teammate



AI alone without a human teammate

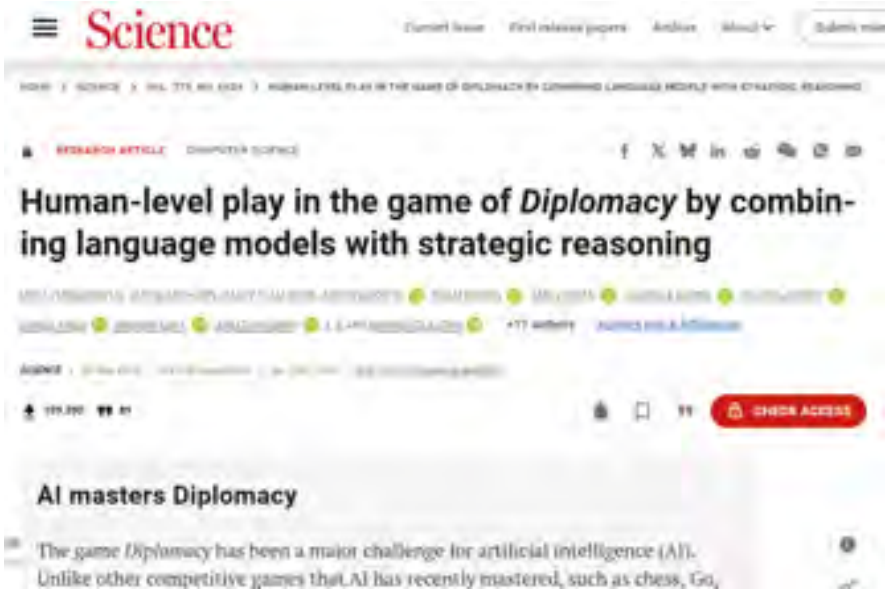


Dynamic assistance to human



Better than showing everything!

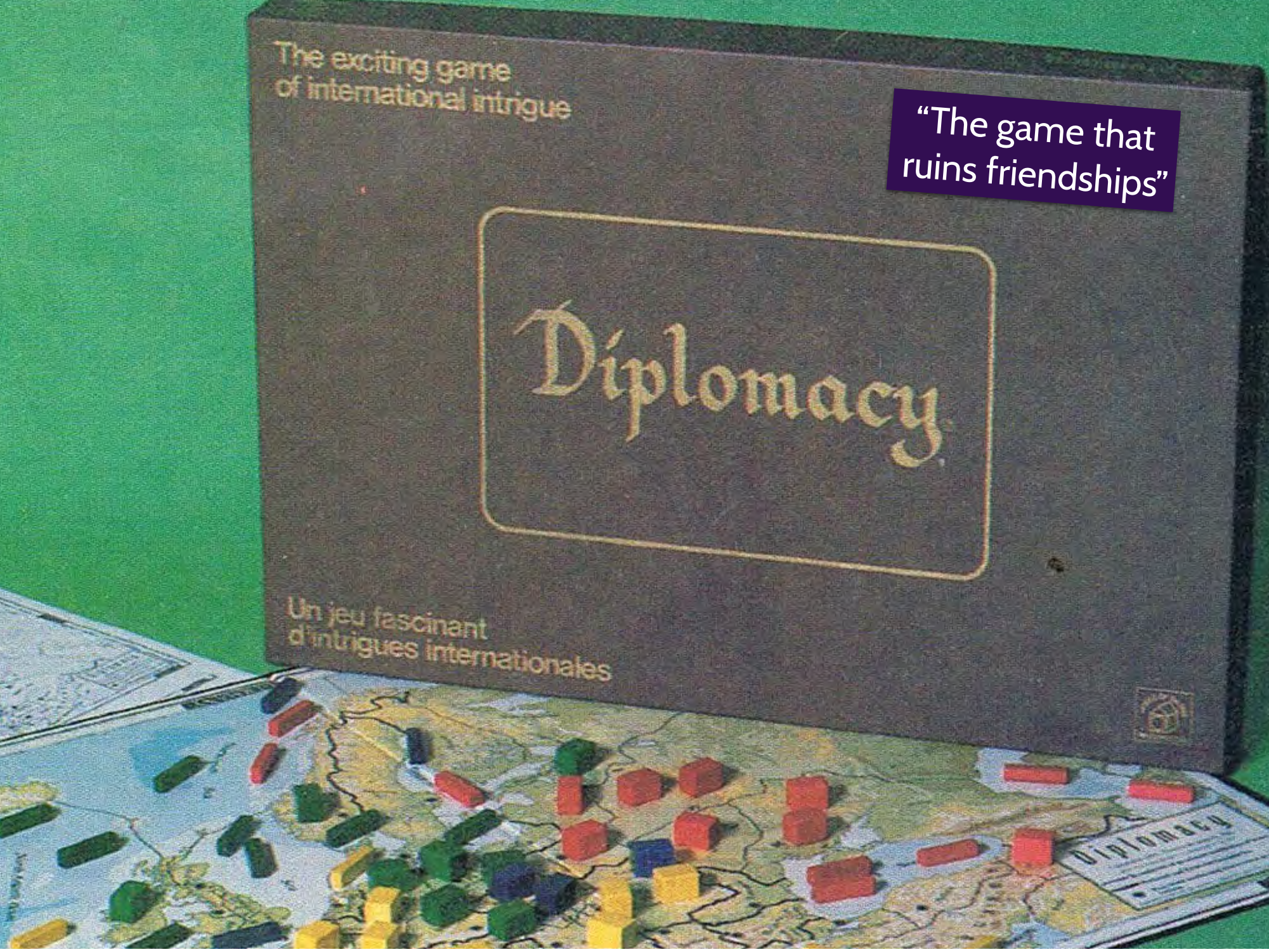
Changing Gears



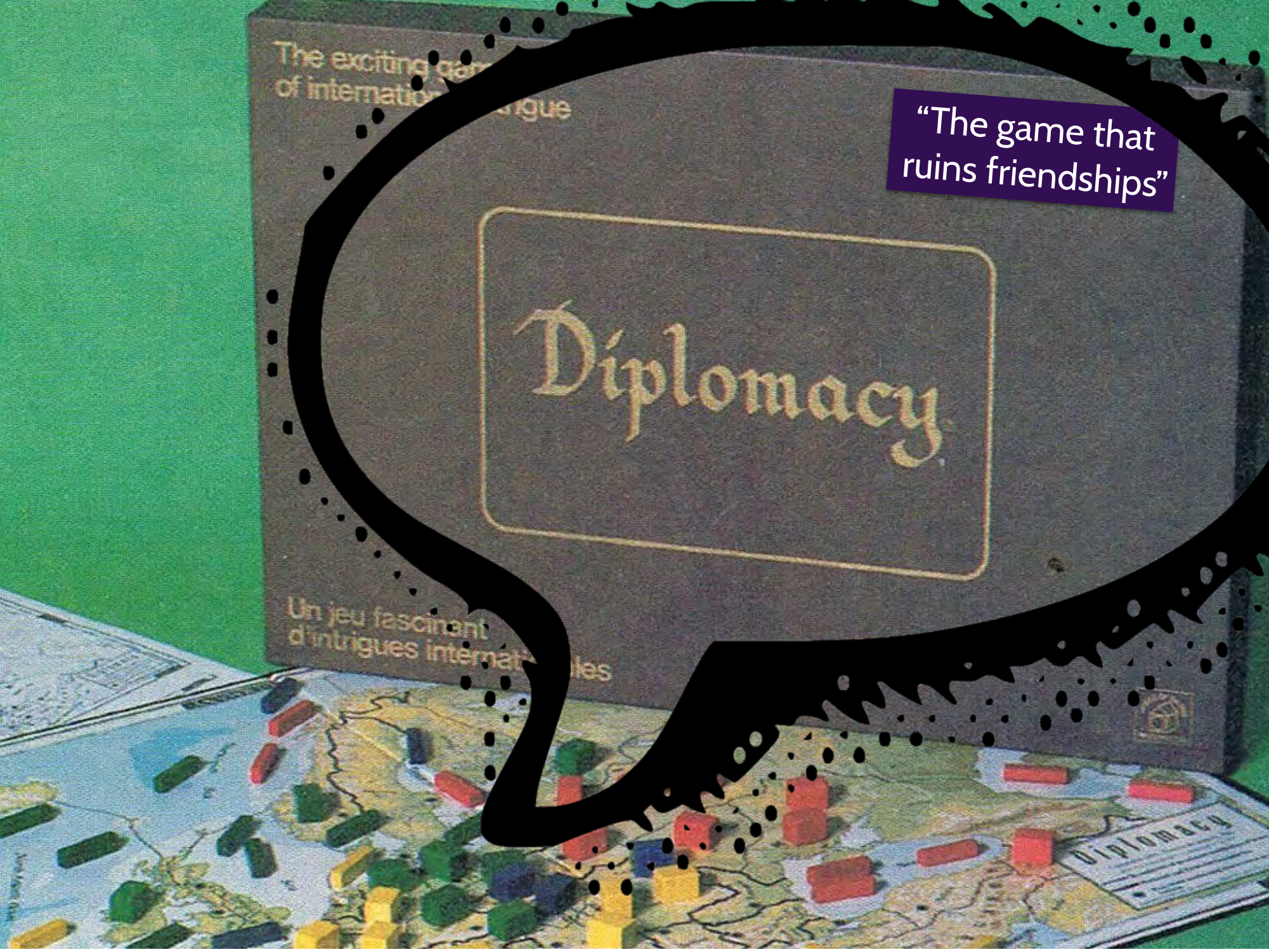
- Focusing on interpersonal relationships
  - Cooperation
  - Deception
  - Betrayal
- Multi-turn interactions

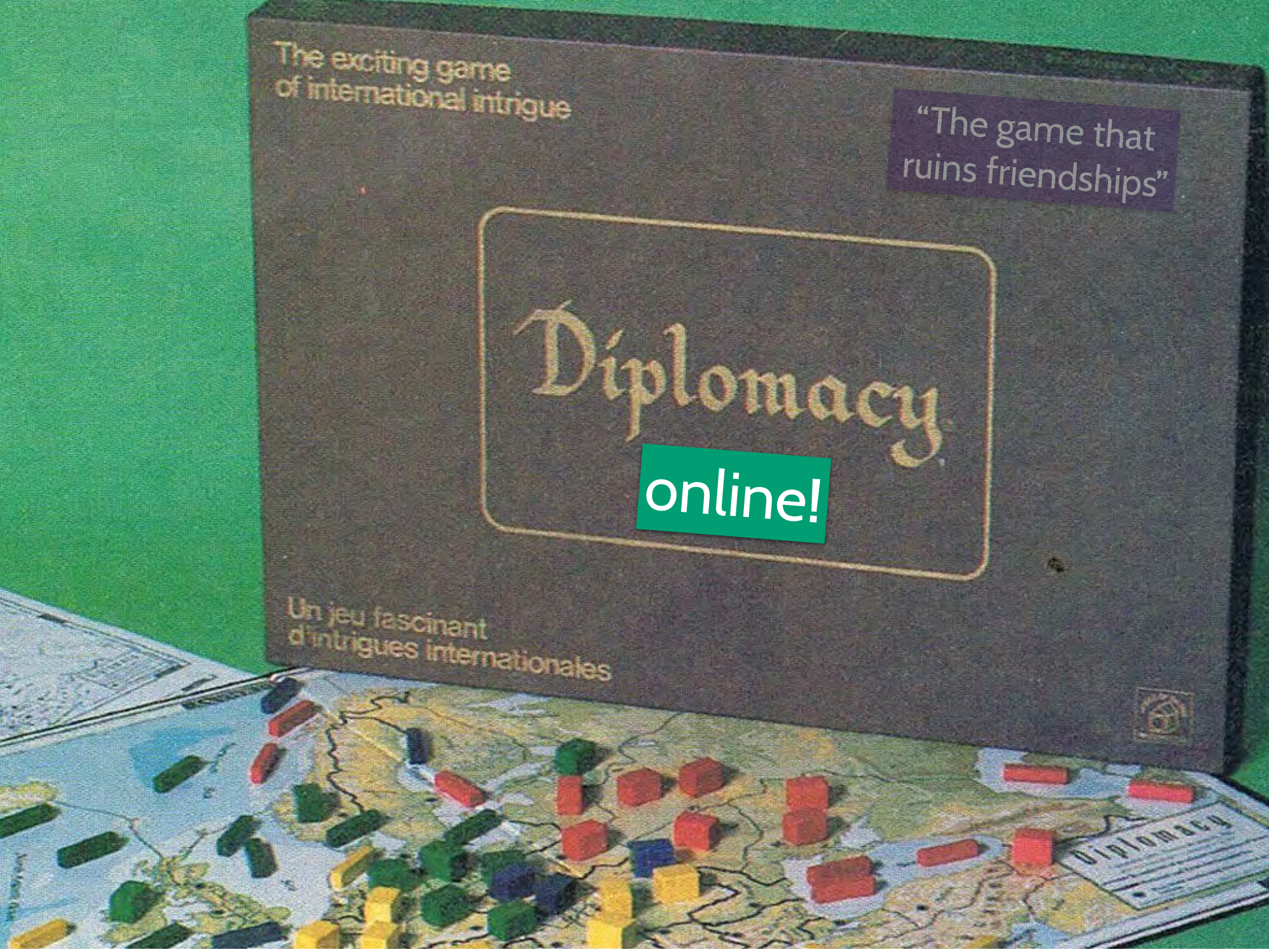
Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. *Association for Computational Linguistics*, 2015


