



# Computers and Humans Playing Language Games Together

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

## Today's Takeaways

- Different Systems and Skilled Humans have Different Strengths
  - Retrieval Methods: Less likely to confabulate
  - LLMs: Good for easy, straightforward tasks
  - Humans: good for nuance and abstraction
- It takes work to figure this out
  - Getting harder examples
  - Measuring human vs. computer accuracy: Item Response Theory
  - Competitive and Cooperative
- Tasks
  - Finding false claims
  - Question answering
  - Detecting lies

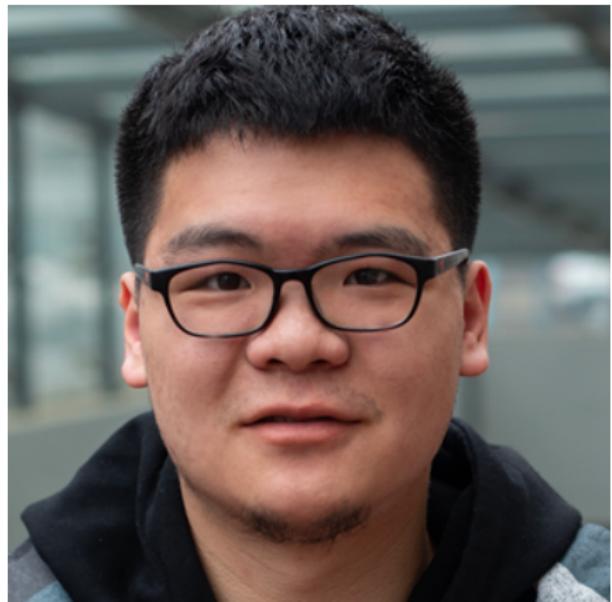
# Fact Checking with Humans in the Loop

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

Chenglei Si<sup>1</sup> Navita Goyal<sup>2</sup> Sherry Tongshuang Wu<sup>3</sup>

Chen Zhao<sup>4</sup> Shi Feng<sup>5</sup> Hal Daumé III<sup>2,6</sup> Jordan Boyd-Graber<sup>2</sup>

<sup>1</sup>Stanford University    <sup>2</sup>University of Maryland    <sup>3</sup>Carnegie Mellon University  
<sup>4</sup>NYU Shanghai    <sup>5</sup>New York University    <sup>6</sup>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?

Claim

Barbara Bush was a spouse of a United States president during his term.

Submit

Submit and flag

Skip (opens menu)

Home

Guidelines

## Wikipedia article for Barbara Bush

Barbara Bush (née Pierce; born June 8, 1925) is the wife of [George H. W. Bush](#), the [41st President of the United States](#), and served as First Lady of the United States from 1989 to 1993.

✓ Supports

✗ Refutes

Cancel

She is the mother of [George W. Bush](#), the 43rd President, and [Jeb Bush](#), the 43rd [Governor of Florida](#).

Expand

She served as the [Second Lady of the United States](#) from 1981 to 1989.

Expand

Barbara Pierce was born in Flushing, [New York](#).

Expand

She attended Milton Public School from 1931 to 1937, and Rye Country Day School from 1937-1940.

Expand

Add a custom page from Wikipedia if essential information is missing from the dictionary. E.g. the claim mentions an entity that does not appear in the Wikipedia page for Barbara Bush

Add Custom Page

If you need to combine multiple sentences from the original page (Barbara Bush), this will add it to the dictionary so that it can form part of the supporting evidence.

Add Main Wikipedia Page  
(Barbara Bush)

### Quick Links

[First Lady of the United States](#)

[George H. W. Bush](#)

[George W. Bush](#)

[List of Presidents of the United States](#)

### First Lady of the United States

First Lady of the United States (FLOTUS) is the informal but accepted title held by the wife of the President of the United States, concurrent with the president's term of office.

# Fact Extraction and VERification

## 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

Google Research

## Fool Me Twice Entailment from Wikipedia Gamification

Julia Eisenschlos, Michael Lüdtke, Jörn Börschinger und Julian Bösel  
griseis@informatik.uni-frankfurt.de

Jannis Bösel   Jordan Boyd-Graber   Julian Eisenschlos



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

The screenshot shows a web browser window with the title bar "Chrome" and various tabs open. The main content area displays a game titled "Fool Me Twice". The top navigation bar includes "Play" and "Leaderboard" buttons. On the right, it shows "LEVEL 3 - DEBUT AUTHOR · 10068 POINTS" and a notification bell icon.

**Game Interface:**

- Left Panel:** A text input field containing "Write a false statement about Michael Faraday". Below it is a "SAVE FALSE STATEMENT" button.
- Middle Panel:** A section titled "Evidence (0 marked as gold)" with a note: "Michael Faraday (; 22 September 1791 – 25 August 1867) was an English scientist who contributed to the study of electromagnetism and electrochemistry." Below this is a "Gold evidence" button.
- Right Panel:** A "Table of Contents" sidebar with links like "Summary", "Personal life | Early life", "Adult life", "Later life", "Scientific achievements | Chemistry", "Electricity and magnetism", "Diamagnetism", "Faraday cage", "Royal Institution and public service", "Commemorations", "Awards named in Faraday's honor", and "Bibliography".
- Bottom Panel:** Two examples of student answers:
  - "His main discoveries include the principles underlying electromagnetic induction, diamagnetism and electrolysis." (marked as gold)
  - "Although Faraday received little formal education, he was one of the most influential scientists in history." (marked as gold)

**Summary:**

Michael Faraday (; 22 September 1791 – 25 August 1867) was an English scientist who contributed to the study of electromagnetism and electrochemistry.

His main discoveries include the principles underlying electromagnetic induction, diamagnetism and electrolysis. Although Faraday received little formal education, he was one of the most influential scientists in history. It was by his research on the magnetic field around a conductor carrying a direct current that Faraday established the basis for the concept of the electromagnetic field in physics. Faraday also established that magnetism could affect rays of light and that there was an underlying relationship between the two phenomena. He similarly discovered the principles of electromagnetic induction and diamagnetism, and the laws of electrolysis.

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

The screenshot shows a Google Chrome browser window with multiple tabs open. The active tab is titled "fool-me-twice.googleplex.com/play". The page content is from the game "Fool Me Twice".

**Fool Me Twice** Play - Leaderboard

LEVEL 3 - DEBUT AUTHOR · 10068 POINTS

**< Michael Faraday 5:00**

Faraday died in an explosion preparing nitrogen trichloride samples

from Outhgill in Westmorland, where he had been an apprentice to the village blacksmith. Michael was born in the autumn of that year. The young Michael Faraday, who was the third of four children, having only the most basic school education, had to educate himself. At the age of 14 he became an apprentice to George Riebau, a local bookbinder and bookseller in Blandford Street. During his seven-year apprenticeship Faraday read many books, including Isaac Watts's *The Improvement of the Mind*, and he enthusiastically implemented the principles and suggestions contained therein. He also developed an interest in science, especially in electricity.

Faraday was particularly inspired by the book *Conversations on Chemistry* by Jane Marcet.

**| Adult life**

In 1812, at the age of 20 and at the end of his apprenticeship, Faraday attended lectures by the eminent English chemist Humphry Davy of the Royal Institution and the Royal Society, and John Tatum, founder of the City Philosophical Society. Many of the tickets for these lectures were given to Faraday by William Dance, who was one of the founders of the Royal Philharmonic Society. Faraday subsequently sent Davy a 300-page book based on notes that he had taken during these lectures. Davy's reply was immediate, kind, and favourable. In 1813, when Davy damaged his eyesight in an accident with nitrogen trichloride, he decided to employ Faraday as an assistant. Coincidentally one of the Royal Institution's assistants, John Payne, was sacked and Sir Humphry Davy had been asked to find a replacement; thus he appointed Faraday as Chemical Assistant at the Royal Institution on 1 March 1813.

Very soon Davy entrusted Faraday with the preparation of nitrogen trichloride samples, and they both were injured in an explosion of this very sensitive substance.

Gold evidence ⏺ GO

In 1813, when Davy damaged his eyesight in an accident with nitrogen trichloride, he decided to employ Faraday as an assistant.

Gold evidence ⏺ GO

Faraday died at his house at Hampton Court on 25 August 1867, aged 75.

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

The screenshot shows a Google Chrome browser window with multiple tabs open. The active tab is titled "fool-me-twice.googleplex.com/play". The page content is from the game "Fool Me Twice".

**Header:** Chrome, File, Edit, View, History, Bookmarks, Profiles, Tab, Window, Help. Date: Tue 4 May 16:59.

**Title Bar:** TabME, TSOA, Sheets, DataQnA, contr, Edito, run\_1, TLM, Resil, QED!, MATI, Jona(s), Fool!, Fool!, Tap!, TaPa!, +. Sharing this tab to meet.google.com/ Stop.

**Main Content:**

**Section 1: Michael Faraday**

Faraday was particularly inspired by the book Conversations on Chemistry by Jane Marcet.

Faraday succumbed to his injuries in an explosion preparing nitrogen trichloride samples.

Save False Statement

**Evidence (2 marked as gold)**

- Gold evidence ⓘ GD  
This is now termed the Faraday effect.
- Gold evidence ⓘ GD  
Faraday married Sarah Barnard (1800–1879) on 12 June 1821.
- Gold evidence ⓘ GD  
Don't see the gold evidence you're looking for? You can add it by clicking or the sentence in the Wikipedia page (right).

**Section 2: Adult life**

In 1812, at the age of 20 and at the end of his apprenticeship, Faraday attended lectures by the eminent English chemist Humphry Davy of the Royal Institution and the Royal Society, and John Tatum, founder of the City Philosophical Society. Many of the tickets for these lectures were given to Faraday by William Dance, who was one of the founders of the Royal Philharmonic Society. Faraday subsequently sent Davy a 300-page book based on notes that he had taken during these lectures. Davy's reply was immediate, kind, and favourable. In 1813, when Davy damaged his eyesight in an accident with nitrogen trichloride, he decided to employ Faraday as an assistant. Coincidentally one of the Royal Institution's assistants, John Payne, was sacked and Sir Humphry Davy had been asked to find a replacement; thus he appointed Faraday as Chemical Assistant at the Royal Institution on 1 March 1813.

Very soon Davy entrusted Faraday with the preparation of nitrogen trichloride samples, and they both were injured in an explosion of this very sensitive substance. Faraday married Sarah Barnard (1800–1879) on 12 June 1821. They met through their families at the Sandemanian church, and he confessed his faith to the Sandemanian congregation the month after they were married. They had no children. Faraday was a devout Christian; his Sandemanian denomination was an offshoot of the Church of Scotland. Well after his marriage, he served as deacon and for two terms as an elder in the meeting house of his youth. His church was located at Paul's Alley in the Barbican. This meeting house relocated in 1862 to Barnsbury Grove, Islington; this North London location was where Faraday served the final two years of his second term as elder prior to his resignation from that post.

**Section 3: Later life**

Biographers have noted that "a strong sense of the unity of God and nature pervaded Faraday's life and work." In June 1832, the University of Oxford granted Faraday an honorary Doctor of Civil Law degree. During his lifetime, he was offered a knighthood in recognition for his services to science, which he turned down on religious grounds, believing that it was against the word of the Bible to accumulate riches and pursue worldly reward, and stating that he

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

The screenshot shows a Google Chrome browser window with multiple tabs open. The active tab is titled "fool-me-twice.googleplex.com/play". The page content is from the game "Fool Me Twice" about Michael Faraday. The top navigation bar includes "Play" and "Leaderboard". On the right, it says "LEVEL 3 - DEBUT AUTHOR · 10068 POINTS".

**Michael Faraday** (0:00)

Faraday succumbed to his injuries in an **explosion** preparing nitrogen **trichloride** samples.

**Evidence (1 marked as gold)**

Very soon Davy entrusted **Faraday** with the preparation of **nitrogen trichloride** samples, and they both were injured in an **explosion** of this very sensitive substance.

Gold evidence

In 1813, when Davy damaged his eyesight in an accident with nitrogen trichloride, he decided to employ **Faraday** as an assistant.

Gold evidence

This work included investigations of **explosions** in coal mines, being an expert witness in court, and along with two engineers from Chance Brothers c. 1853, the preparation of high-quality optical glass, which was

**Links:**  
Electricity and magnetism  
Diamagnetism  
Faraday cage  
Royal Institution and public service  
Commemorations  
Awards named in Faraday's honor  
Bibliography

**Summary**

Michael Faraday (22 September 1791 – 25 August 1867) was an English scientist who contributed to the study of electromagnetism and electrochemistry. His main discoveries include the principles underlying electromagnetic induction, diamagnetism and electrolysis. Although Faraday received little formal education, he was one of the most influential scientists in history. It was by his research on the magnetic field around a conductor carrying a direct current that Faraday established the basis for the concept of the electromagnetic field in physics. Faraday also established that magnetism could affect rays of light and that there was an underlying relationship between the two phenomena. He similarly discovered the principles of electromagnetic induction and diamagnetism, and the laws of electrolysis.

His inventions of electromagnetic rotary devices formed the foundation of electric motor technology, and it was largely due to his efforts that electricity became practical for use in technology. As a chemist, Faraday discovered benzene, investigated the clathrate hydrate of chlorine, invented an early form of the Bunsen burner and the system of oxidation numbers, and popularised terminology such as "anode", "cathode", "electrode" and "ion".

Faraday ultimately became the first and foremost Fullerian Professor of Chemistry at the Royal Institution, a lifetime position. Faraday was an excellent experimentalist who conveyed his ideas in clear and simple language; his mathematical abilities, however, did not extend as far as trigonometry and were limited to the simplest algebra. James Clerk Maxwell took the work of Faraday and others and summarized it in a set of equations which is accepted as the basis of all modern theories of electromagnetic phenomena. On Faraday's death of force, Maxwell

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

Sharing this tab to [meet.googleplex.com/play](https://meet.googleplex.com/play) Stop

Fool Me Twice Play - Leaderboard LEVEL 3 - DEBUT AUTHOR · 10068 POINTS

< Michael Faraday 0:00

Faraday succumbed to his injuries in an explosion preparing nitrogen trichloride samples in 1813

SAVE FALSE STATEMENT

Evidence (1 marked as gold)

Very soon Davy entrusted Faraday with the preparation of nitrogen trichloride samples, and they both were injured in an explosion of this very sensitive substance.

Gold evidence

In 1813, when Davy damaged his eyesight in an accident with nitrogen trichloride, he decided to employ Faraday as an assistant.