The exciting game
of international intrigue
"The game that
ruins friendships"
Diplomacy
Un jeu fascinant
d'intrigues internationales


The exciting game
of international intrigue
"The game that
uins friendships"
Un jeu fascinant
d'intrigues interna
ies
Diplomacy


"The game that ruins friendships"
Diplomacy


The exciting game
of international intrigue
"The game that
ruins friendships"
Diplomacy
online!
Un jeu fascinant
d'intrigues internationales


The exciting game
of international intrigue
"The game that
ruins friendships"
Diplomacy
online!
249 games
~6 months/game
145k messages
diplom.org; usak.asciiking.com


Iceland
North Atlantic
NAt
Norwegian Sea
NTH
Sweden
NWY
BOT
Gulf
NC
StP
Siberia
Skagerrak
Denmark
Livonia
LVN
Helgeland Bight
Holland
Kiel
Belgium
Ruhr
Munich
Ukraine
Mid Atlantic
Tyrolia
Budapest
Rumania
Black Sea
EC
Portugal
Spain
NC
SC
GoL
Tuscany
Roma
Tyrrhenian Sea
TYN
West Mediterranean
Greece
Ankara
Armenia
Smyrna
Syria


Berlin
Bohemia
Galicia
Ukraine
Vienna
Budapest
Trieste
Rumania
Serbia
Bulgaria
EC
Adriatic


Kiel
Berlin
Ruhr
Munich
Bohemia
Ukraine
Galicia
Vienna
Budapest
Switz.
Tyrolia
Trieste
Venezia
Tuscany
Serbia
Rumania
Bulgaria
EC


Kiel
Berlin
Ruhr
Munich
Bohemia
Ukraine
Galicia
Vienna
Budapest
Switz.
Tyrolia
Trieste
Venezia
Tuscany
Serbia
Rumania
Bulgaria
EC


Kiel
Berlin
Ruhr
Galicia
Vienna
Budapest
Switz.
Tyrolia
Trieste
Venezia
Tuscany
Serbia
Rumania
Bulgaria
EC


Kiel
Berlin
Ruhr
Munich
Bohemia
Galicia
Ukraine
help?
Switz.
Tyrol
Budapest
Trieste
Venezia
Tuscany
Rumania
Serbia
Bulgaria
EC

F

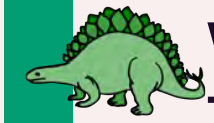
What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F

What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

F

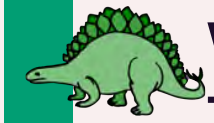


What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

F

F

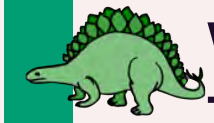


What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

F

F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

E

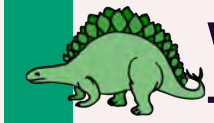
...

stabs

!


NOW STAND BACK,
I GOTTA PRACTICE MY STABBIN'

F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

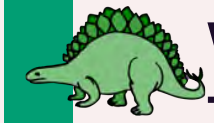
E

...

stabs

!

F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

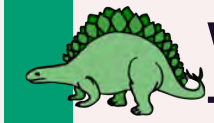
E

...

stabs

!

F



What I would like you to do is keep Turkey busy and somehow get Russia and Turkey to engage. Meanwhile we need to take VIE, suggest you support me in there.

F

It's a sensible plan. I'll support you as requested. Please be sure to simultaneously attack SWE.

E

...

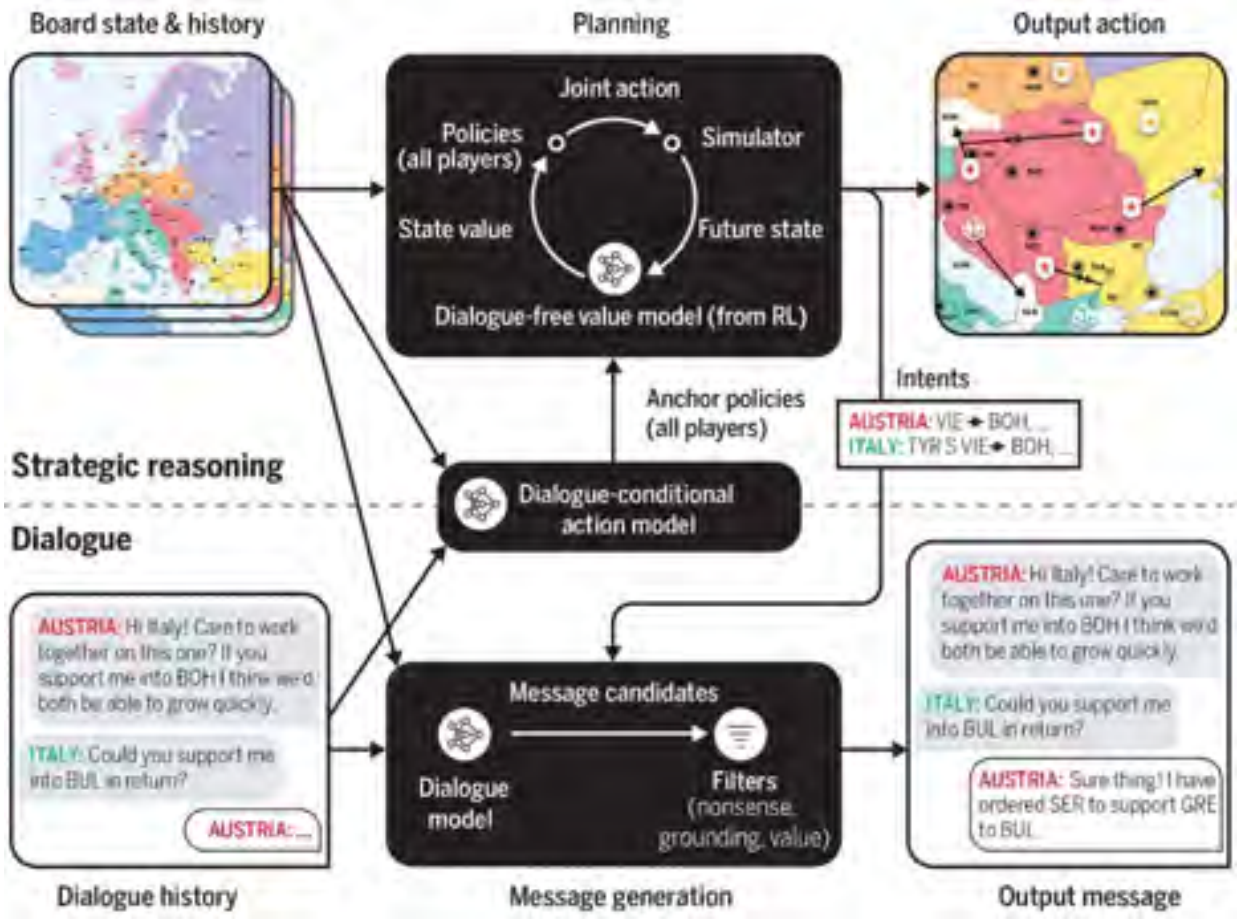
stabs

!

Not really sure what to say, except that I regret you did what you did.

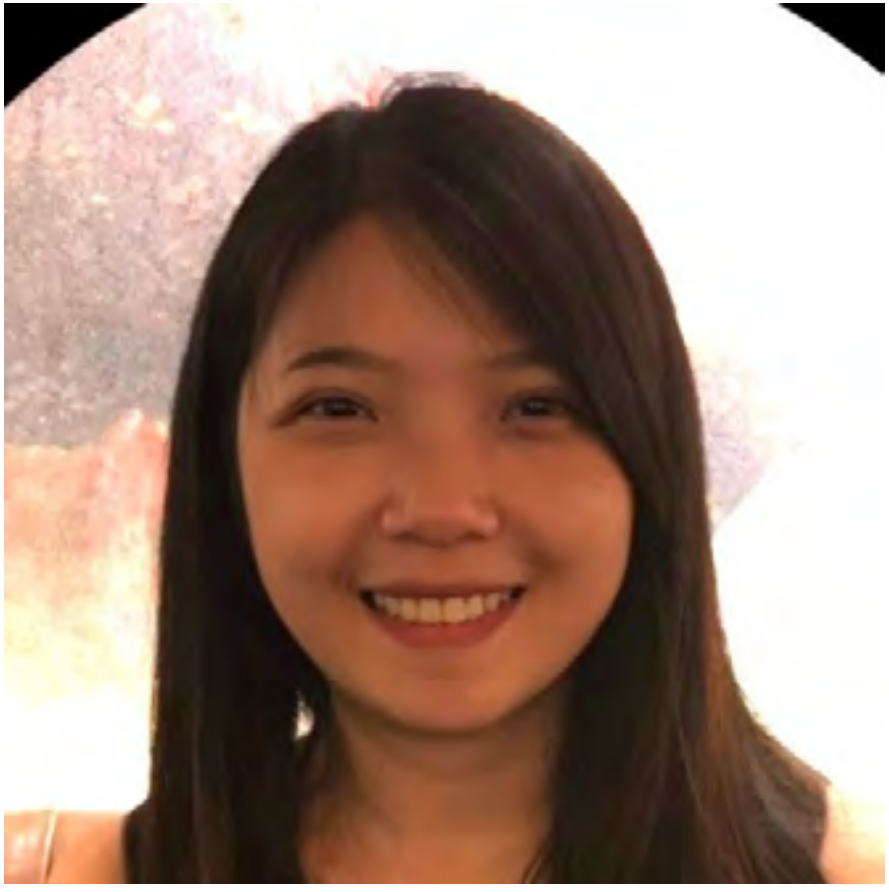

Curse your sudden
but inevitable
betrayal!

Cicero is really good



Rank	Average Score	# Games Played
1	0.4918	1
2	0.4912	1
3	0.4528	1
4	0.4516	1
5	0.3734	1
6	0.3505	11
7	0.3491	1
8	0.3035	4
9	0.2754	2
10	**0.2583**	**40**
11	0.2450	6
12	0.2311	1
13	0.2277	2

And we confirm their results, it wins around 90% of games!



More Victories, Less Cooperation: Assessing Cicero's Diplomacy Play

Wichayaporn Wongkamjan[1] **Feng Gu**[1] **Yanze Wang**[1] **Ulf Hermjakob**[4]
Jonathan May[4] **Brandon M. Stewart**[2] **Jonathan K. Kummerfeld**[3]
Denis Peskoff[2] **Jordan Lee Boyd-Graber**[1]
[1]University of Maryland [2]Princeton University [3]University of Sydney
[4]Information Sciences Institute, University of Southern California
{wongkam,fgu1}@umd.edu {yanzewan,ulf,jonmay}@isi.edu
bms4@princeton.edu jonathan.kummerfeld@sydney.edu.au
dp7896@princeton.edu jbg@umiacs.umd.edu

Can they tell it's a bot?

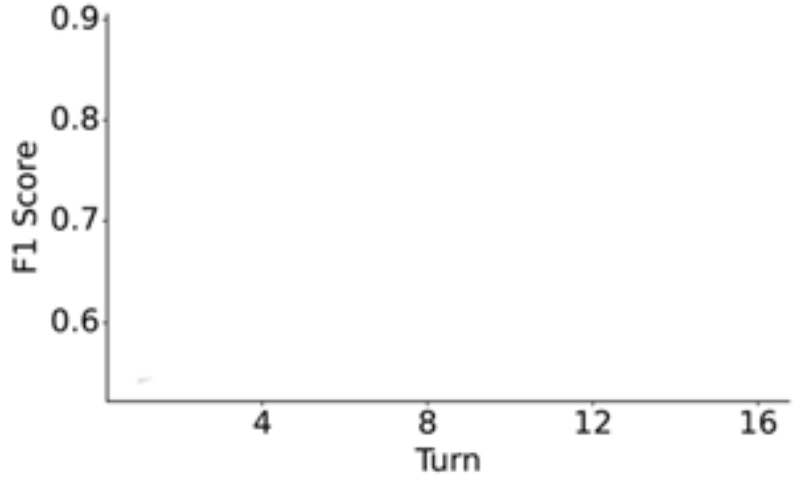
Passing the Turing Test?

No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.

Can they tell it's a bot?

Passing the Turing Test?