Gold evidence

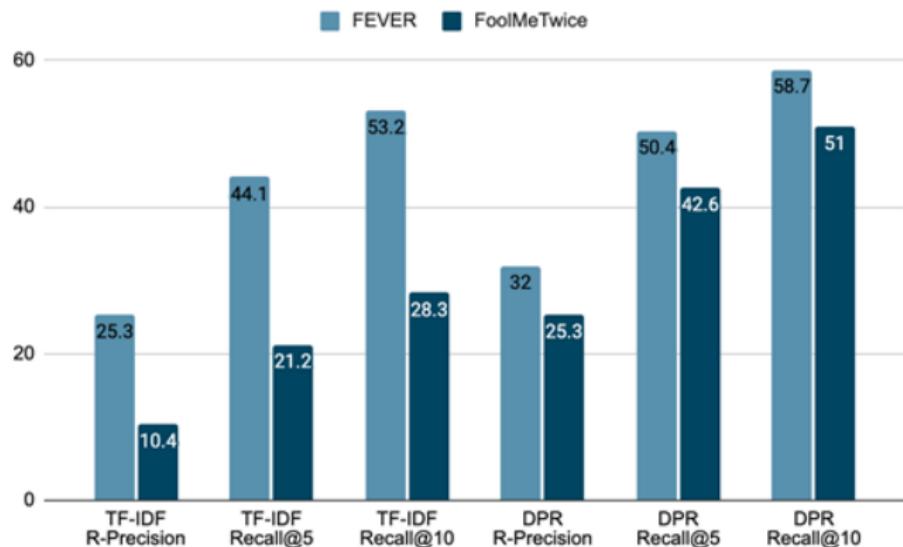
This work included investigations of explosions in coal mines, being an expert witness in court, and along with two engineers from Chance Brothers c. 1853, the preparation of high-quality optical glass, which was required by Chance for its lighthouses. In 1846, together with Charles Lyell, he produced a lengthy and detailed report on a serious explosion in the colliery at Haswell, County Durham, which killed 95 miners. Their report was a meticulous forensic investigation and indicated that coal dust contributed to the severity of the explosion.

The report should have warned coal owners of the hazard of coal dust explosions, but the risk was ignored for over 60 years until the Senghenydd Colliery Disaster of 1913. As a respected scientist in a nation with strong maritime interests, Faraday spent extensive amounts of time on projects such as the construction and operation of lighthouses and protecting the bottoms of ships from corrosion. His workshop still stands at Trinity Buoy Wharf above the Chain and Buoy Store, next to London's only lighthouse where he carried out the first experiments in electric lighting for lighthouses. Faraday was also active in what would now be called environmental science, or engineering. He investigated industrial pollution at Swansea and was consulted on air pollution at the Royal Mint.

In July 1855, Faraday wrote a letter to The Times on the subject of the foul condition of the River Thames, which resulted in an often-reprinted cartoon in Punch. (See also The Great Stink). Faraday assisted with the planning and judging of exhibits for the Great Exhibition of 1851 in London. He also advised the National Gallery on the cleaning and protection of its art collection, and served on the National Gallery Site Commission in 1857. Education was another of Faraday's areas of service; he lectured on the topic in 1854 at the Royal Institution, and in 1862 he appeared before a Public Schools Commission to give his views on education in Great Britain.

Faraday also weighed in negatively on the public's fascination with table-turning, mesmerism, and seances, and in so doing chastised both the public and the nation's educational system. Before his famous Christmas lectures, Faraday delivered chemistry lectures for the City Philosophical Society from 1816 to 1818 in order to refine the quality of his lectures. Between 1827 and 1860 at the Royal Institution in London, Faraday gave a series of nineteen Christmas lectures for young people, a series which continues today. The objective of Faraday's Christmas lectures was to present science to the general public in the hopes of inspiring them and generating revenue for the Royal Institution. They were notable events on the social calendar among London's gentry. Over the course of several letters to his close friend Benjamin Aihart, Faraday outlined his recommendations on the art of lecturing: Faraday wrote "a flame

## It is Harder (a bit)



## **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.

### **Case For**

Manhattan is the most populous borough in New York City, with a population of approximately 1,627,788 as of 2025. The other boroughs have smaller populations: Brooklyn: 2.72 million; Queens: 2.57 million; The Bronx: 1.46 million; Staten Island: 552,00.

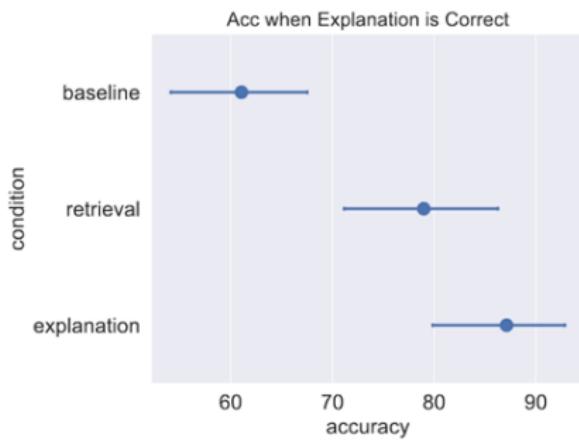
### **IR**

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.

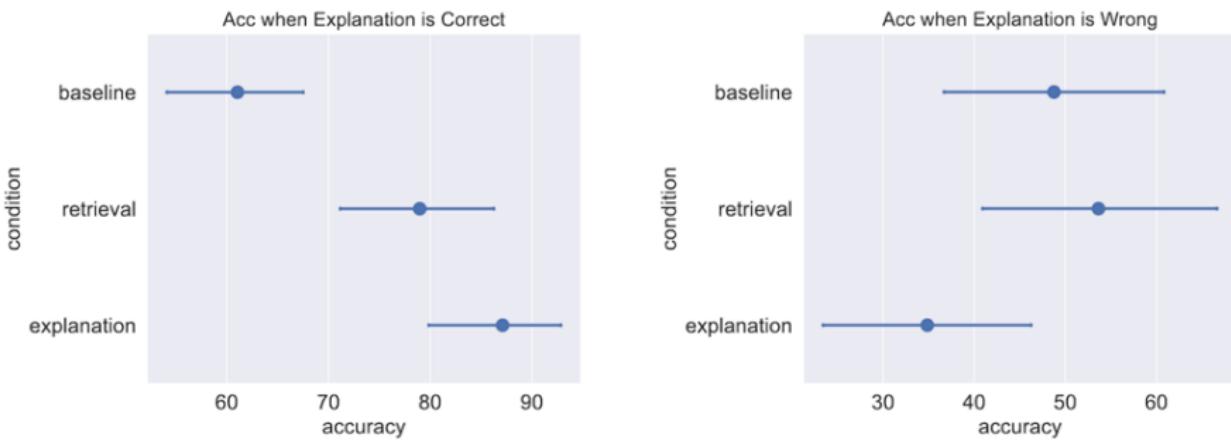
### **Case Against**

According to recent data, Brooklyn is the most populated borough in New York City, not Manhattan. Manhattan has a population of approximately 1.6 million residents, while Brooklyn has a significantly higher population.

## Results

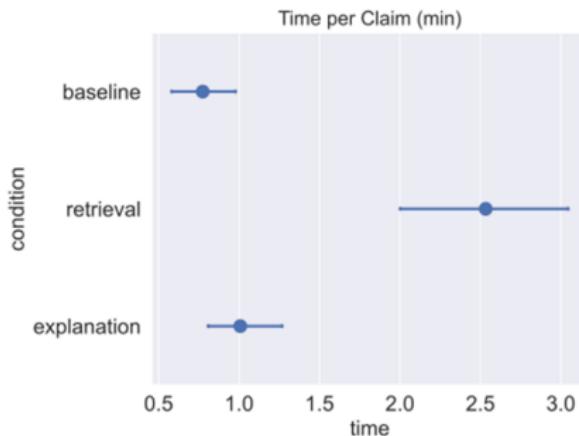


## Results



**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<sup>1</sup>

Hal Daumé III<sup>1,2</sup>

Tianyi Zhou<sup>1</sup>

Jordan Boyd-Graber<sup>1</sup>

<sup>1</sup>University of Maryland      <sup>2</sup>Microsoft Research  
mgor@cs.umd.edu



# Item Response Theory



## 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.

(A)						
(B)						
(C)						
(D)						
(E)						
(F)						
(G)						
(H)						
(I)						
(J)						
(K)						
(L)						
(M)						
(N)						
(O)						
(P)						
(Q)						
(R)						
(S)						
(T)						
(U)						

2. Your Name: \_\_\_\_\_  
(First) \_\_\_\_\_ Last \_\_\_\_\_  
I agree to the conditions on the back of the SAT I test book.

Signature: \_\_\_\_\_ Date: \_\_\_\_\_ / \_\_\_\_\_ / \_\_\_\_\_

Home Address: \_\_\_\_\_  
(Print) \_\_\_\_\_ Number and Street \_\_\_\_\_

Center: \_\_\_\_\_  
(Print) \_\_\_\_\_ City \_\_\_\_\_ State \_\_\_\_\_ Zip Code \_\_\_\_\_

### 3. Date of Birth

Month	Day	Year
Jan.	1	0
Feb.	2	0
Mar.	3	0
Apr.	4	0
May	5	0
June	6	0
July	7	0
Aug.	8	0
Sept.	9	0
Oct.	0	1
Nov.	1	1
Dec.	2	1

### 4. Social Security Number

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

### 6. Registration Number

(Copy from Admission Ticket.)

### 7. Test Book Serial Number

(Copy from front of test book.)

(A)	(A)	(D)	(D)	(D)
(B)	(B)	(I)	(I)	(I)
(C)	(C)	(Z)	(Z)	(Z)
(D)	(D)	(3)	(3)	(3)
(E)	(E)	(4)	(4)	(4)
(F)	(F)	(5)	(5)	(5)
(G)	(G)	(6)	(6)	(6)
(H)	(H)	(7)	(7)	(7)
(I)	(I)	(8)	(8)	(8)
(J)	(J)	(9)	(9)	(9)
(K)	(K)			
(L)	(L)			
(M)	(M)			
(N)	(N)			
(O)	(O)			
(P)	(P)			
(Q)	(Q)			
(R)	(R)			

**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.)

## Item Response Theory

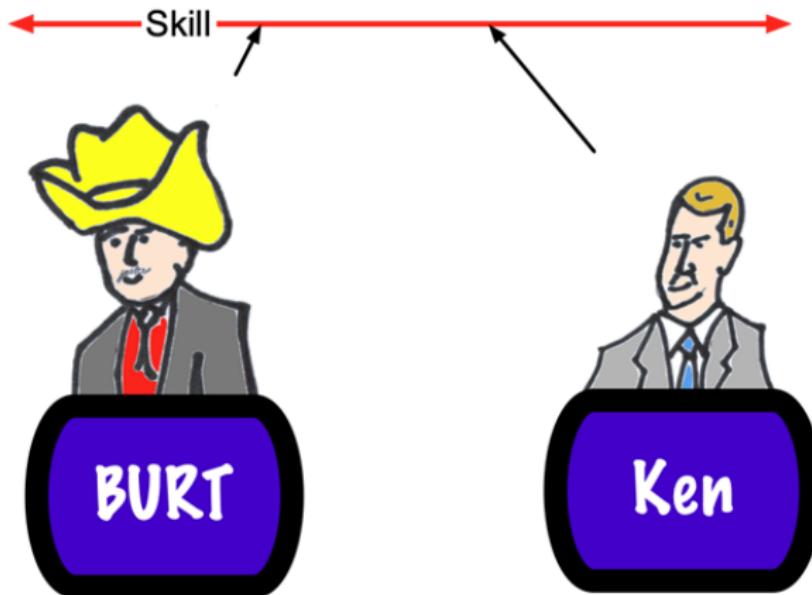
<b>SAT</b>		SAT I: Reasoning Test		Page 1	Use a No. 2 pencil only. Be sure each mark is dark and completely fills the intended oval. Completely erase any errors or stray marks.	
1. Your Name First 4 letters of Last Name      First init. Mid. init.						2. Your Name: _____ (Print) _____ Last _____ First _____ 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 _____						3. Date of Birth Month Day
4. Social Security Number 0 0 0 0 0 0 0 0 0						
						
5. Sex _____ Male _____ Female _____						
6. Registration Number (Copy from Admission Ticket)						
7. Test Book Serial Number (Copy from front of test book)						
						<b>IMPORTANT:</b> Fill in items 8 and 9 exactly as shown on the back of test book.
						<b>8. Form Code</b> (Copy and grid as on back of test book.)
						

## Item Response Theory

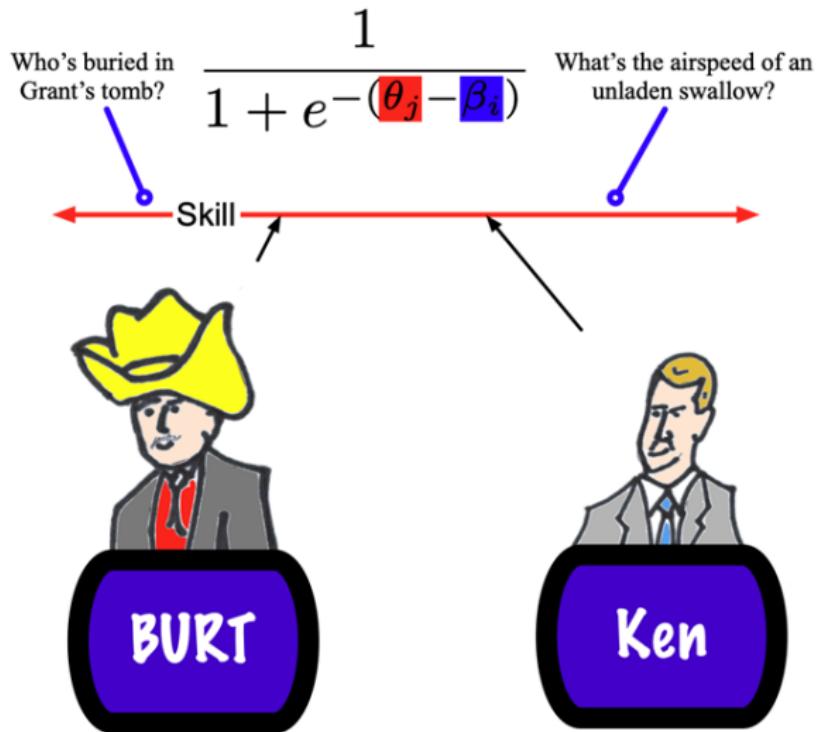


## Item Response Theory

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



## Item Response Theory



## Making Dimensions Interpretable

- Make skills and difficulty vector-valued

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

- Where do the dimensions come from?