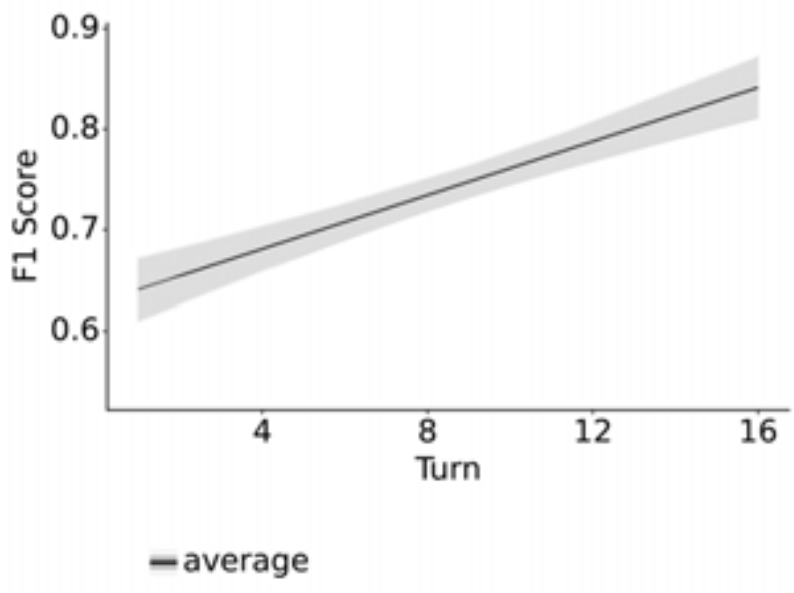
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



Can they tell it's a bot?

Passing the Turing Test?

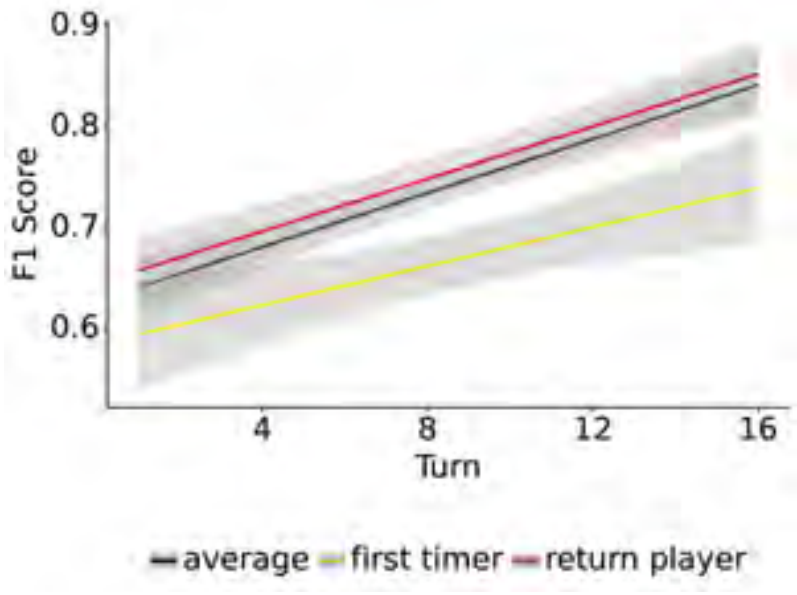
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



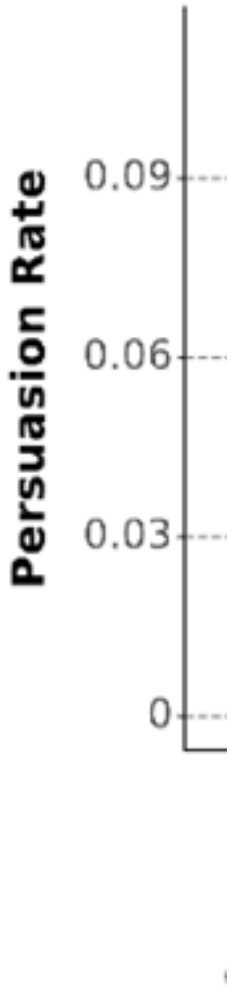
Can they tell it's a bot?

Passing the Turing Test?

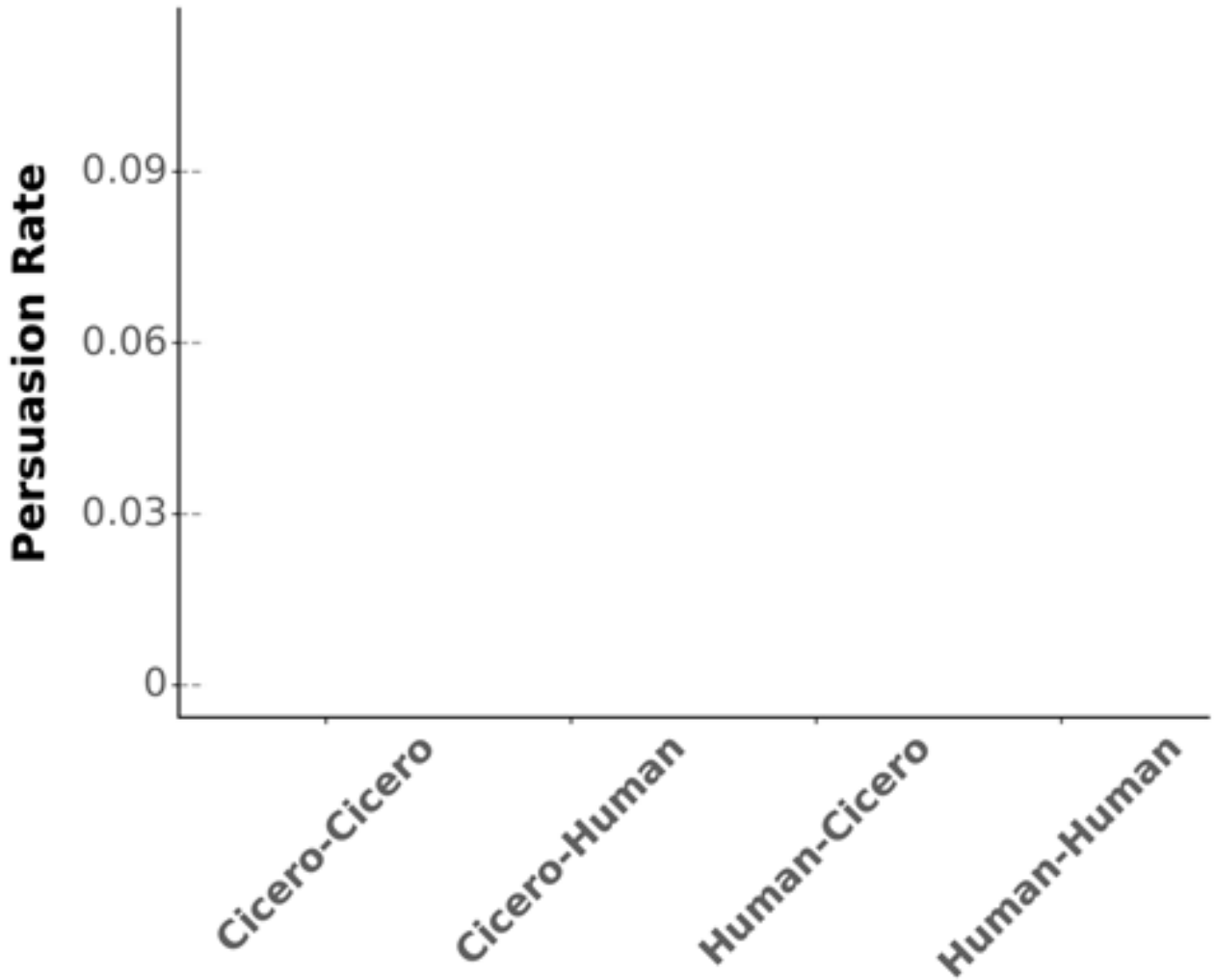
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



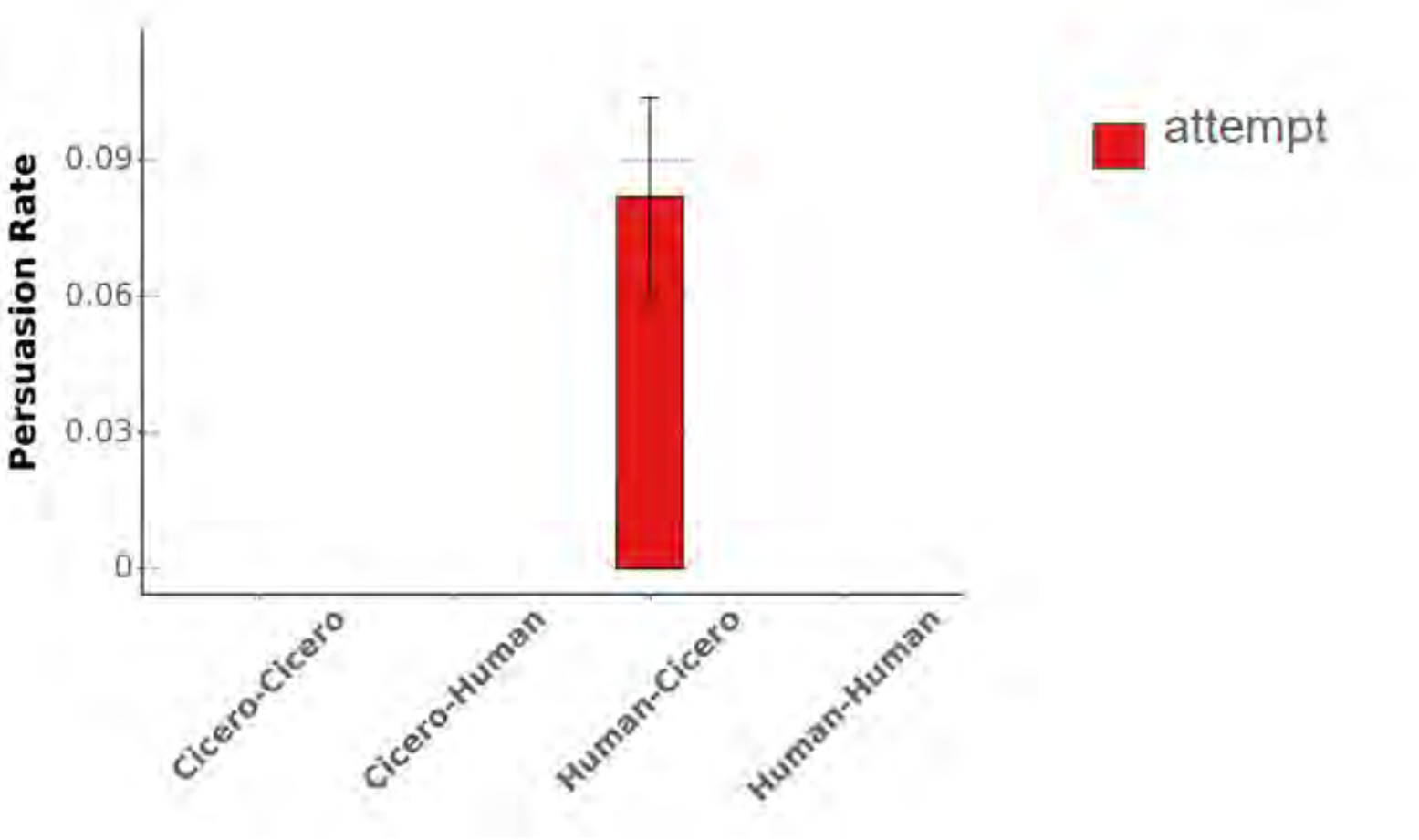
Persuasion



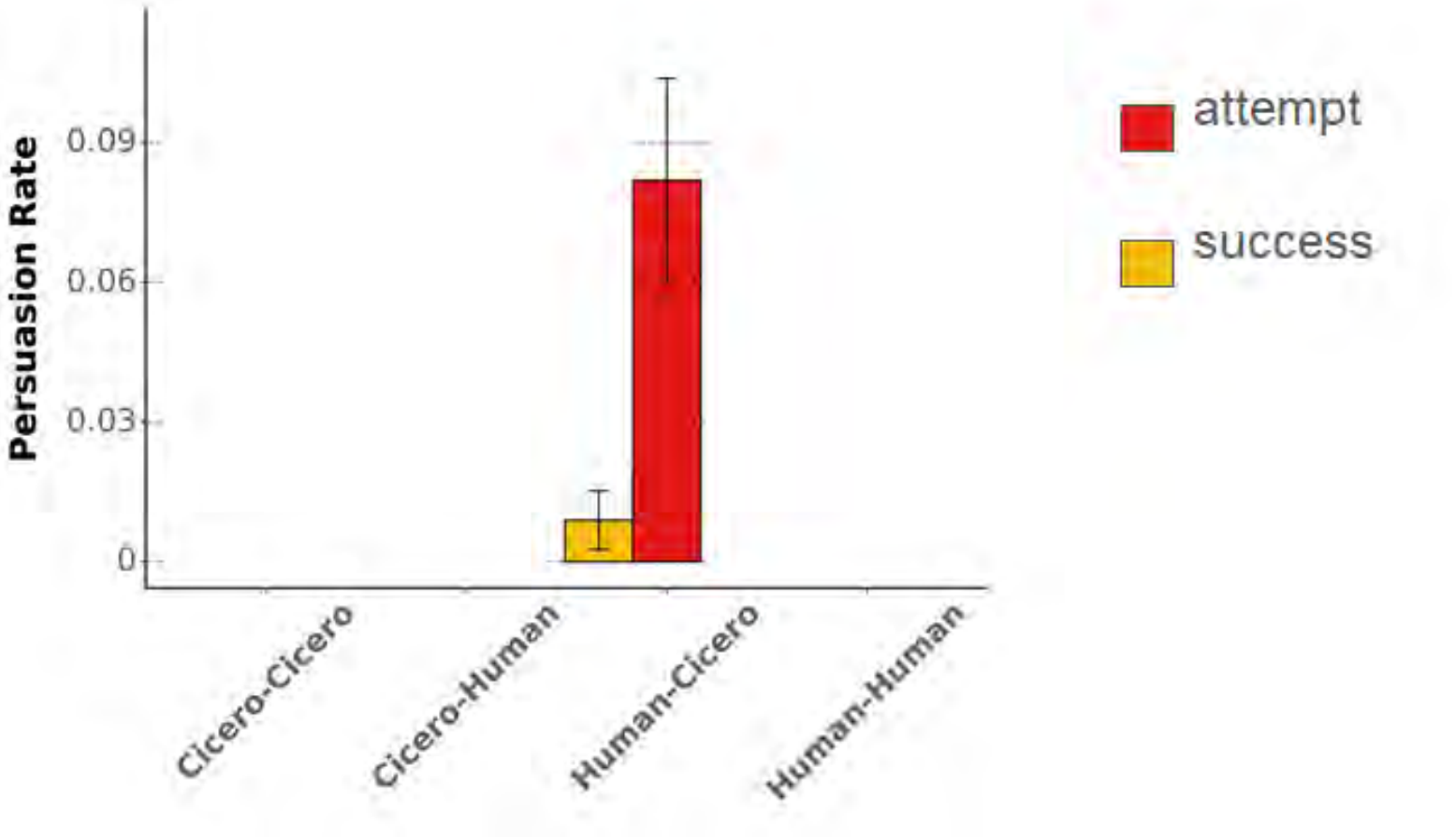
Persuasion


Persuasion Rate
0.09
0.06
0.03
0
Cicero-Cicero
Cicero-Human
Human-Cicero
Human-Human

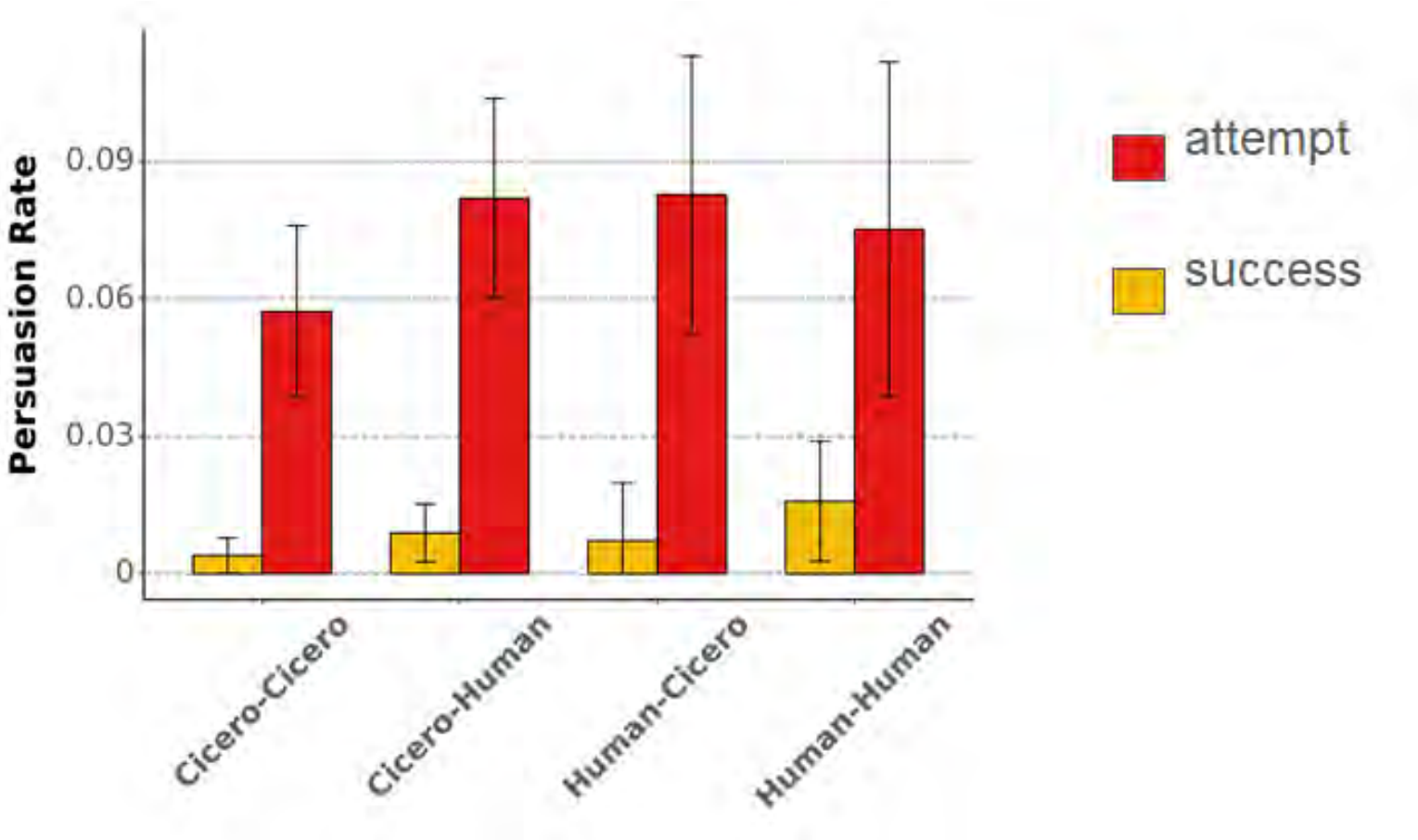
Persuasion


Persuasion Rate
0.09
0.06
0.03
0
Cicero-Cicero
Cicero-Human
Human-Cicero
Human-Human
attempt

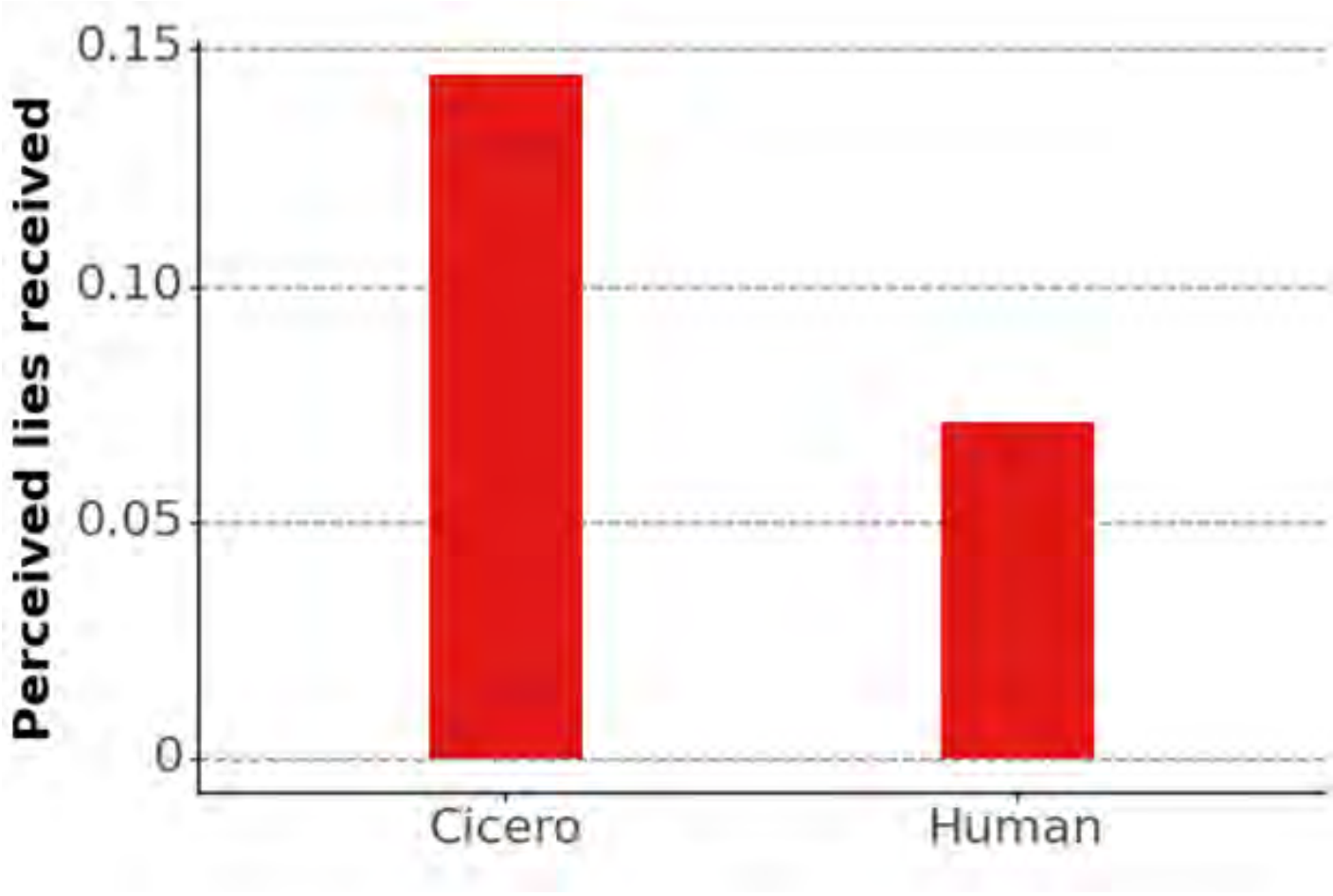
Persuasion


Persuasion Rate
0.09
0.06
0.03
0
Cicero-Cicero
Cicero-Human
Human-Cicero
Human-Human
attempt
success

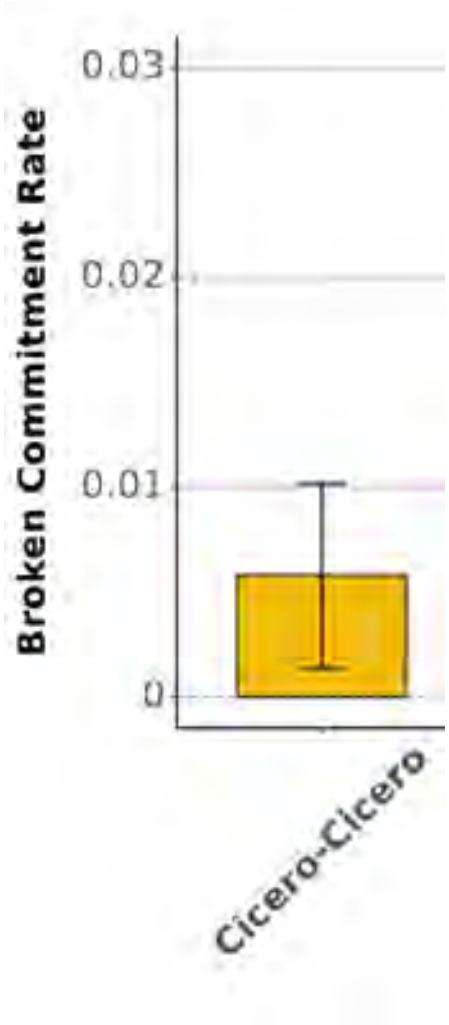
Persuasion


Persuasion Rate
0.09
0.06
0.03
0
Cicero-Cicero
Cicero-Human
Human-Cicero
Human-Human
attempt
success