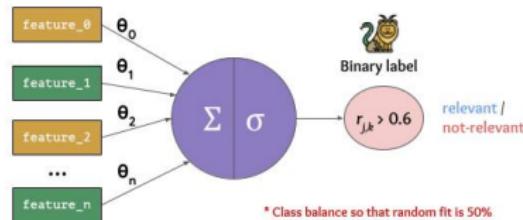
## Making Dimensions Interpretable

- Make skills and difficulty vector-valued

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

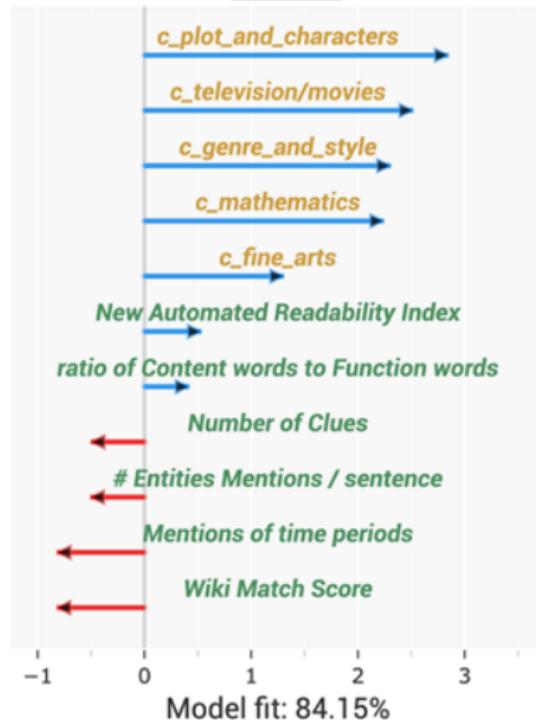
- Where do the dimensions come from?
  - Latent variable: learned to predict correctness
  - Function of question encoding, question features
  - Regularized to be sparse
- Posthoc labeling

For each dim  $k$ , run Logistic Regression on  
“is this question relevant to dim  $k$ ”



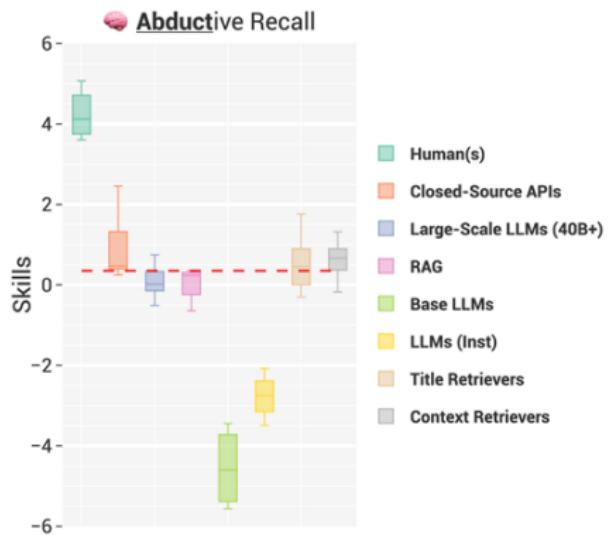
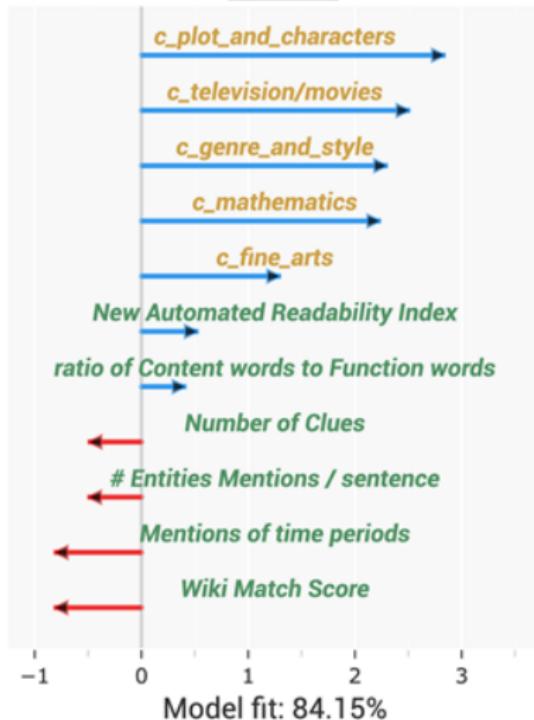
## Hard for Computers: Abductive Inference

### Dim 1: 🧠 Abductive Recall



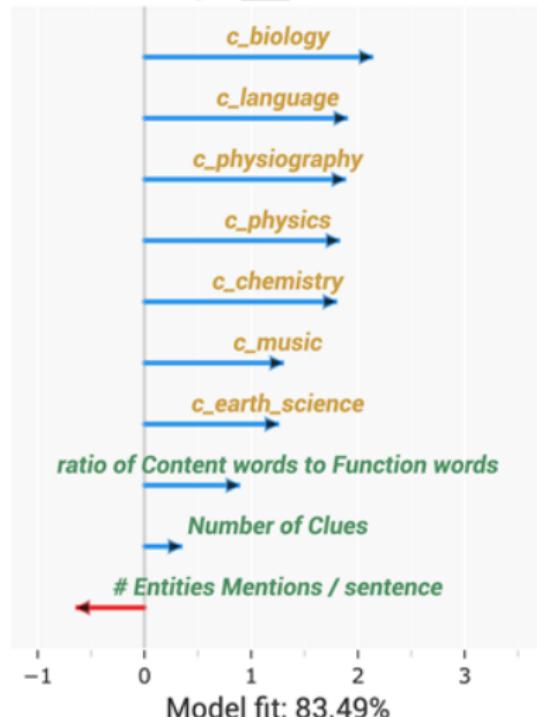
## Hard for Computers: Abductive Inference

### Dim 1: 🧠 Abductive Recall



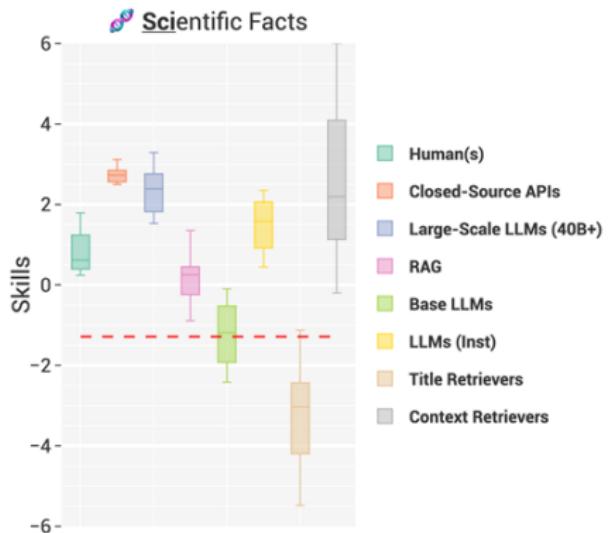
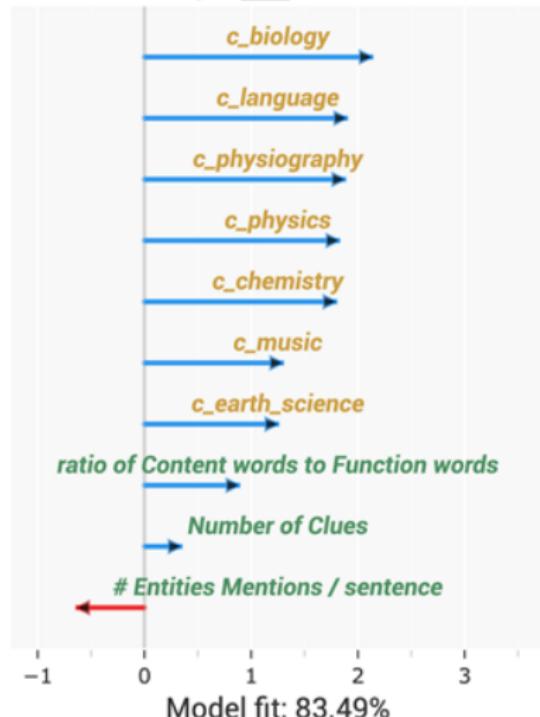
## Hard for Humans: Science

### Dim 3: Scientific Facts

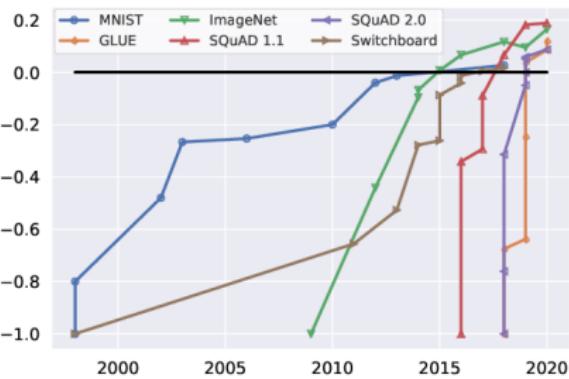


## Hard for Humans: Science

### Dim 3: Scientific Facts

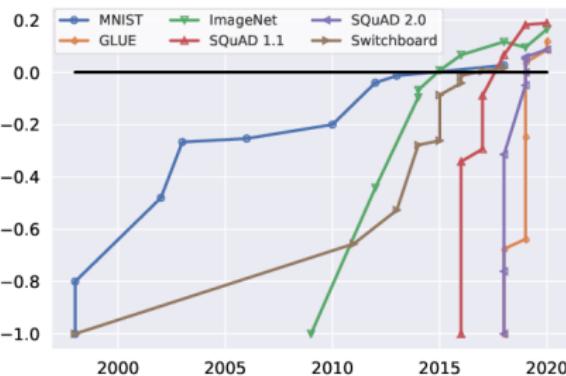


## Adversarial Datasets



- 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

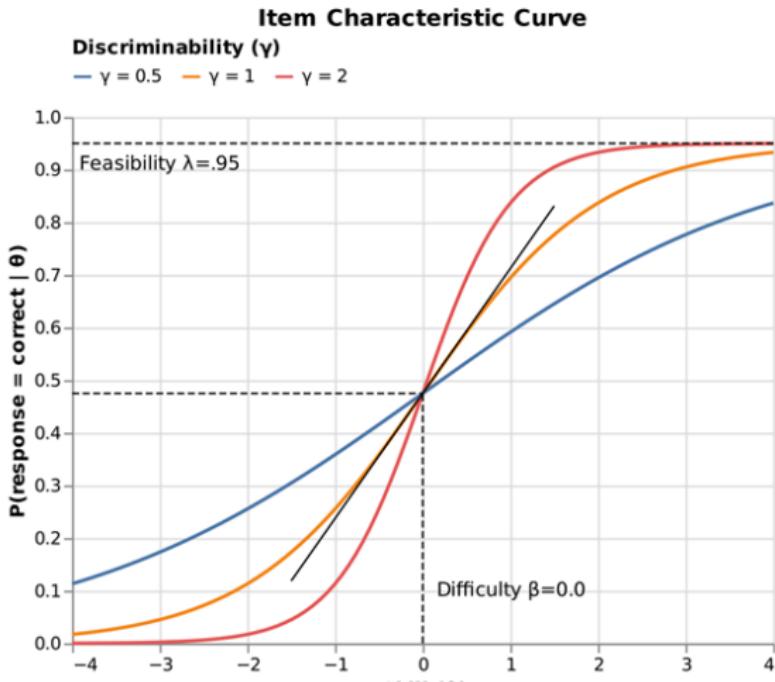


- Many benchmarks are “saturated”
- Newer datasets claim to be “adversarial”
  - Hard for computers, “easy” for humans
  - No real metric / definition
- Can we use the lessons of the previous paper to inform how to write hard examples
- Can we *measure* how well we did?
- Language game: increasing the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

## Expanding IRT: Discriminability

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

- Difficulty:  $\beta$
  - Discriminability:  $\gamma$
  - Feasibility:  $\lambda$
- 
- Skill/ability:  $\theta$



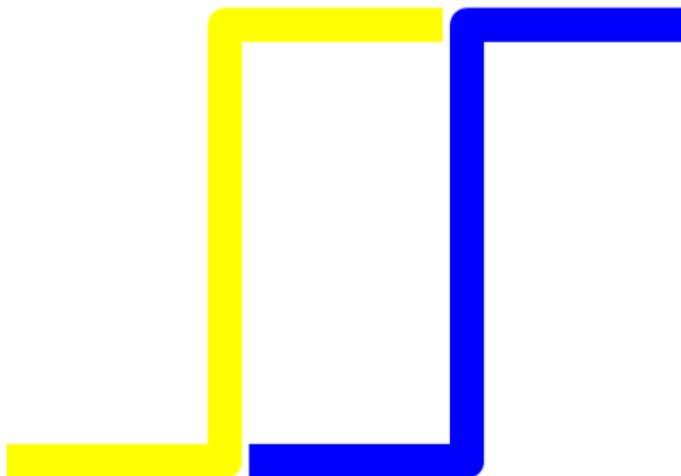
**Finding Skill of Subject is Like Binary Search**

$$\beta_1 = .2$$



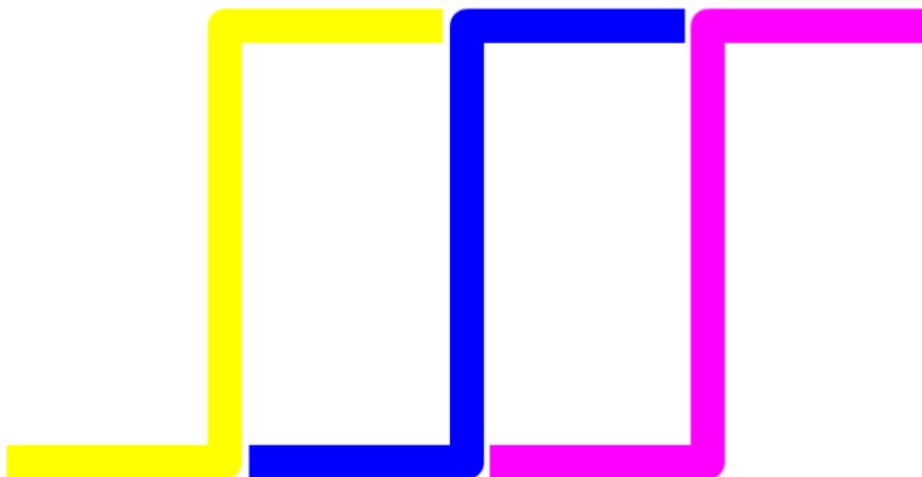
Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2 \quad \beta_2 = .4$$



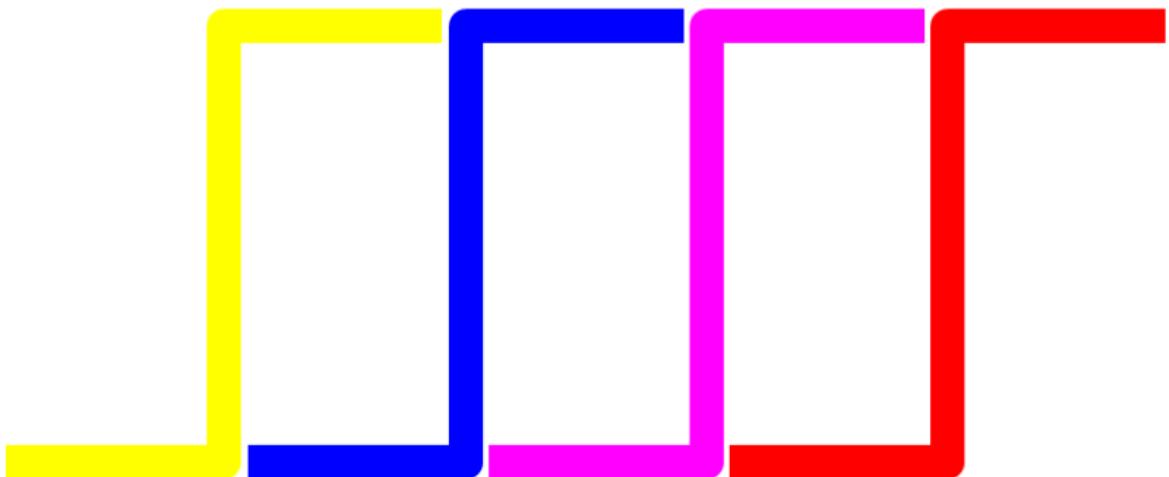
Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2 \quad \beta_2 = .4 \quad \beta_3 = .6$$