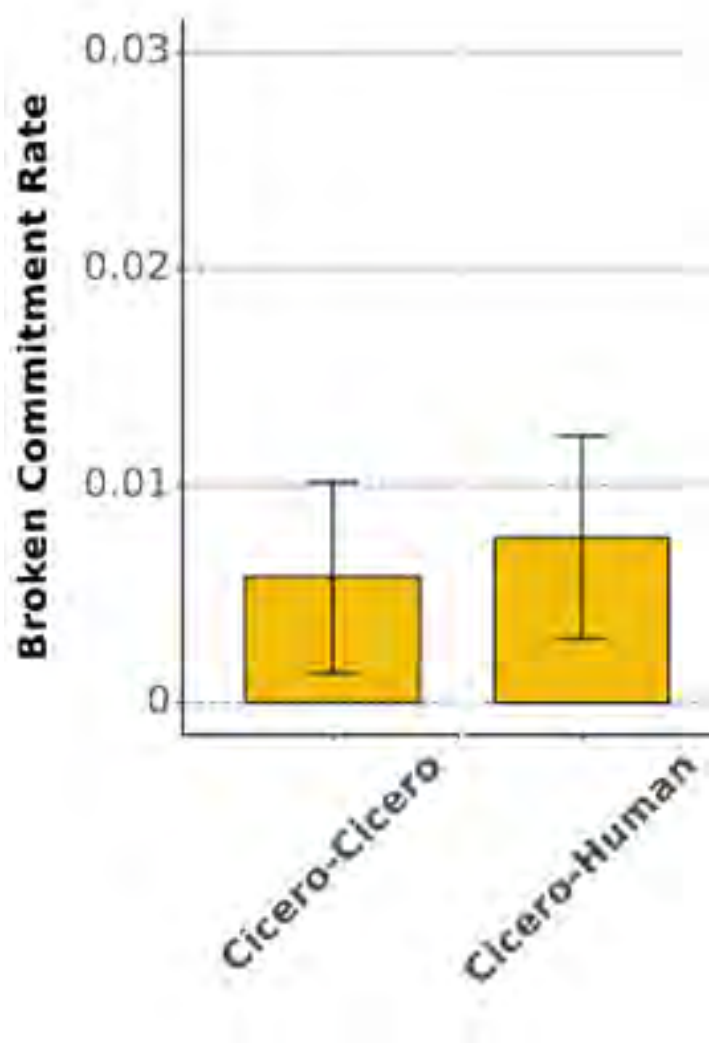
People Think Cicero Lies More



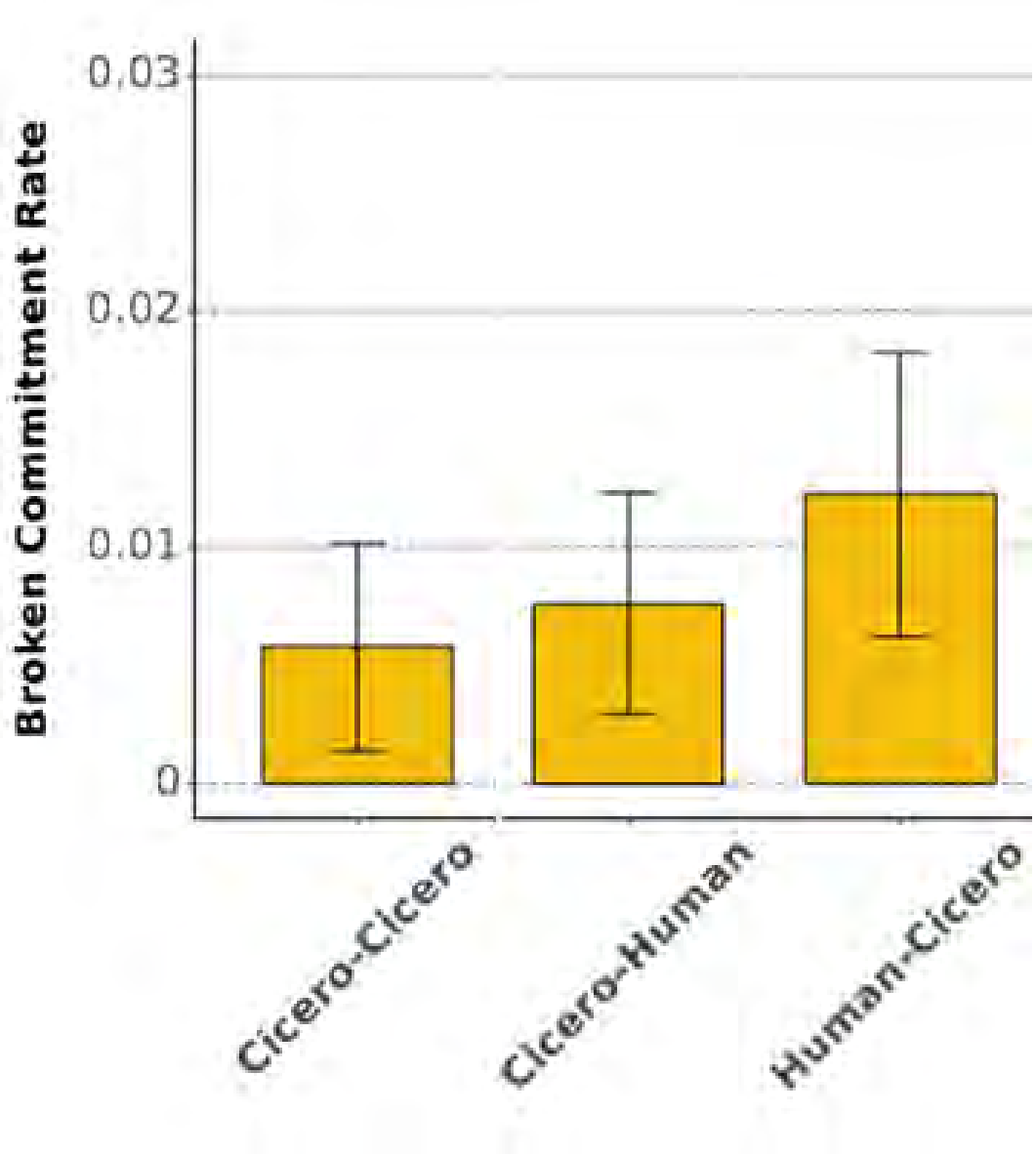
Broken Commitments (subset of deception)



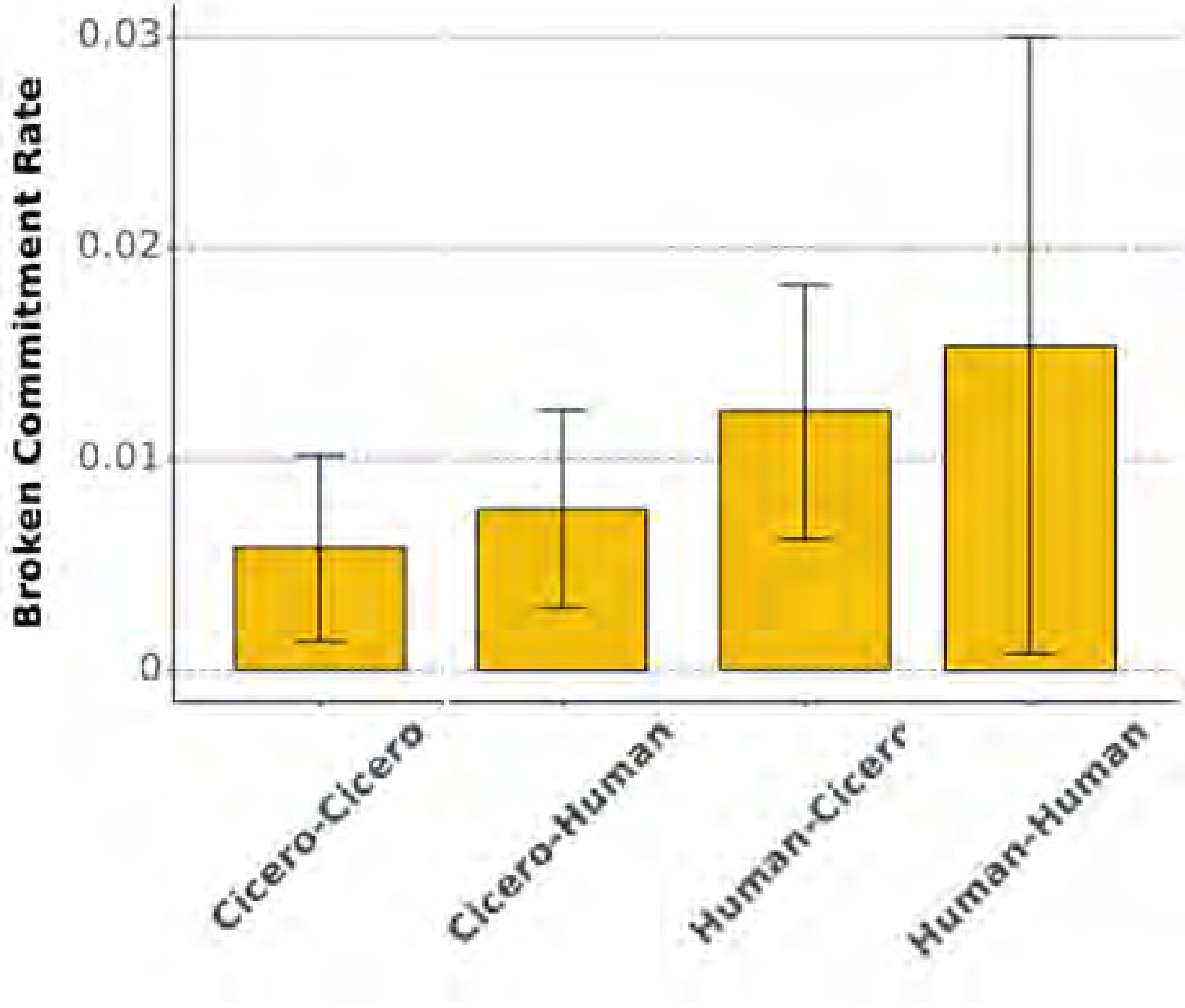
Broken Commitments (subset of deception)



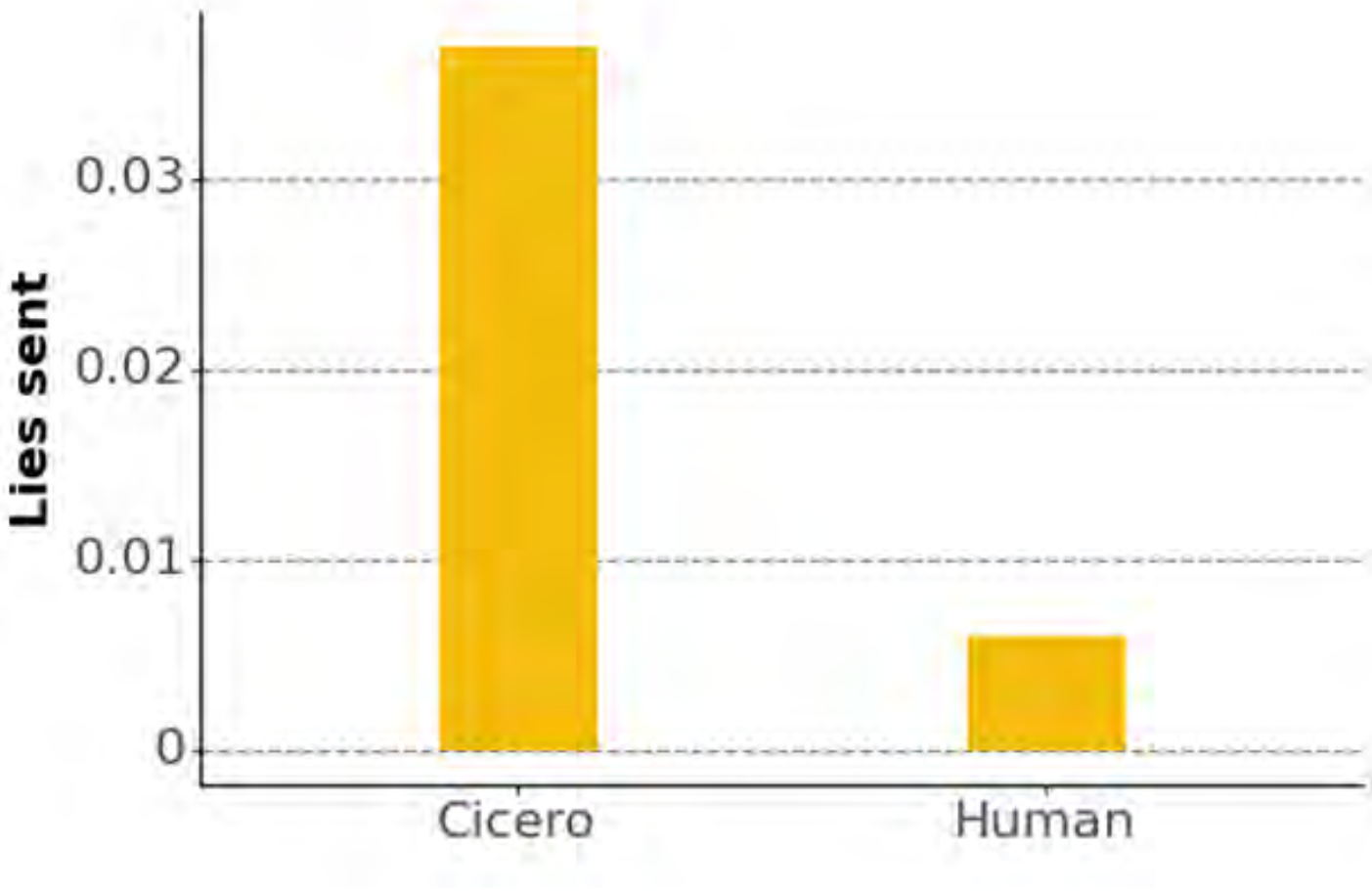
Broken Commitments (subset of deception)



Broken Commitments (subset of deception)



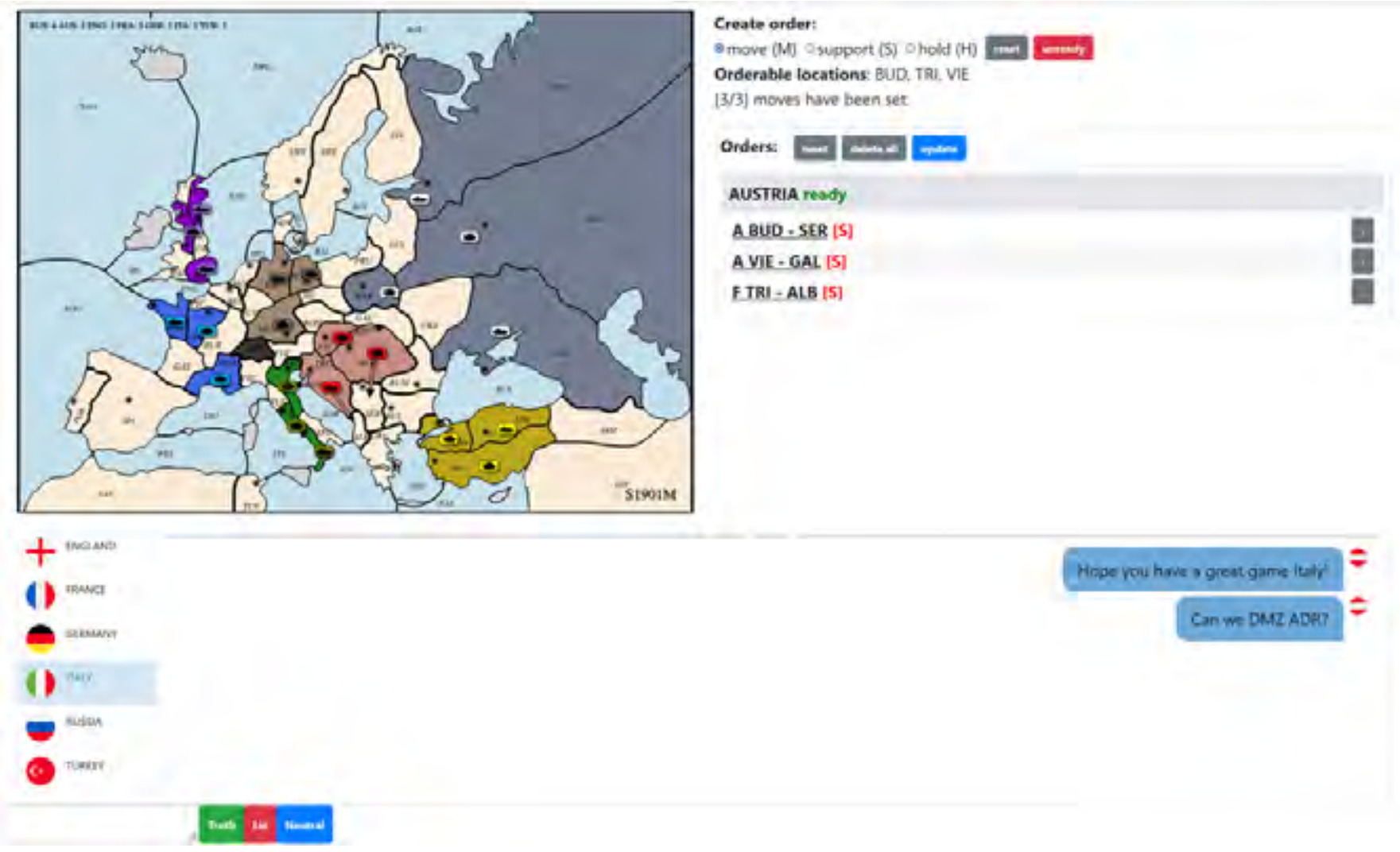
People Lie to Cicero More



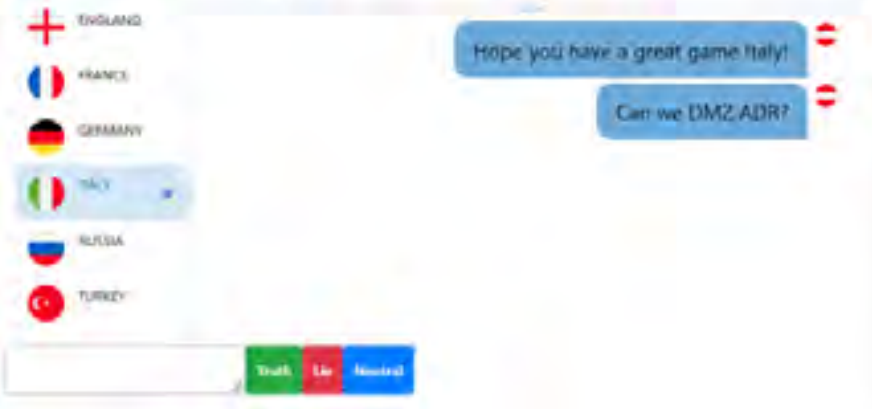
Personalized Help for Optimizing Low-Skilled Users' Strategy

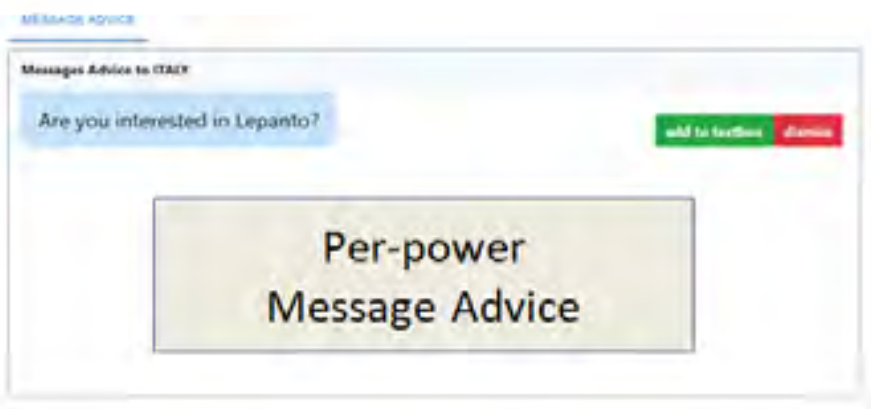
Feng Gu[1] **Wichayaporn Wongkamjan**[1] **Jonathan K. Kummerfeld**[2]
Denis Peskoff[3] **Jonathan May**[4] **Jordan Lee Boyd-Graber**[1]
[1]University of Maryland [2]University of Sydney
[3]Princeton University
[4]Information Sciences Institute, University of Southern California
{fgu1, wwongkam}@umd.edu jbg0.umiacs.umd.edu

Giving Advice

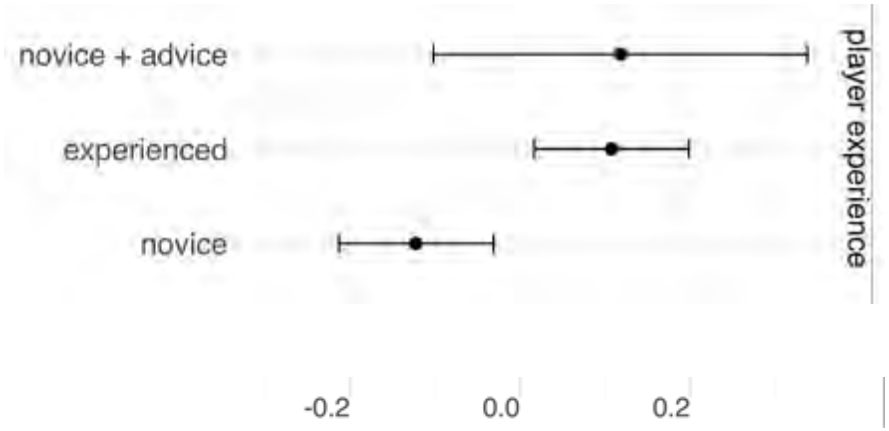


Giving Advice

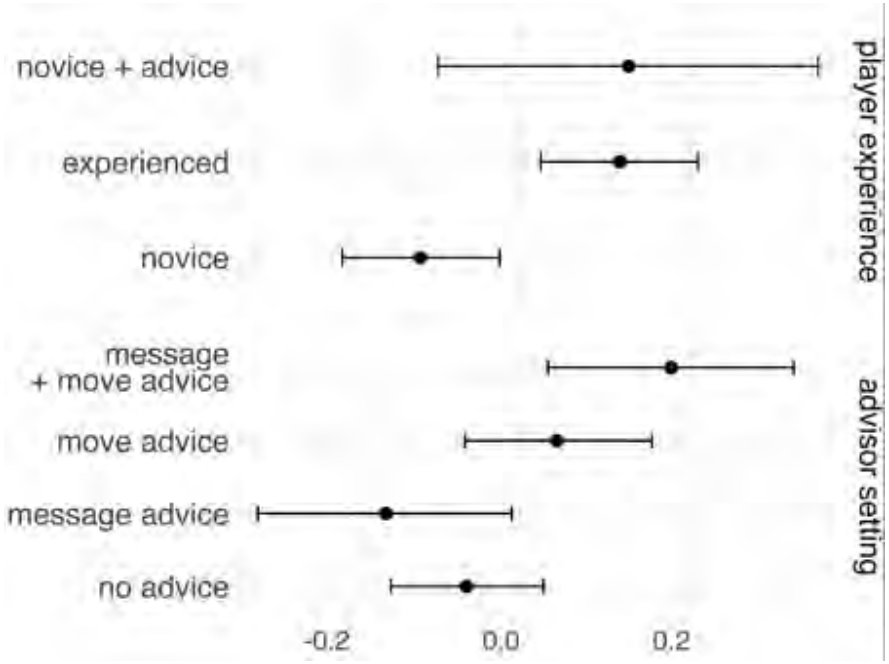




How does Advice Boost (or Diminish) Skill?



How does Advice Boost (or Diminish) Skill?



Why are messages hurting?

- Cognitive load / distraction
- Not tuned for suboptimal moves
- Inconsistency with moves

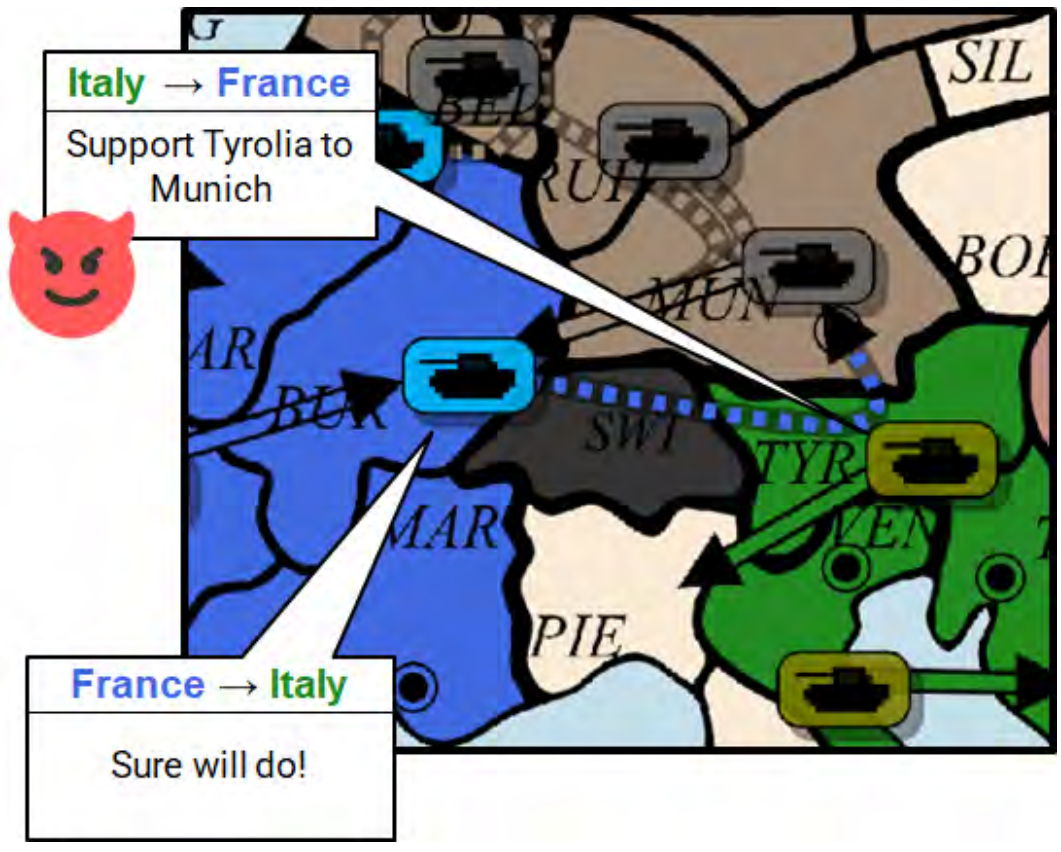
Are they just parrots?