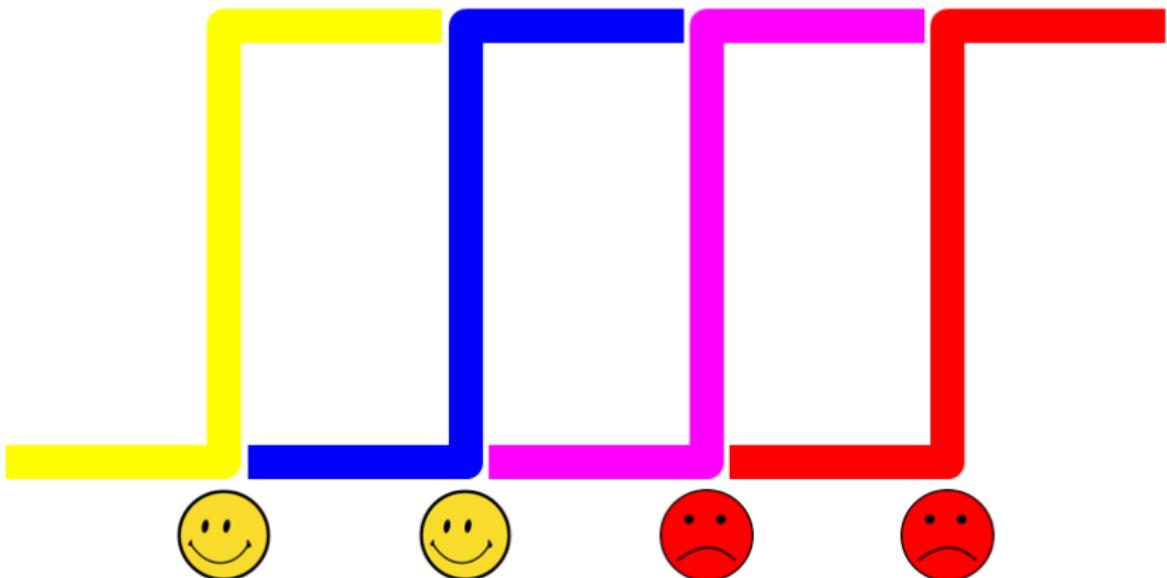
## Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2 \quad \beta_2 = .4 \quad \beta_3 = .6 \quad \beta_4 = .8$$



## Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2 \quad \beta_2 = .4 \quad \beta_3 = .6 \quad \beta_4 = .8$$



## Adversarial Score

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

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

## Adversarial Score

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

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

### 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!)

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

- Skilled humans should agree on the answer

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

## Adversarial Score

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

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

- Skilled humans should agree on the answer

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

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

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

$$p_j(\theta) = \frac{1}{1 + \exp(-\text{TIE}_j)} = 1 - \exp(-\text{TIE}_j) \quad (5)$$

## Adversarial Score

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

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

- Skilled humans should agree on the answer

$$\delta_j = \sum_{i \sim H_{(1)}} \left[ \frac{\frac{1}{\exp \left\{ -\gamma_j \left[ \beta_i^{H(1)} - \theta_j \right] \right\}} - \rho_{H_{(1)}}(r_{i,j})}{|H_{(1)}|} \right]. \quad (3)$$

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

$$\kappa_j = 1 - \exp(-\text{TIF}_j). \quad (4)$$

## Adversarial Score

$$\text{ADVSCORE}_j = \frac{\mu_j}{1 + \delta_j} \cdot (1 + \kappa_j) \quad (2)$$

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

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

- Skilled humans should agree on the answer

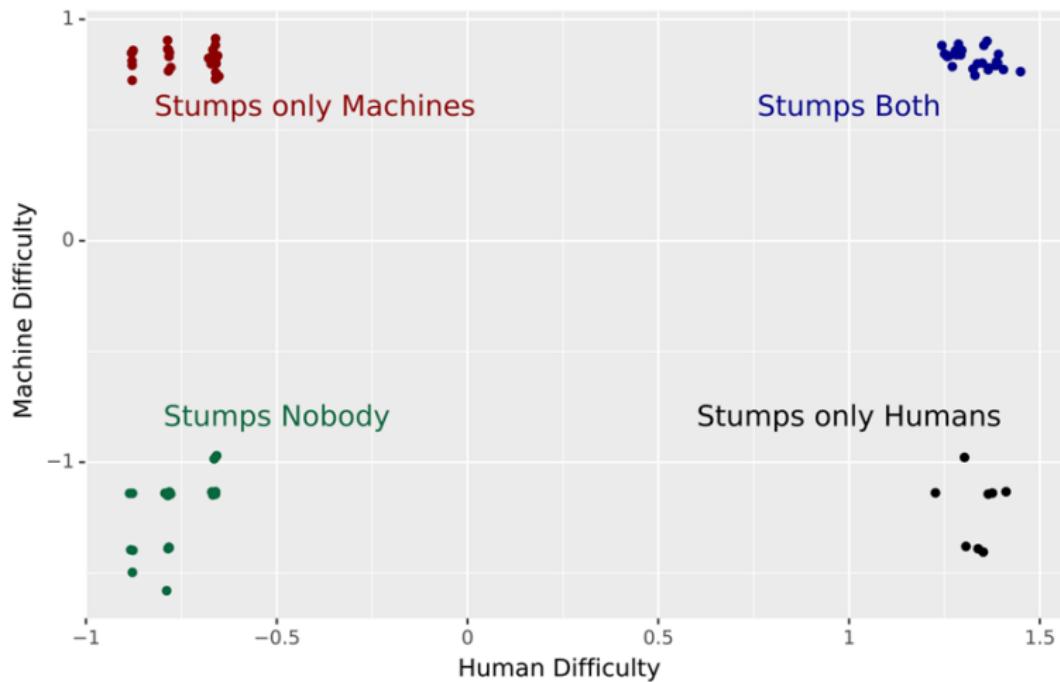
$$\delta_j = \sum_{i \sim H_{(1)}} \left[ \frac{\exp\left\{-\gamma_j \left[\beta_i^{H_{(1)}} - \theta_j\right]\right\} - p_{H_{(1)}}(r_{i,j})}{|H_{(1)}|} \right]. \quad (4)$$

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

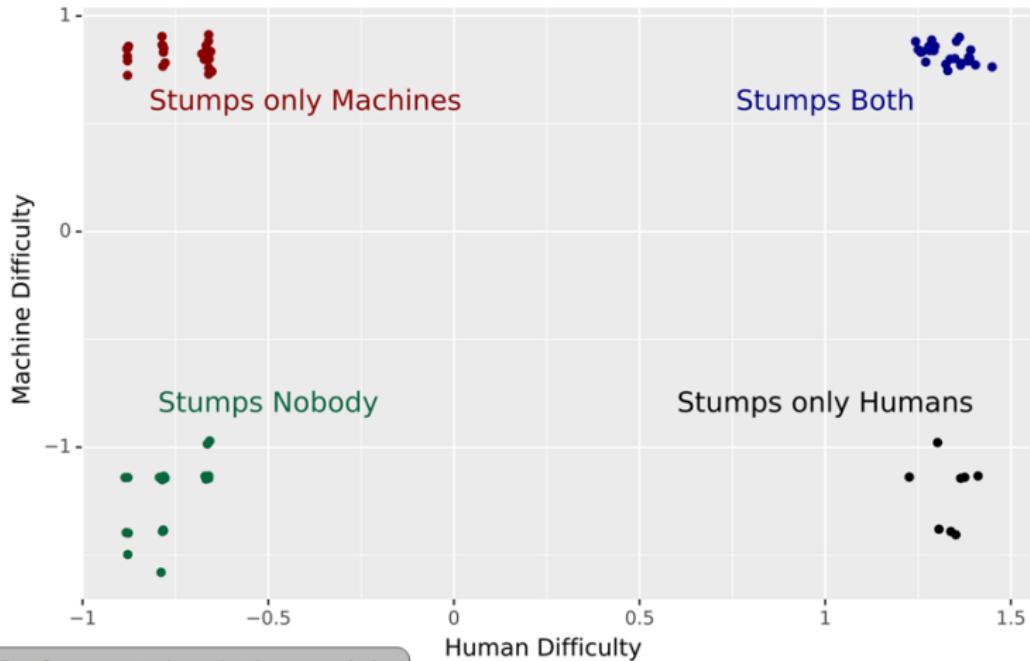
## AdvQA: Is this a viable incentive structure?

- 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

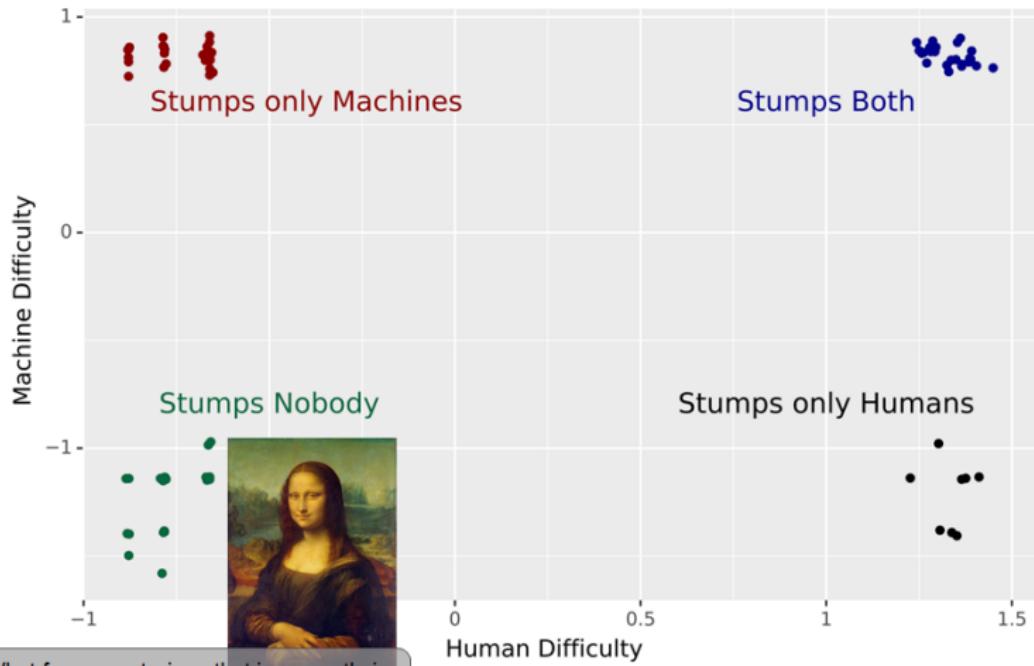


## What makes for Adversarial Example



What famous art piece that is currently in France is referred to as La Gioconda?

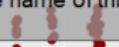
## What makes for Adversarial Example



What famous art piece that is currently in France is referred to as La Gioconda?

## What makes for Adversarial Example

This cheerful duo sings in a musical which repeats a phrase about the lack of trouble, in an East African language, while on the protagonist's journey home. What is the name of this duo?



Stumps only Machines



Stumps Both

Machine Difficulty

0

-1

-1

0

0.5

1

1.5

Stumps Nobody



Stumps only Humans

What famous art piece that is currently in France is referred to as La Gioconda?

Human Difficulty

## What makes for Adversarial Example

This cheerful duo sings in a musical which repeats a phrase about the lack of trouble, in an East African language, while on the protagonist's journey home. What is the name of this duo?



Machine Difficulty

0

Stumps Nobody

-1



-1

0

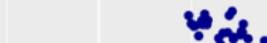
0.5

1

1.5

Human Difficulty

What famous art piece that is currently in France is referred to as La Gioconda?



Stumps Both



Stumps only Humans

## What makes for Adversarial Example

This cheerful duo sings in a musical which repeats a phrase about the lack of trouble, in an East African language, while on the protagonist's journey home. What is the name of this duo?



Stumps only Machines

Machine Difficulty

0

Stumps Nobody

-1



-1

What famous art piece that is currently in France is referred to as La Gioconda?

What is the name of the cricket team that is owned by the founder of Poomalaai and is considered to have one of the best bowling sides?



Stumps Both



Stumps only Humans

0

0.5

1

1.5

Human Difficulty

## What makes for Adversarial Example

This cheerful duo sings in a musical which repeats a phrase about the lack of trouble, in an East African language, while on the protagonist's journey home. What is the name of this duo?



Stumps only Machines

Machine Difficulty

0

-1

-1

Stumps Nobody



What famous art piece that is currently in France is referred to as La Gioconda?

What is the name of the cricket team that is owned by the founder of Poomaalai and is considered to have one of the best bowling sides?



Stumps Both



Stumps only Humans



Which of the first Adidas Yeezy Boost 350 designs had an out of this world themed name?