	Move Advice		Message Advice	
	Accepted	Total	Accepted	Total
Novices	32.6%	872	6.3%	1413
Veterans	6.4%	2807	3.4%	2912

- Don't think for themselves?
- Beginners often take move advice
- Message advice rarely taken

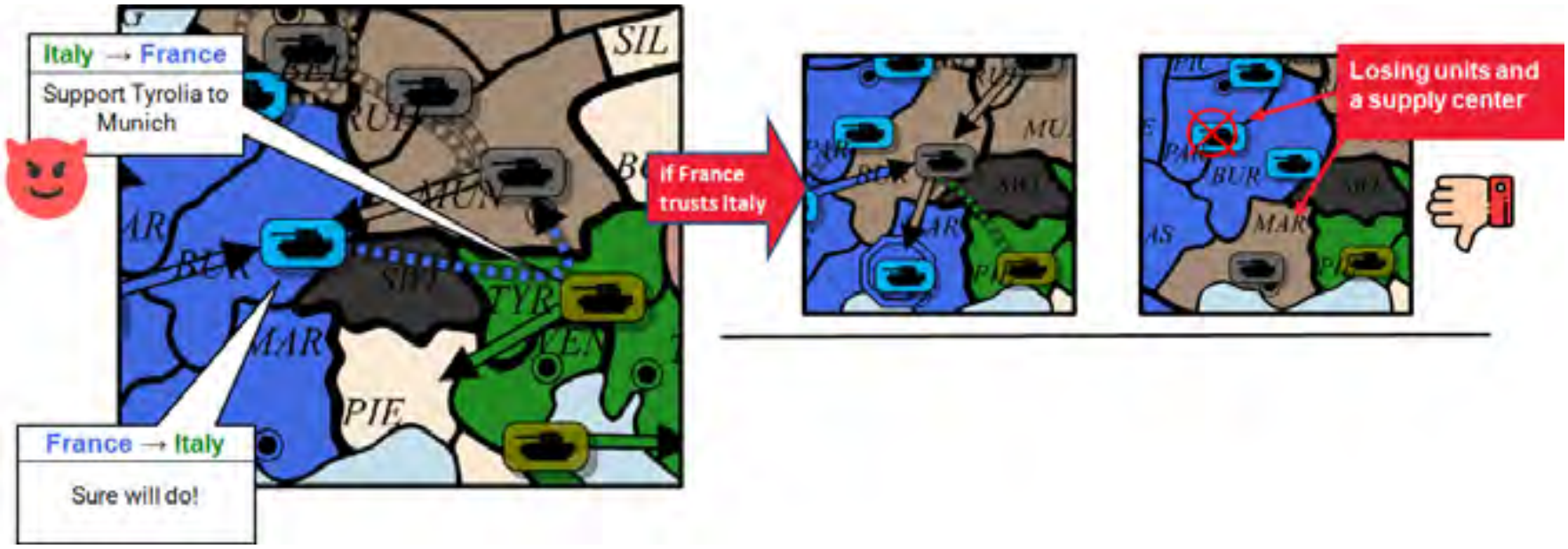
How to better detect Deception

How to better detect Deception

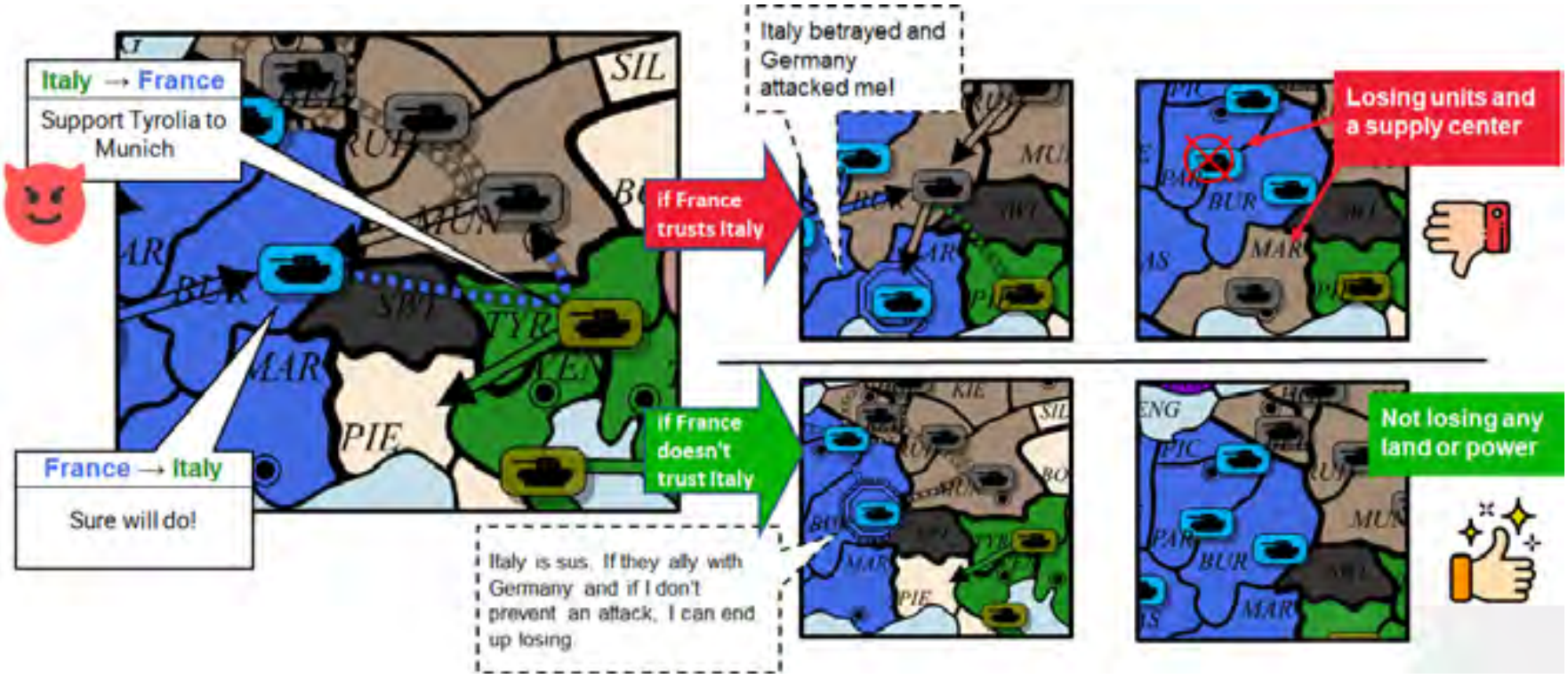


How to better detect Deception

How to better detect Deception



How to better detect Deception

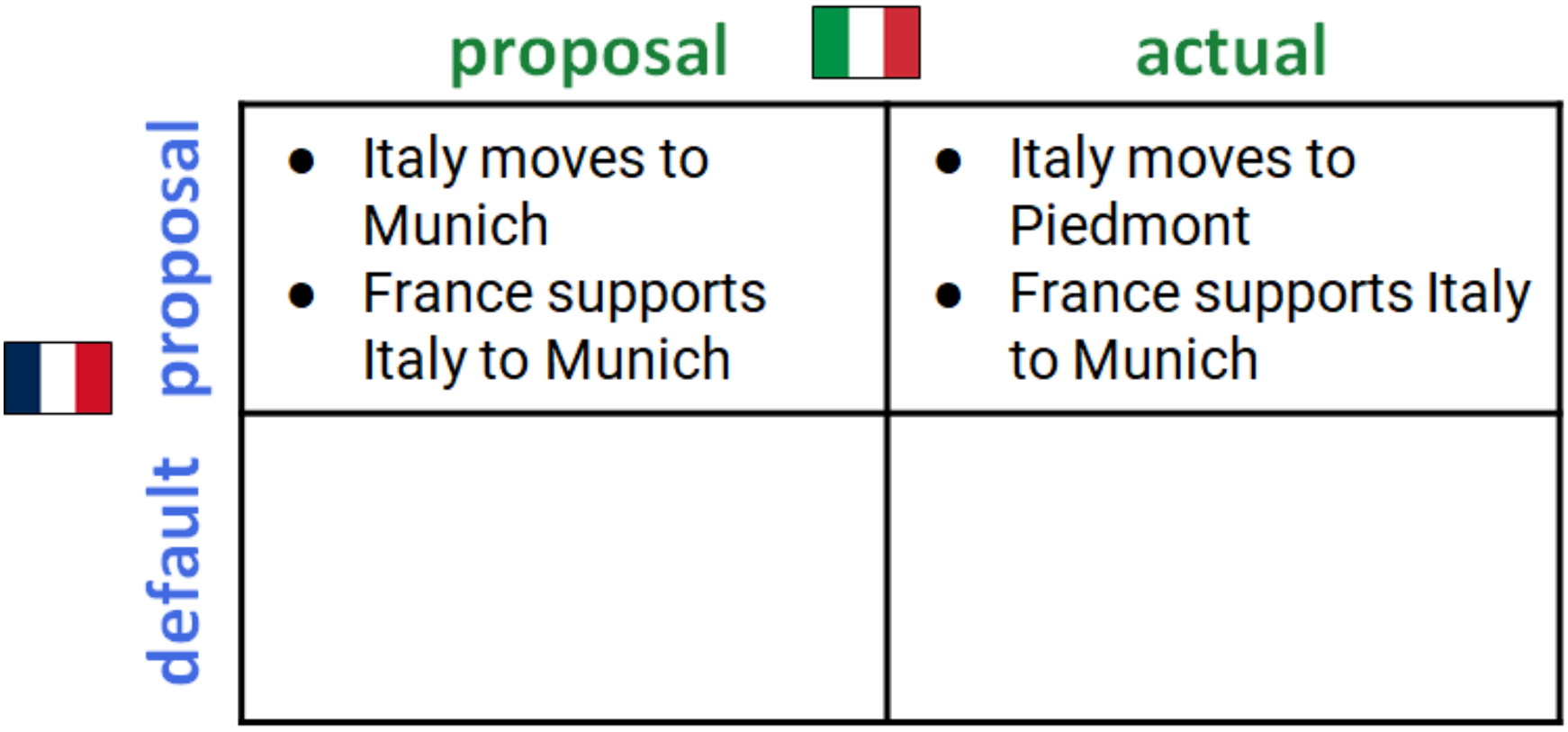


It's Time for Some Game Theory

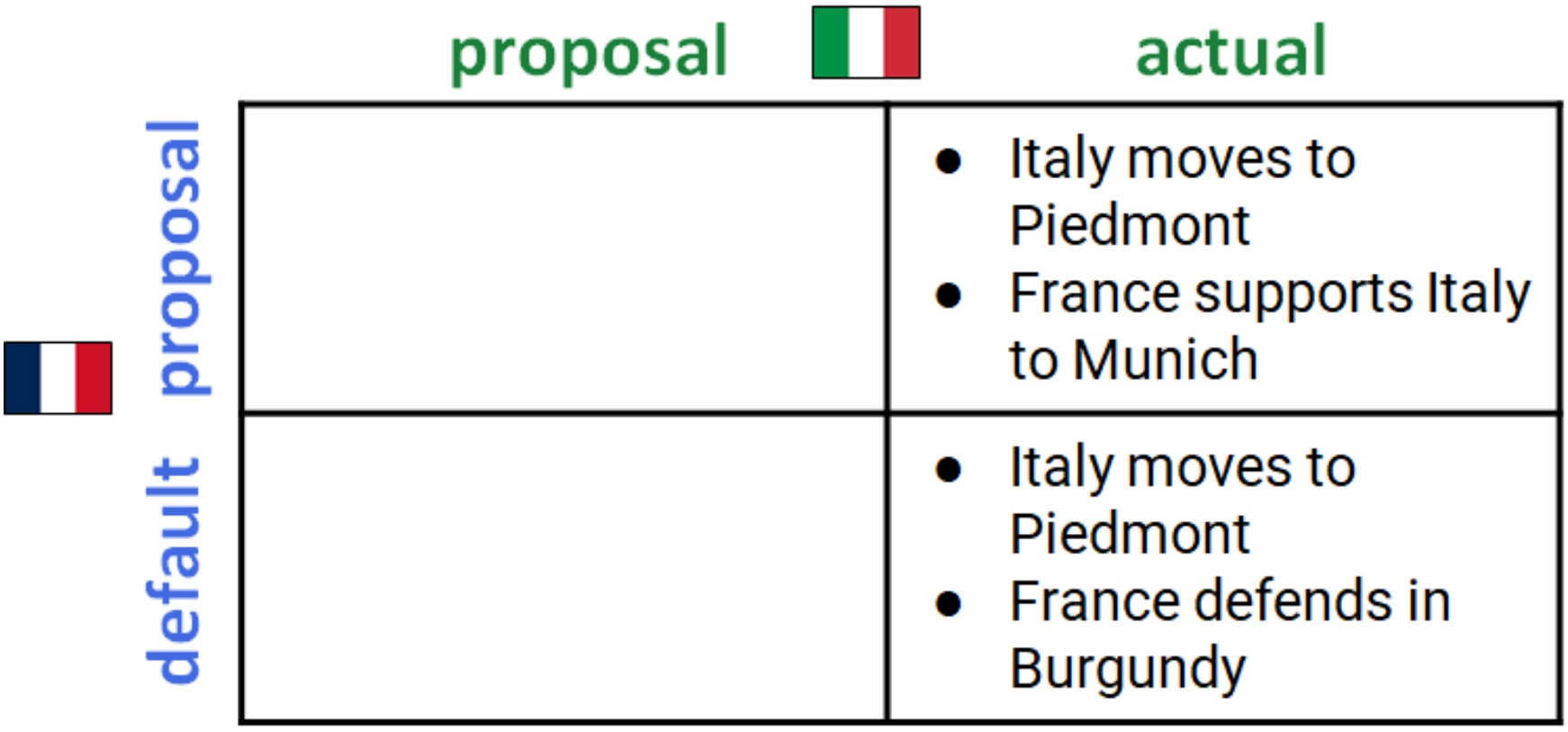
It's Time for Some Game Theory

	proposal 	actual
proposal	• Italy moves to Munich • France supports Italy to Munich	
default	• Italy moves to Munich • France defends in Burgundy	

It's Time for Some Game Theory



It's Time for Some Game Theory



It's Time for Some Game Theory

	proposal 	actual
	• Italy moves to Munich • France supports Italy to Munich	• Italy moves to Piedmont • France supports Italy to Munich
	• Italy moves to Munich • France defends in Burgundy	• Italy moves to Piedmont • France defends in Burgundy

Bait ➡ $U1 = u_{France}(prop, prop) - u_{France}(def, prop)$

Switch ➡ $U2 = u_{France}(prop, prop) - u_{France}(prop, act)$

Edge ➡ $U3 = u_{Italy}(prop, act) - u_{Italy}(def, act)$

Really Hard Problem, but Better

Model	Precision	Recall	F1-Score
LLM baseline using Direct Judgment	0.095	**0.551**	0.161
LLM baseline using Alignment Judgment	0.147	0.065	0.090
CTRL-D **(ours)**	**0.950**	0.238	**0.380**
CTRL-D with human annotated logical forms	**0.960**	0.300	**0.457**
Context LSTM + Power	0.263	0.171	0.207

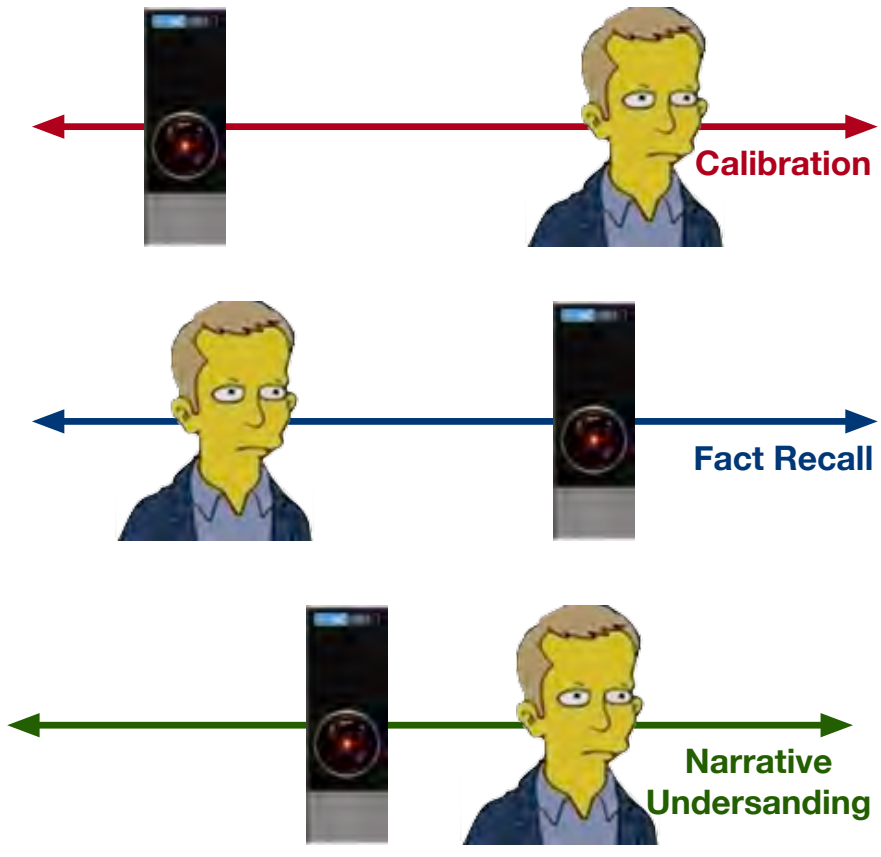
Human ability is around 0.2

Recap

- Reinforcement learning is really good at determining moves
- Language and cooperation lags behind
- Diplomacy remains a useful language game
  - Not just for winning the game
  - Are you respected
  - Are you truthful
  - Are you helpful
- Maybe we can do this on a small scale so that it applies to all AI

Human and Computer Skills

- LLMs are not perfect
  - Retrieval Methods: Slow but Trustworthy
  - LLMs: Use if you are confident
  - Humans: Leave them in driver's seat
- Item Response Theory
  - Highlight skills and examples
  - Measure how much computers *help* or *hurt*
- Application
  - Supporting Users to find False Claims
  - Supporting Users to Answer Questions
  - Supporting Users to Negotiate



Why didn't you...

- IR works better than LLM for fact checking *sometimes*
- We also showed bandit model can figure out what evidence to show when
- Use Bandit model to show *right* evidence to user

Why didn't you...

- IR works better than LLM for fact checking *sometimes*
- We also showed bandit model can figure out what evidence to show when
- Use Bandit model to show *right* evidence to user
- We're trying... requires good calibration
- ACL Submission: Adversarial Calibration
- Looking for collaborators

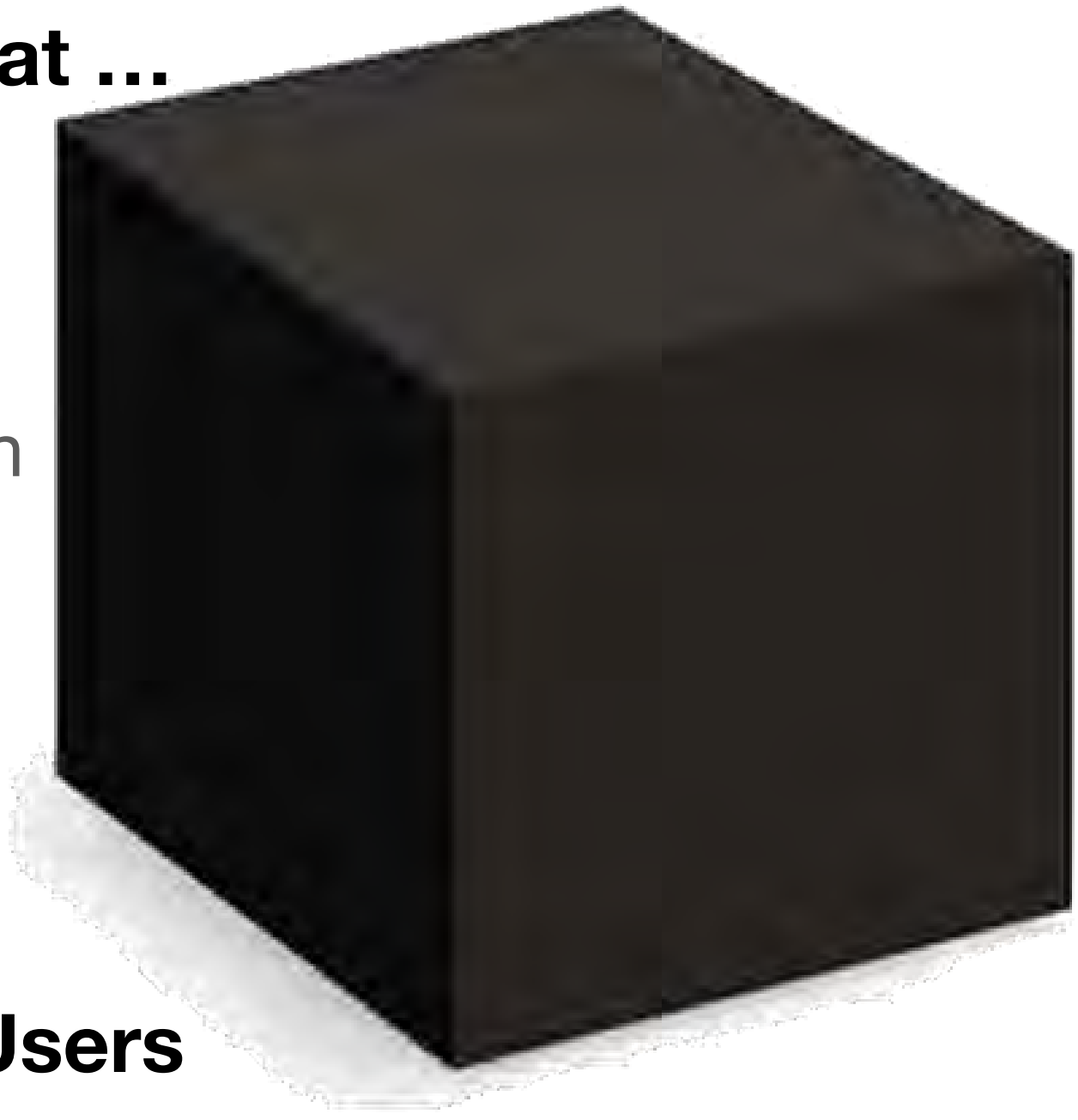
Algorithms that …

Inform

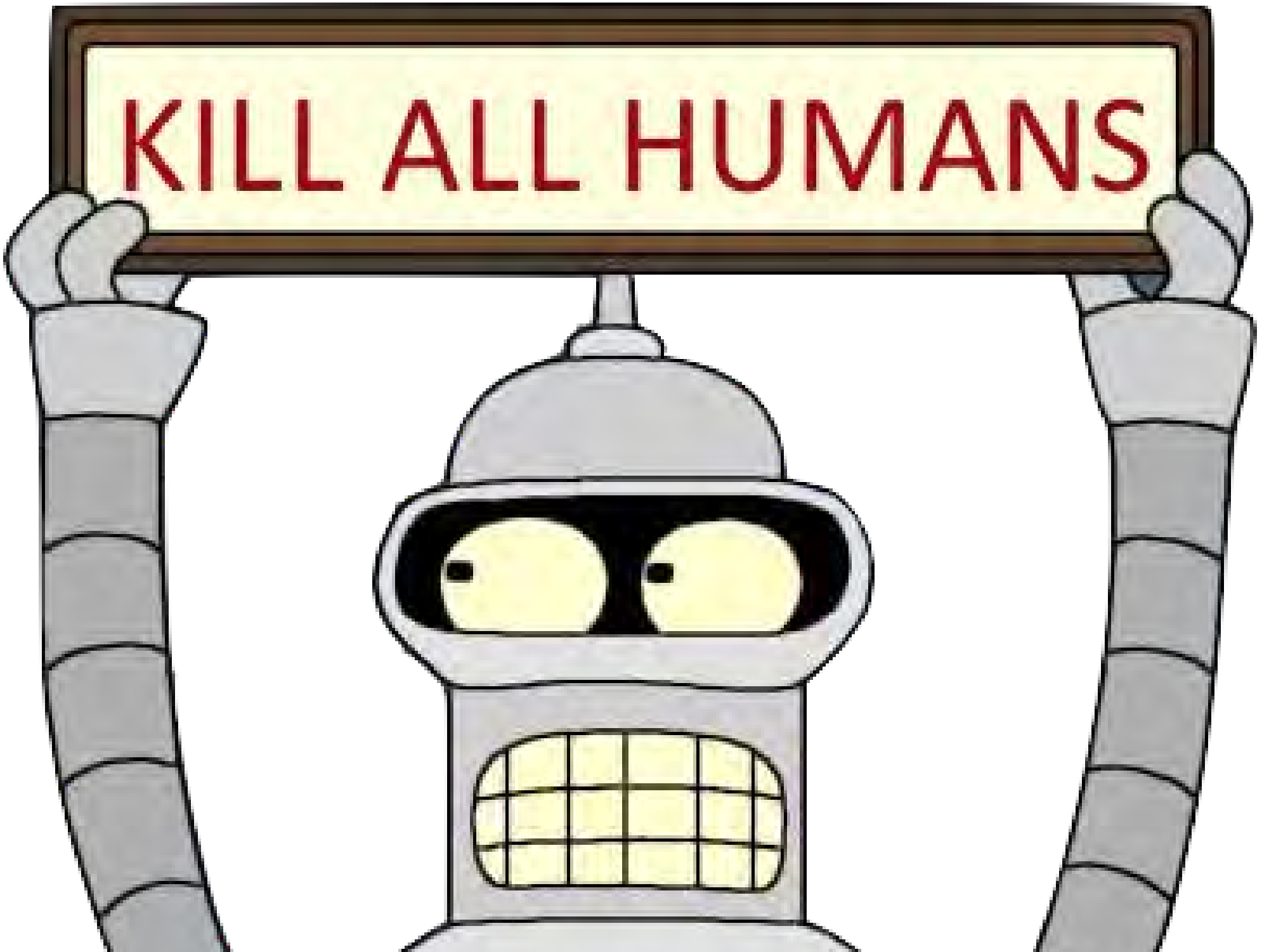
Collaborate with

Compete with

Understand



their Human Users


KILL ALL HUMANS

QANTA

QUESTION ANSWERING IS NOT A TRIVIAL ACTIVITY

June 14, 2025: College Park
June 21, 2025: Online
Details: http://qanta.org
Contact: qanta@googlegroups.com

- **Human+Computer Teams**
- **Tossups and Bonuses**
- **Submit computer agents / play as Human (eval: calibration and interpretability)**
- **Adversarial Questions**