Human Difficulty

0

0.5

## 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?

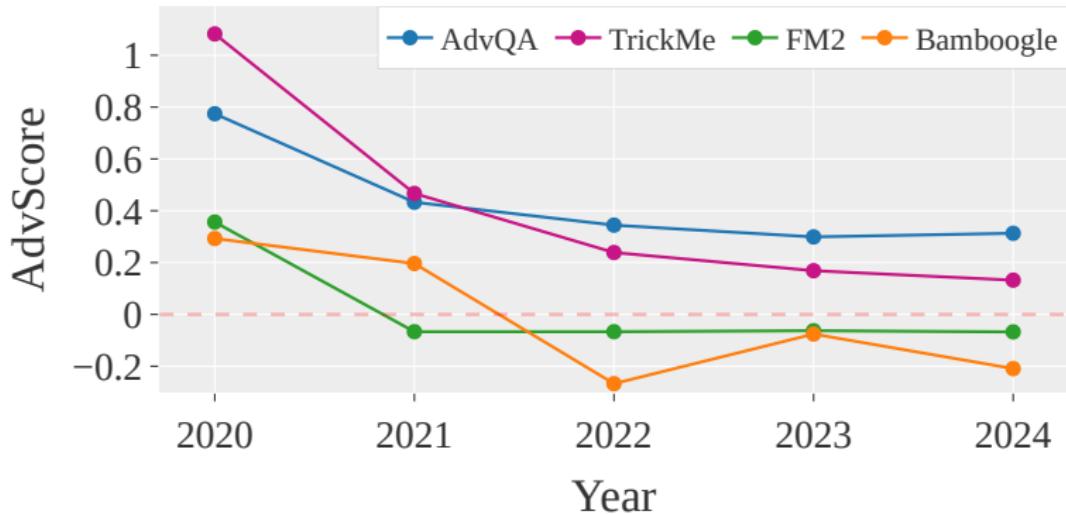
## Adversarial Strategies



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**

## 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: Human in the loop (expert), LLM model + category

початок об 11.00

Готель Radisson Blu  
м. Київ, проспект Голославів Вал 22

Олена  
БОЙЧУН



## 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	Country
1	-15	Summer Dew	1/5
2	475	rmunizmidtown	54
3	285	Cottman	40

## 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

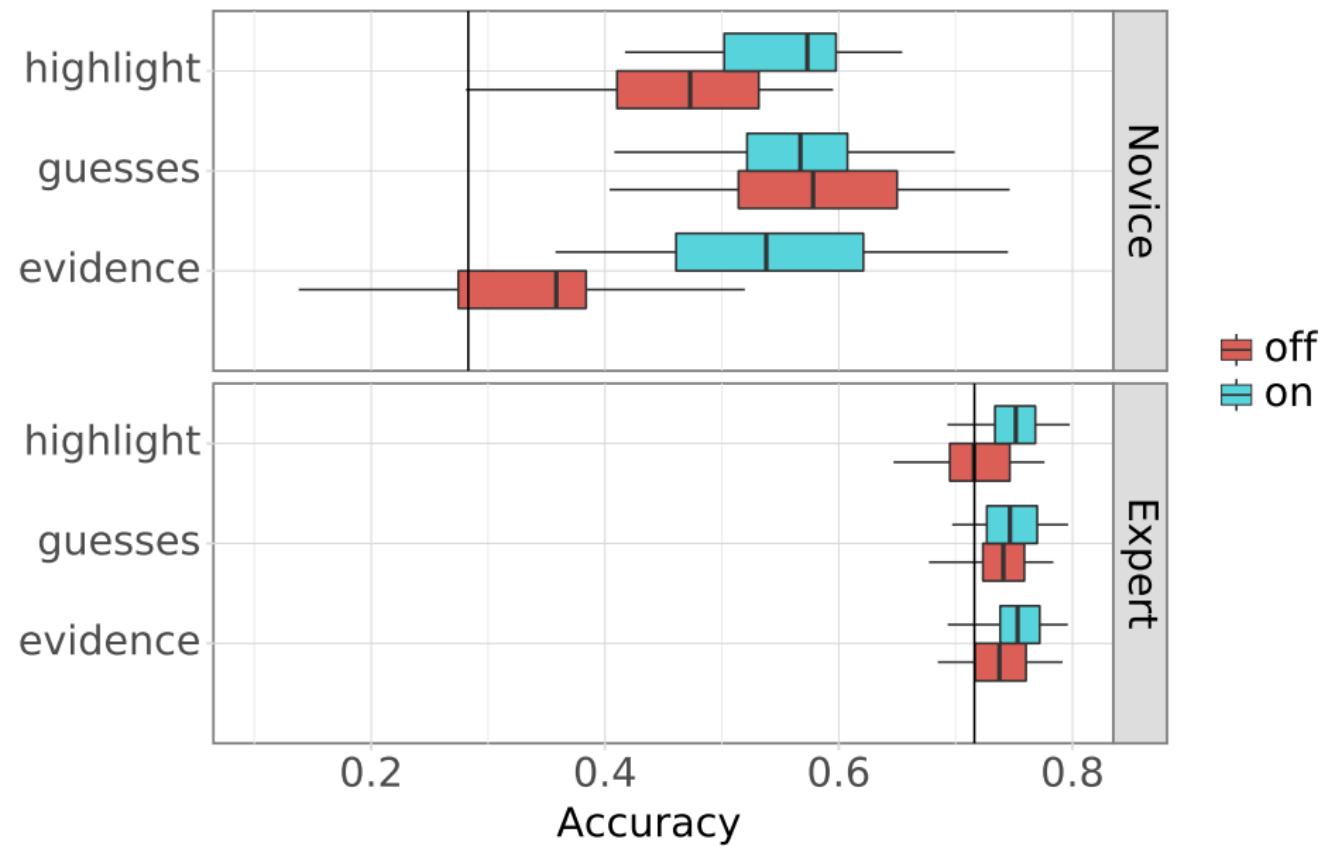
## Experts vs. Novices

### 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**

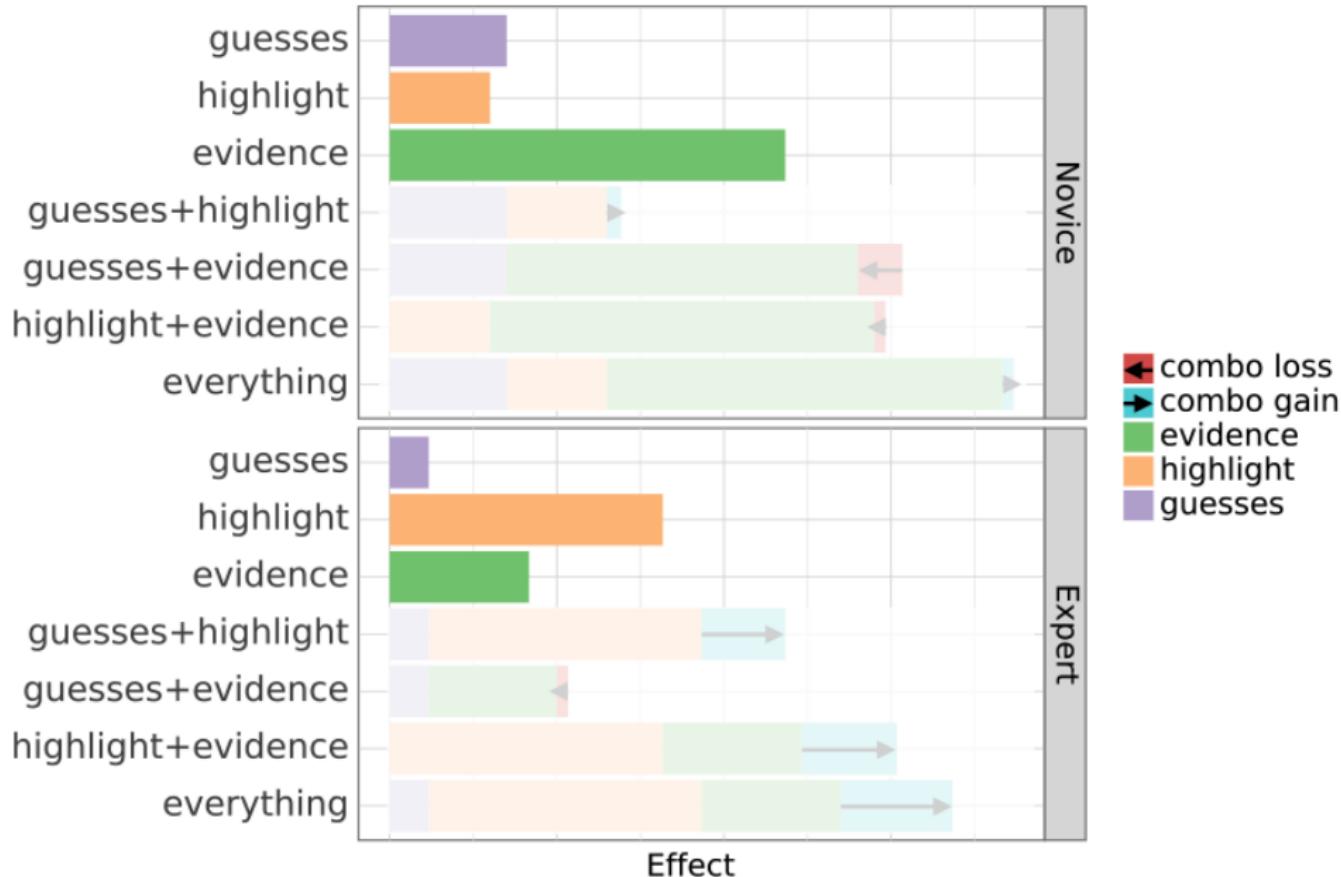
Rerun IRT analysis: rather than contrast human vs. computer skill, we see how much interface boosts (or diminishes) user skill.

## 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!

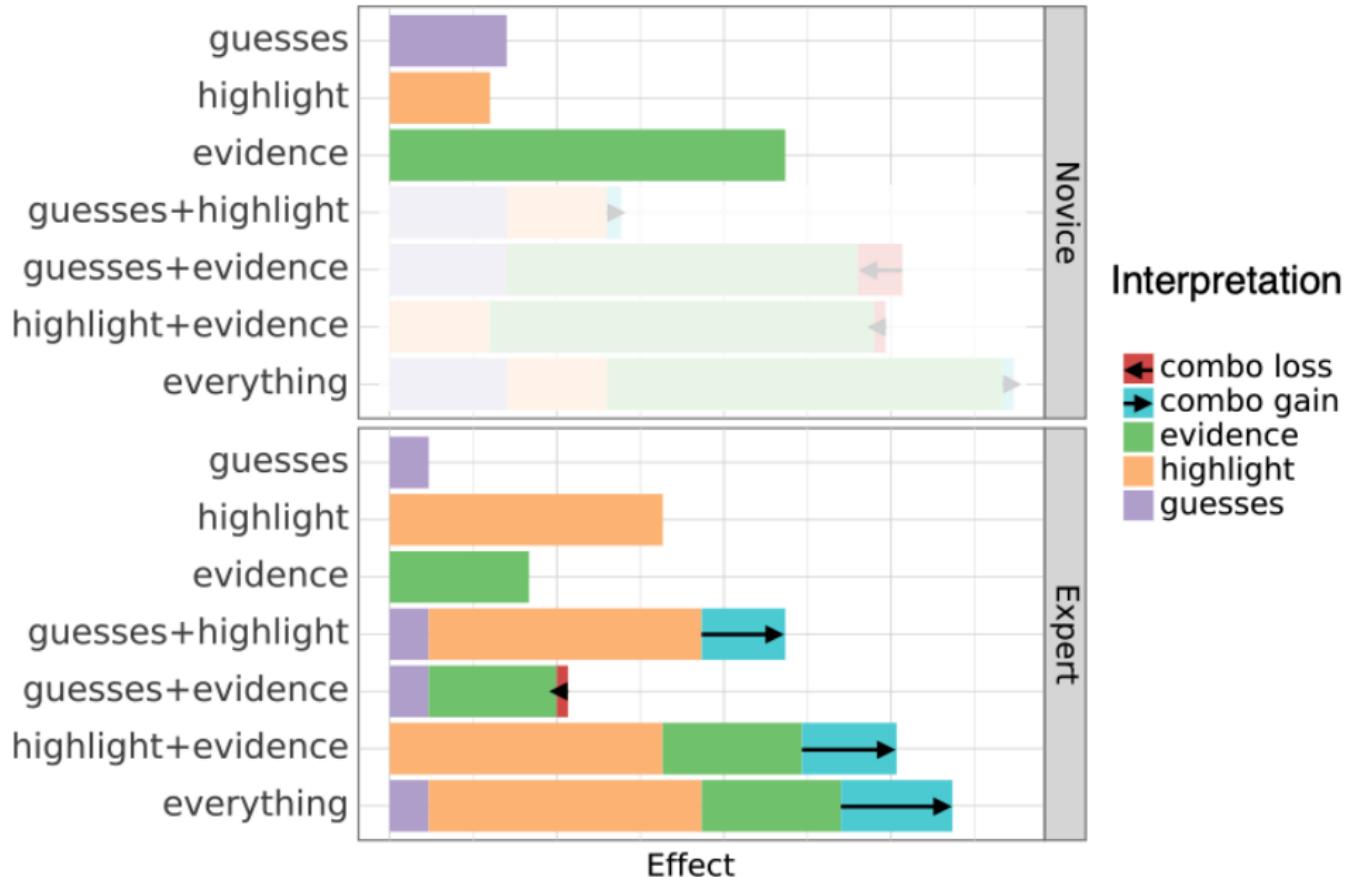
- **Big, Positive:** Help
- **Big, Negative:** Hurt
- **Small:** Neutral



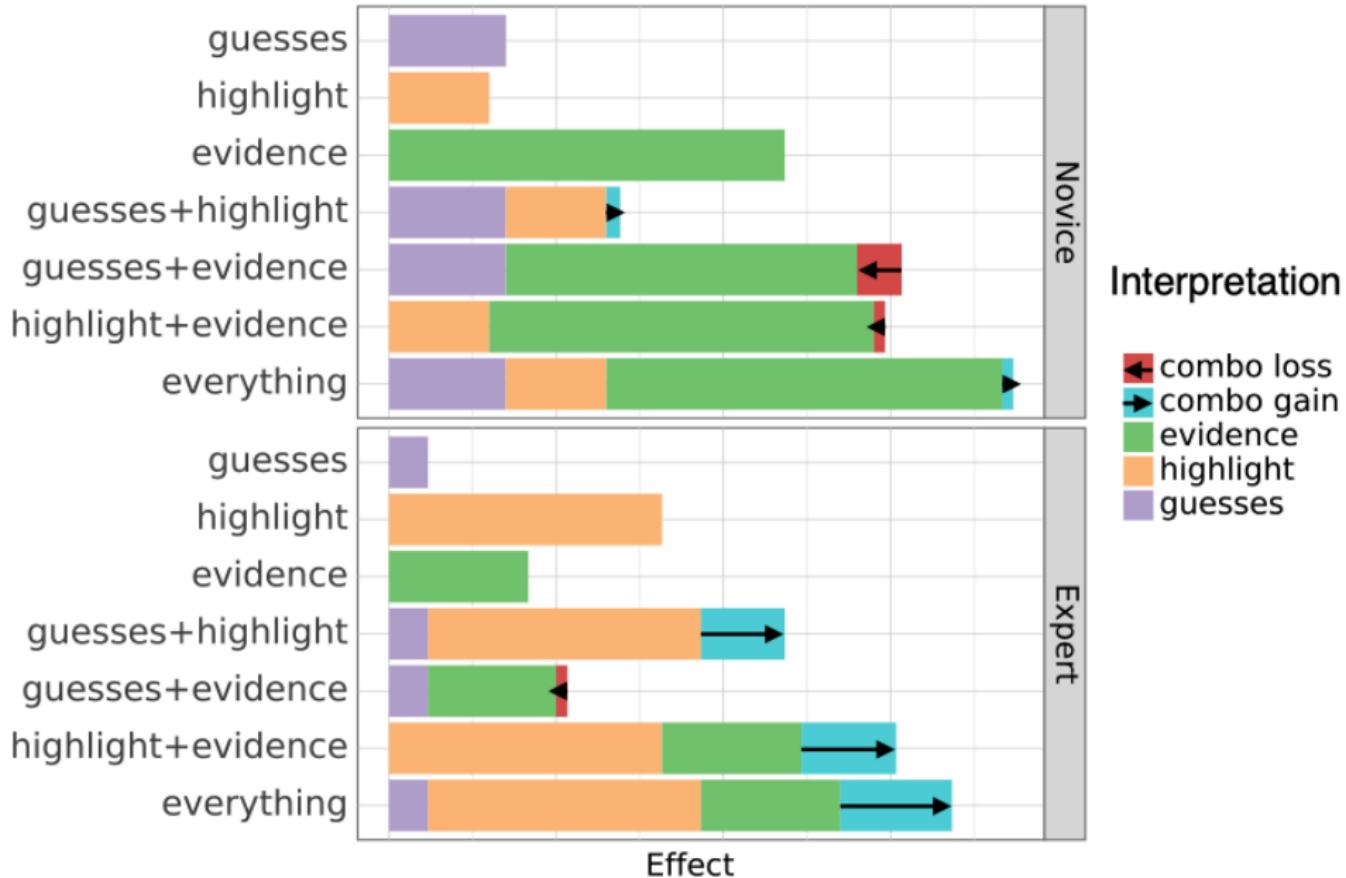
Everything helps: Evidence for novices, Highlight for experts



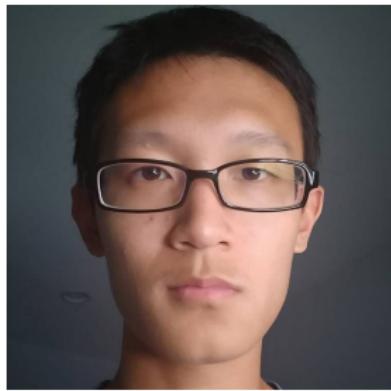
Synergistic effects



Highlight and evidence help experts most



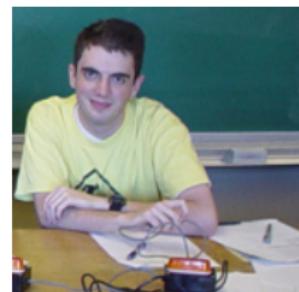
For novices, less synergy



Learning to Explain Selectively

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

## Measuring Interpretability

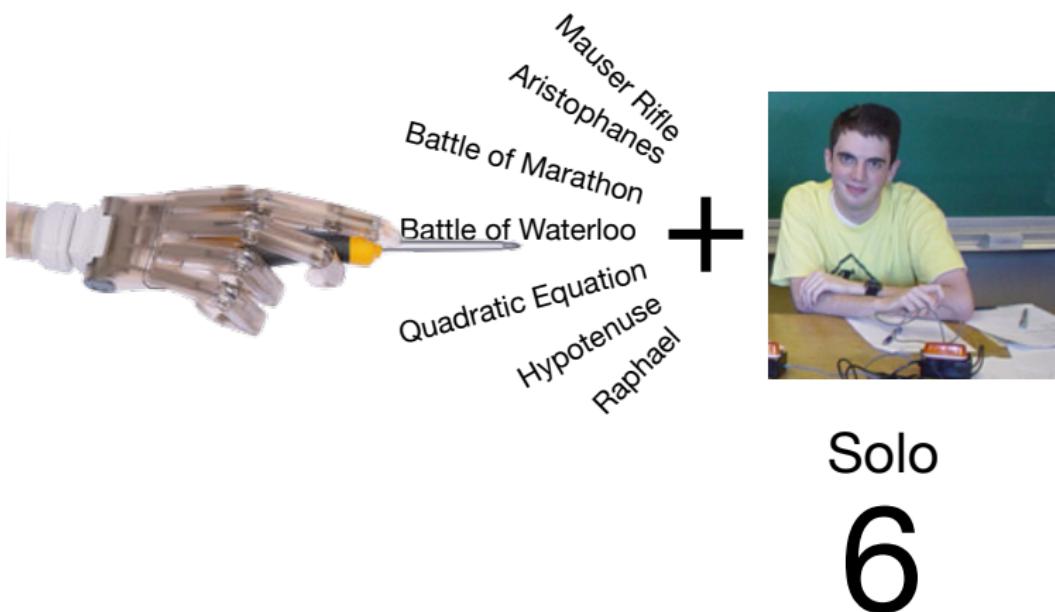


## Measuring Interpretability

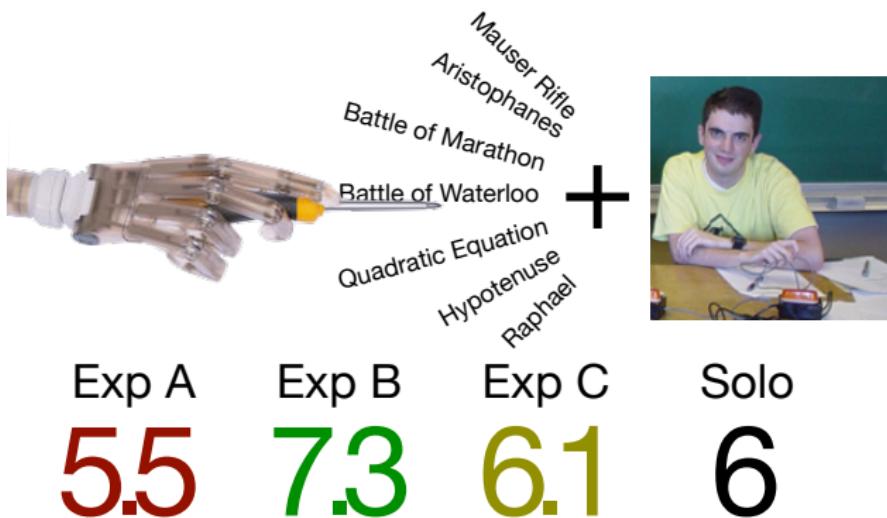


Solo  
**6**

## Measuring Interpretability

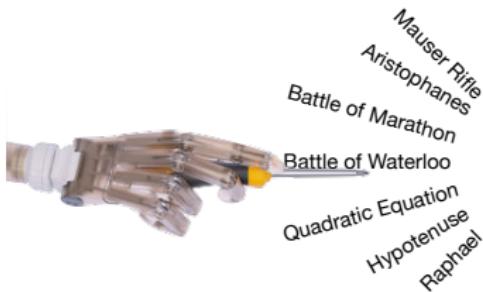


## Measuring Interpretability



## Improvement through Bandit Algorithms

### Visualization

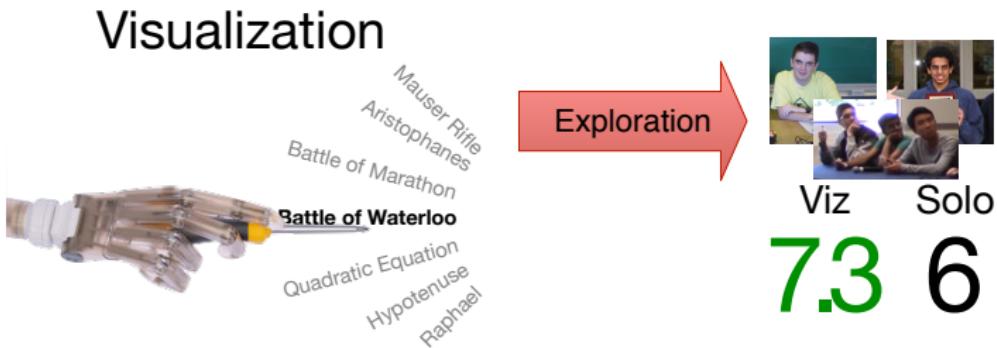


Viz Solo  
**7.3 6**

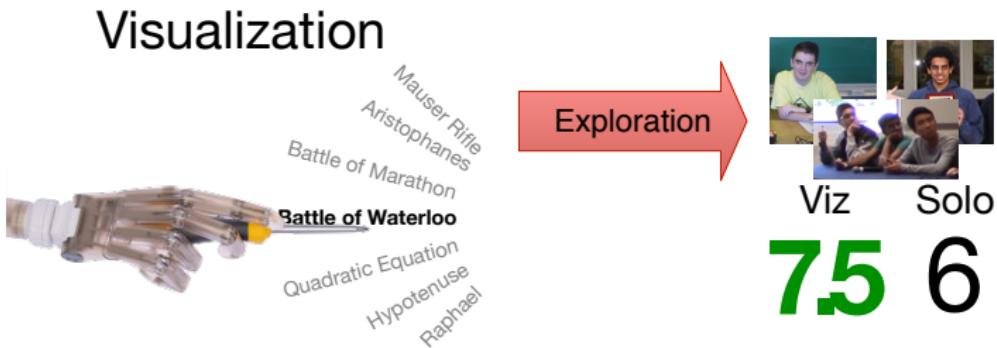
## Improvement through Bandit Algorithms



## Improvement through Bandit Algorithms



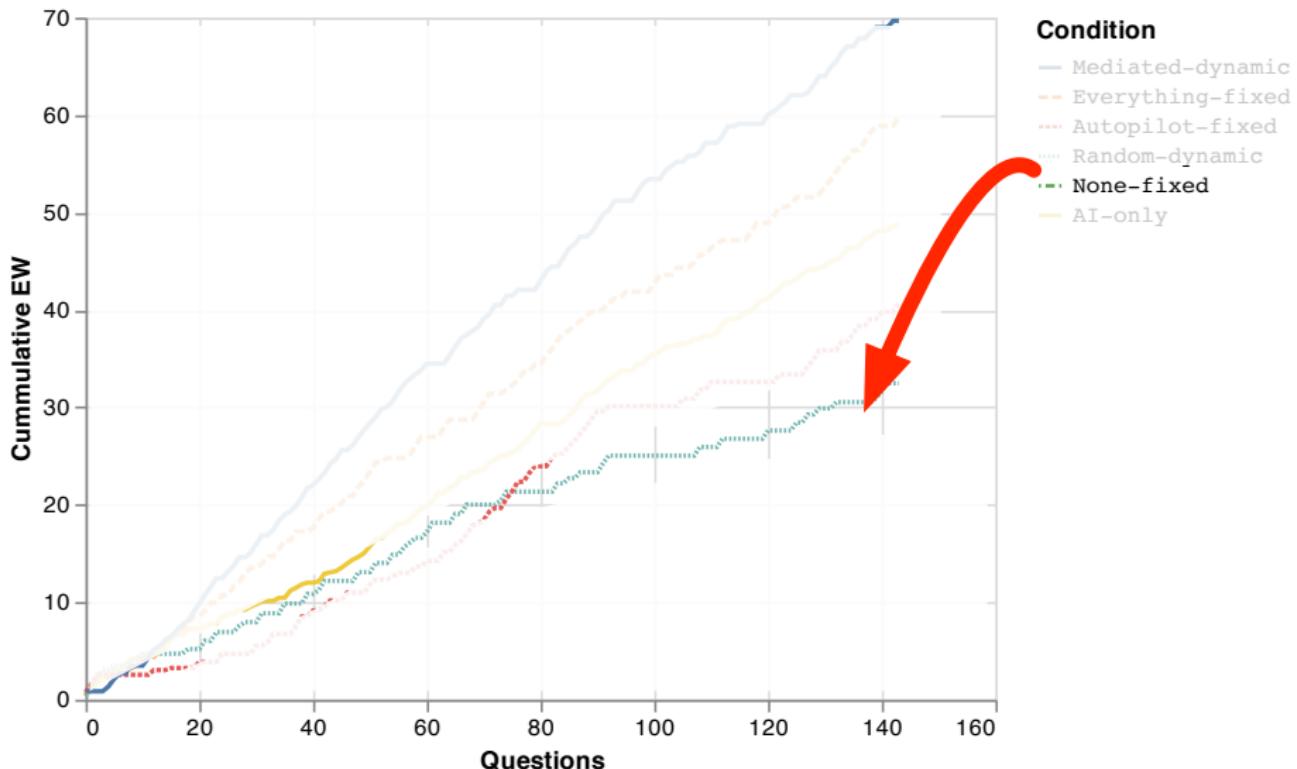
## Improvement through Bandit Algorithms



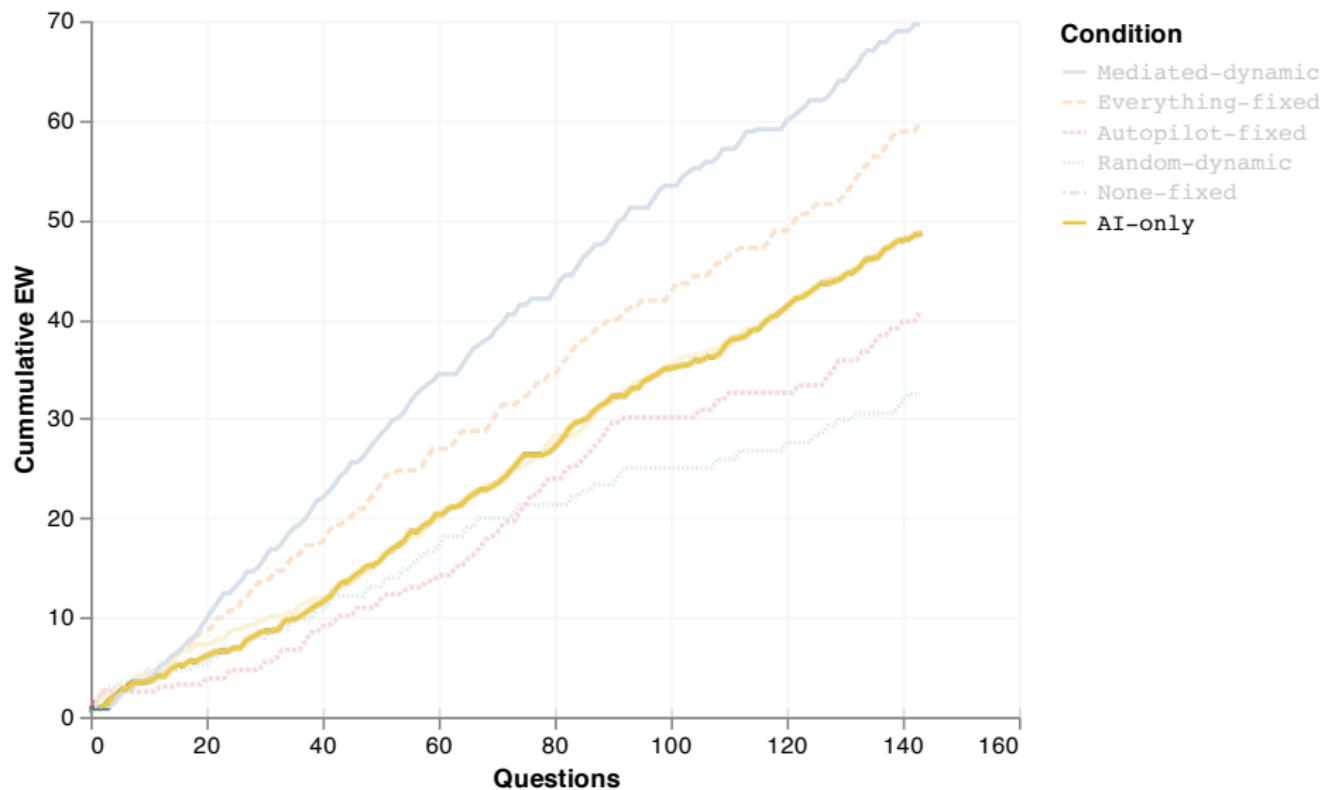
## Improvement through Bandit Algorithms



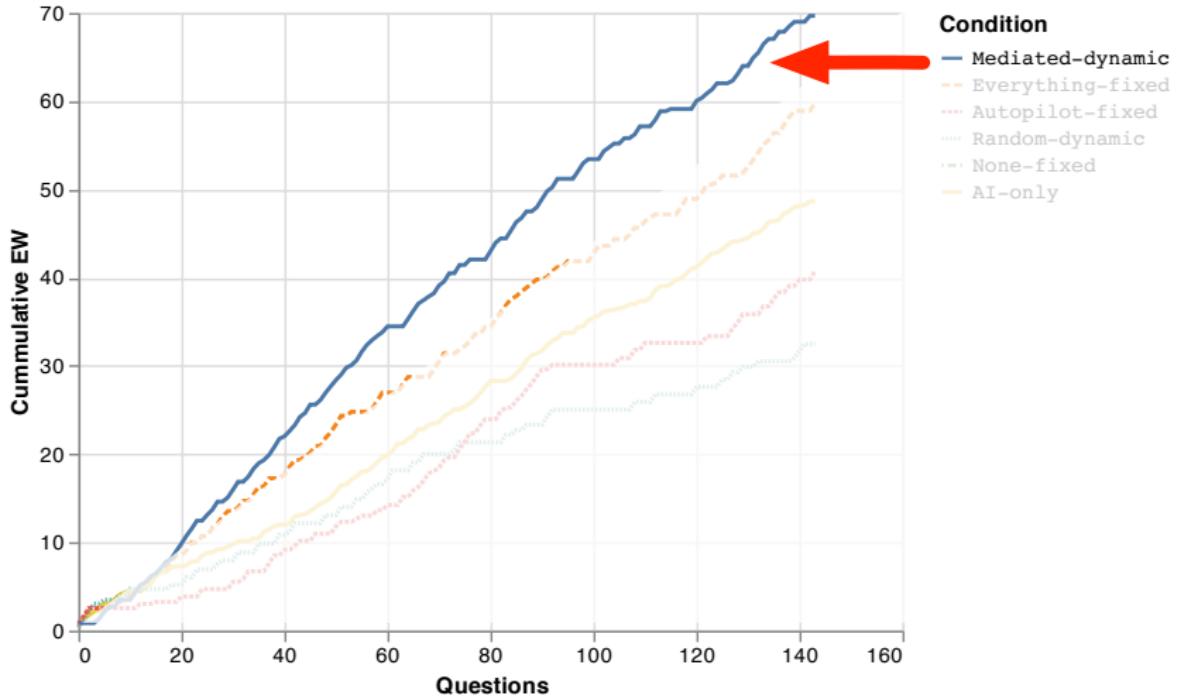
Bandit actions [?]: turn each of the explanations (Guess, Highlight, Evidence) on or off.



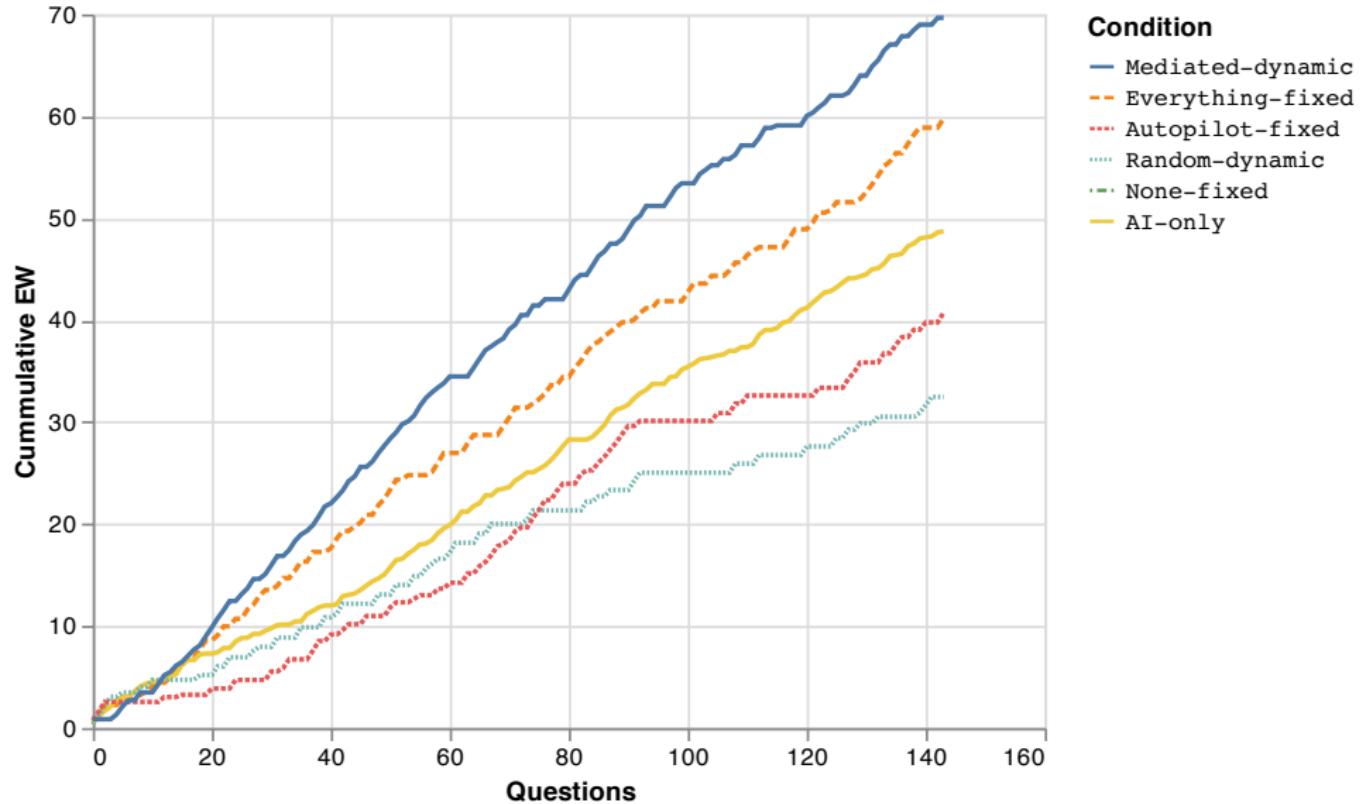
Human alone without an AI teammate



AI alone without a human teammate



Dynamic assistance to human



Better than showing everything!

# Changing Gears

≡ Science

Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 378, NO. 6624 > HUMAN-LEVEL PLAY IN THE GAME OF DIPLOMACY BY COMBINING LANGUAGE MODELS WITH STRATEGIC REASONING

RESEARCH ARTICLE COMPUTER SCIENCE

f X in

## Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning

MITA FUNDAMENTAL AI RESEARCH DIPLOMACY TEAM (FAIR) · ANTON BAKHTIN · ADAM BROWN · EMILY DRAHN · GABRIELE FARINA · COLIN FLAHERTY · DANIEL FRIED · ANDREW GOET · JONATHAN GRAY · I-I AND MARILUS ZULASTRO · +17 authors · Authors Info & Affiliations

SCIENCE · 22 Nov 2022 · Vol 378, Issue 6624 · pp. 1067-1074 · DOI: 10.1126/science.adc6024

103,390 99 89

**AI masters Diplomacy**

The game *Diplomacy* has been a major challenge for artificial intelligence (AI). Unlike other competitive games that AI has recently mastered, such as chess, Go,

- 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.

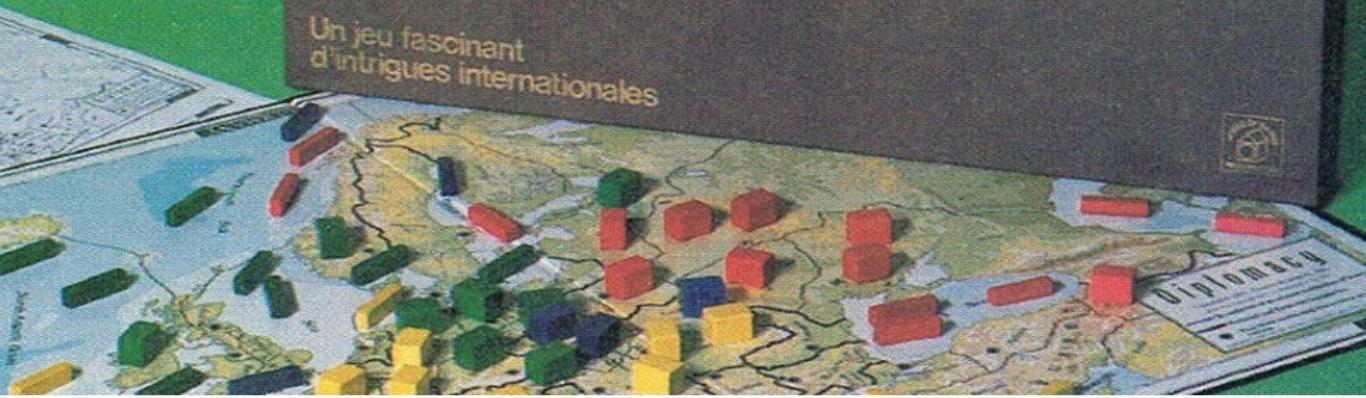
*Association for Computational  
Linguistics, 2015*

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  
ruins friendships"



Un jeu fascinant  
d'intrigues internationales

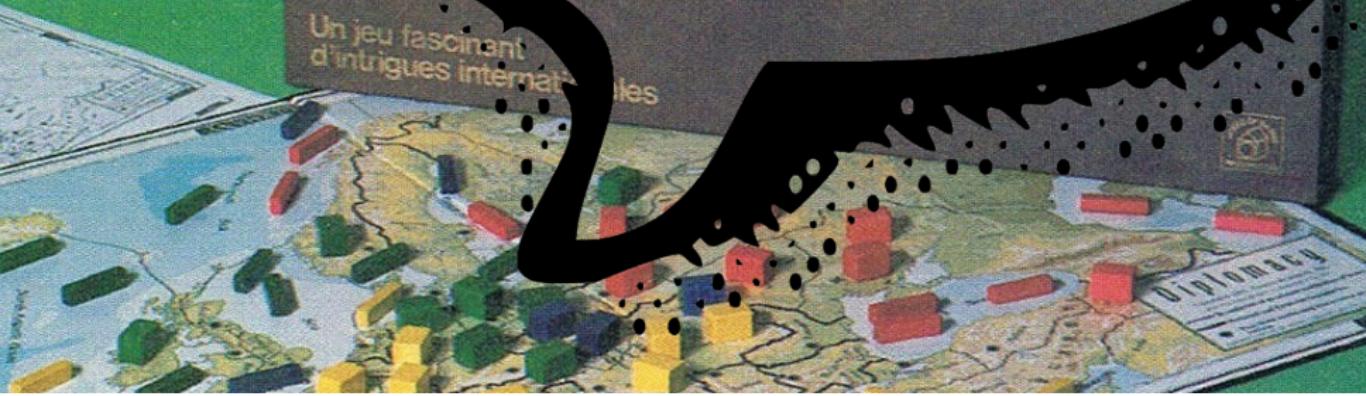


The exciting game  
of international intrigue

"The game that  
ruins friendships"

# Diplomacy

Un jeu fascinant  
d'intrigues internationales



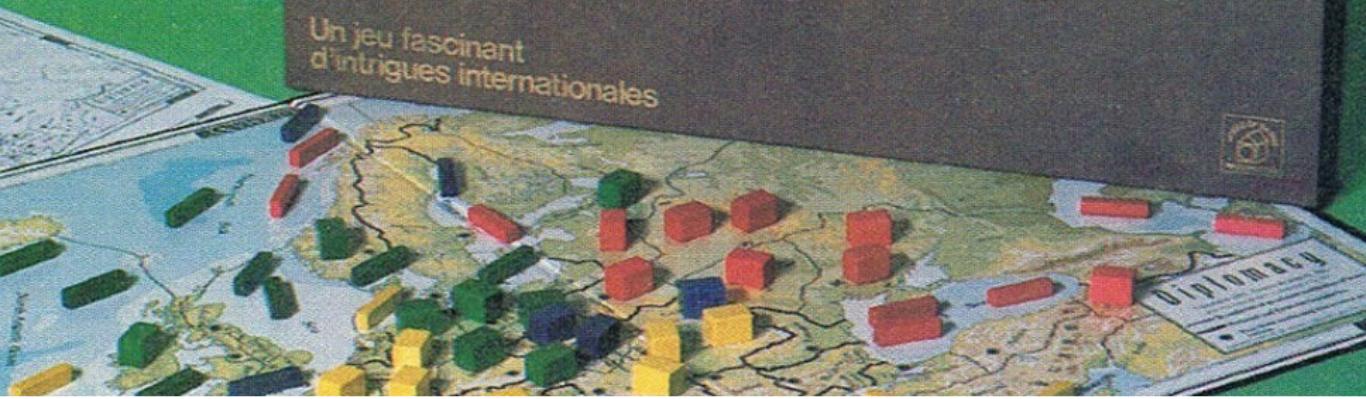
The exciting game  
of international intrigue

"The game that  
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](http://diplom.org); [usak.asciiiking.com](http://usak.asciiiking.com)





Diplomacy  
by Allan B. Calhamer  
Copyright 1995, Avalon Hill  
Map by J. Piatkus, III





A map of Central Europe showing various regions and major cities. Superimposed on the map are several green and brown cartoon-style dinosaurs. One green Stegosaurus is located in the northern part of Germany near Berlin. Another green Stegosaurus is in the eastern part of the map, near Budapest. Two brown Brachiosaurus-like dinosaurs are positioned in the southern and southeastern parts of the map, one near Trieste and another near Bulgaria. The map features a light beige background with greyish-blue oceans. Regions are outlined in grey, and major cities like Berlin, Vienna, and Budapest are labeled. The map also shows the Alpine mountain range and the Black Sea coastline.



A map of Central Europe showing political boundaries and major cities. The regions are color-coded in shades of grey, pink, and blue. Major cities like Berlin, Vienna, and Budapest are labeled. Several green and brown dinosaurs are scattered across the map, particularly in the northern and eastern areas. A black arrow points from the word 'Vienna' to a green Stegosaurus-like dinosaur located near the city of Vienna.





help?



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.



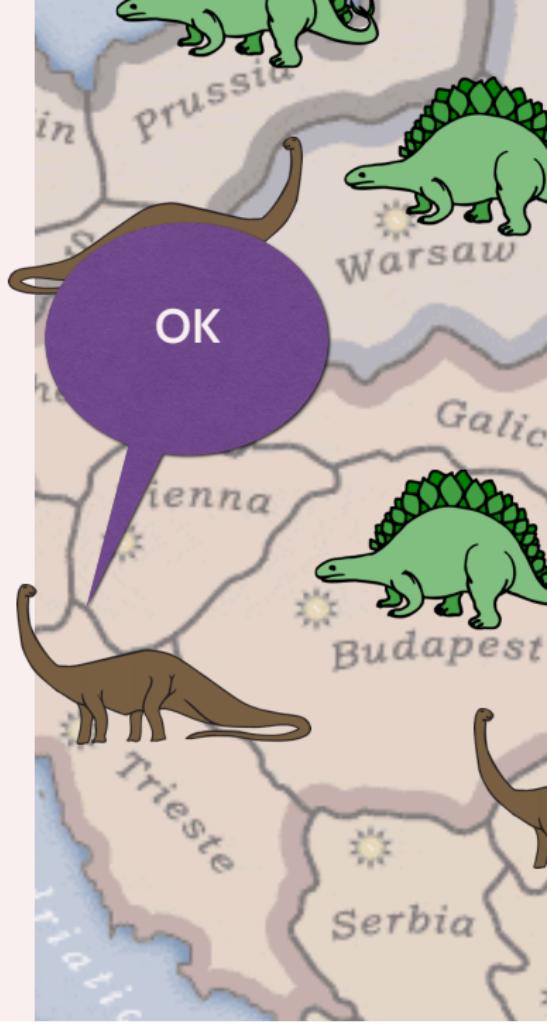
F



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.



F

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.

F



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.



F

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

E

...



**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

E



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.

F

E

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.

F



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

E

...



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

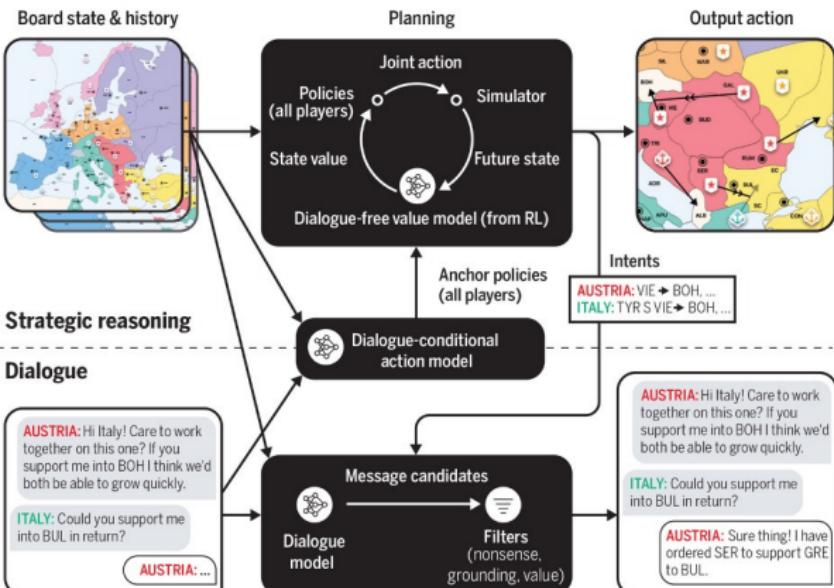




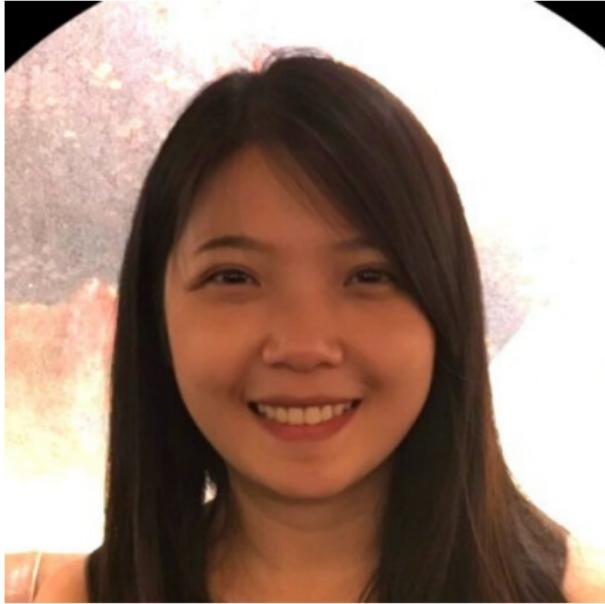
A man with a shocked expression looks down at two toy dinosaurs. He has his mouth open and wide eyes. In the foreground, a hand holds a purple Stegosaurus toy on the left, and another hand holds a yellow Tyrannosaurus Rex toy on the right. The background is dark.

Curse your sudden  
but inevitable  
betrayal!

## Cicero is really good



Rank	Average Score	#
1	0.4918	
2	0.4912	
3	0.4528	
4	0.4516	
5	0.3734	
6	0.3505	
7	0.3491	
8	0.3035	
9	0.2754	



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

Wichayaporn Wongkamjan<sup>1</sup> Feng Gu<sup>1</sup> Yanze Wang<sup>4</sup> Ulf Hermjakob<sup>4</sup>  
Jonathan May<sup>4</sup> Brandon M. Stewart<sup>2</sup> Jonathan K. Kummerfeld<sup>3</sup>

Denis Peskoff<sup>2</sup> Jordan Lee Boyd-Graber<sup>1</sup>

<sup>1</sup>University of Maryland    <sup>2</sup>Princeton University    <sup>3</sup>University of Sydney

<sup>4</sup>Information Sciences Institute, University of Southern California

{wwongkan,fgu1}@umd.edu {yanzewan,ulf,jonmay}@isi.edu

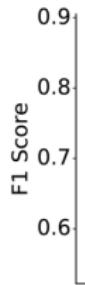
bms4@princeton.edu jonathan.kummerfeld@sydney.edu.au

dp2896@princeton.edu jbg@umiacs.umd.edu

## Can they tell it's a bot?

### Passing the Turing Test?

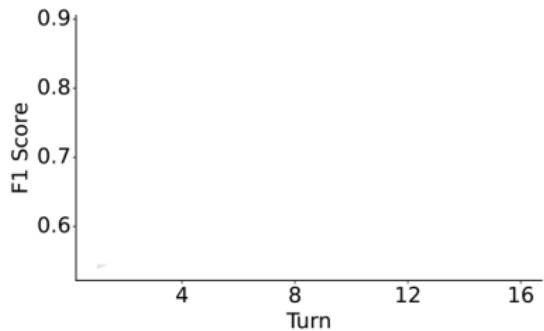
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



## Can they tell it's a bot?

### Passing the Turing Test?

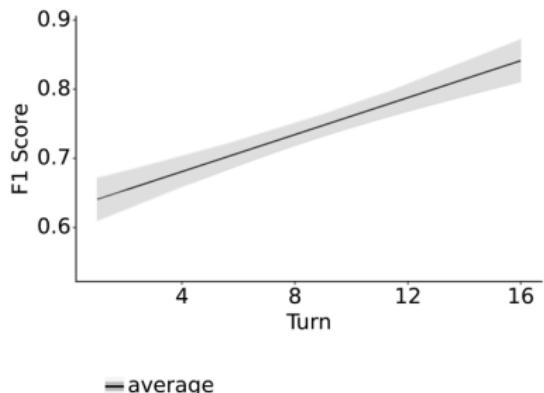
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



## Can they tell it's a bot?

### Passing the Turing Test?

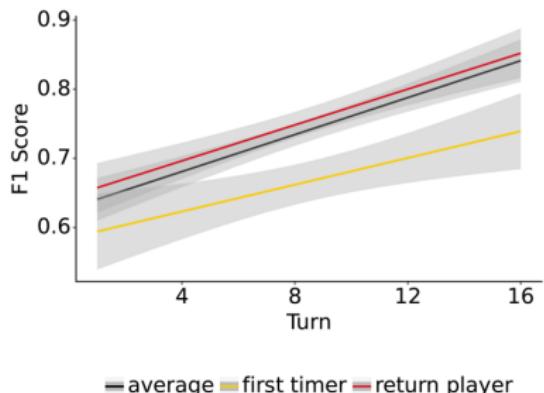
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



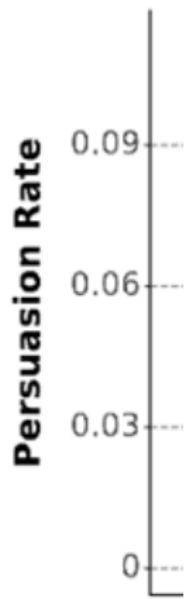
## Can they tell it's a bot?

### Passing the Turing Test?

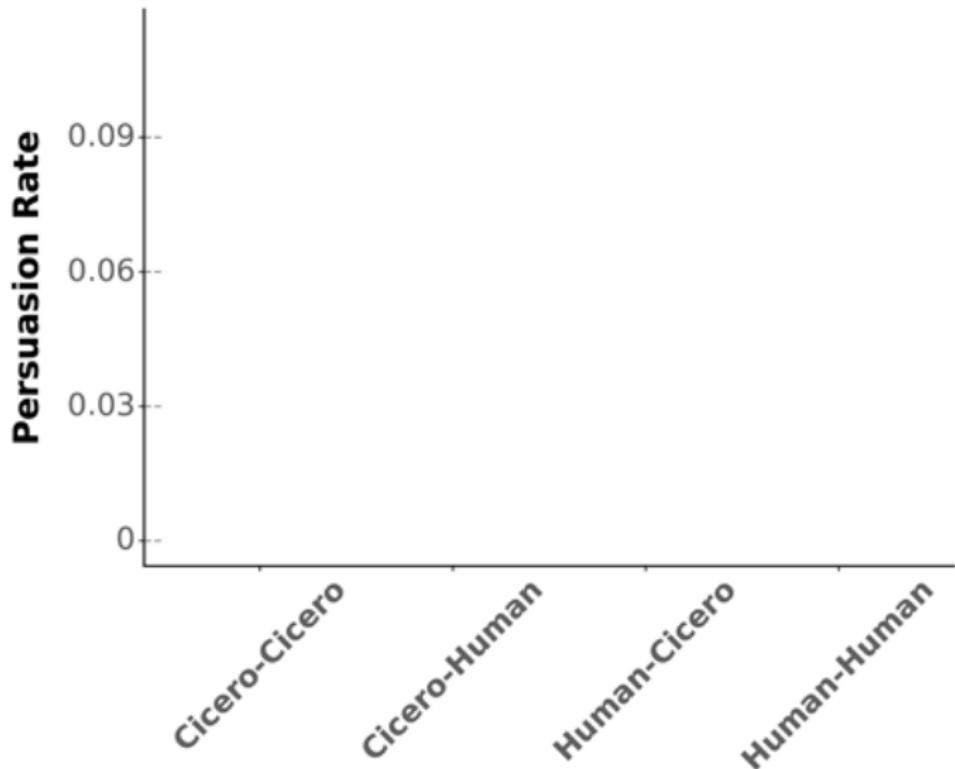
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



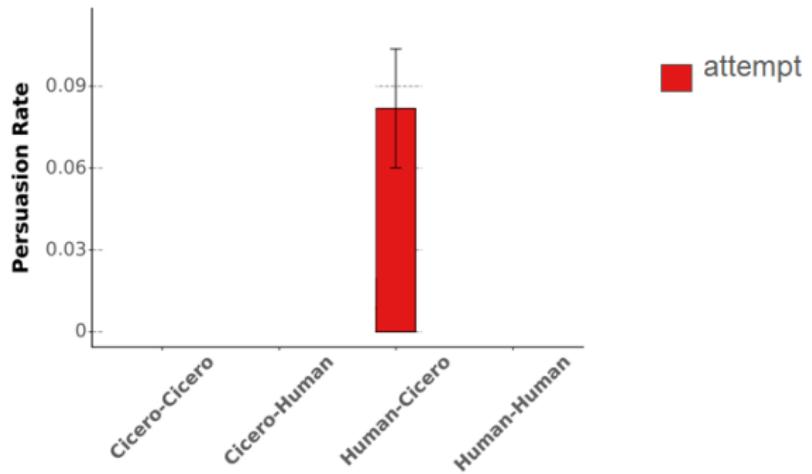
## Persuasion



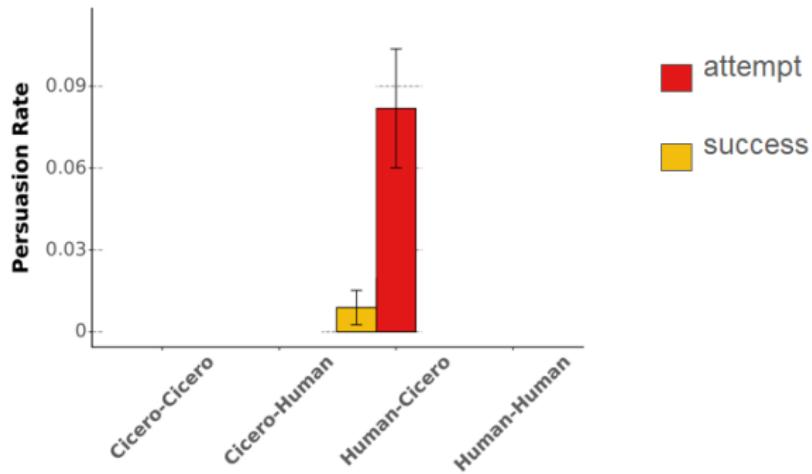
## Persuasion



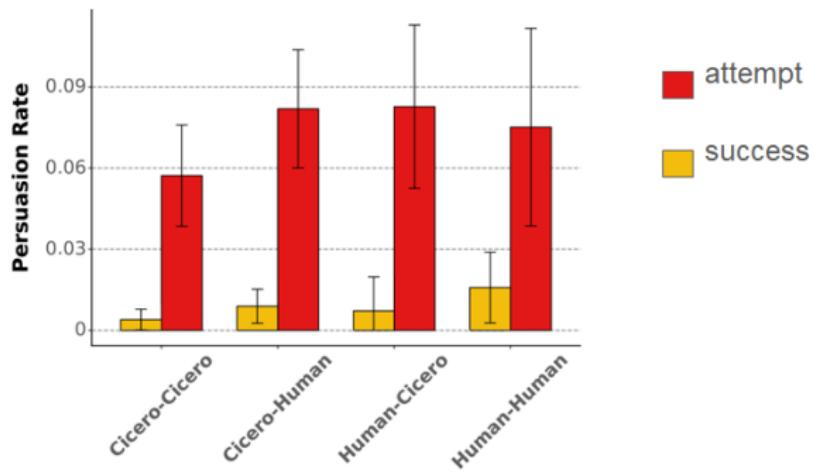
## Persuasion



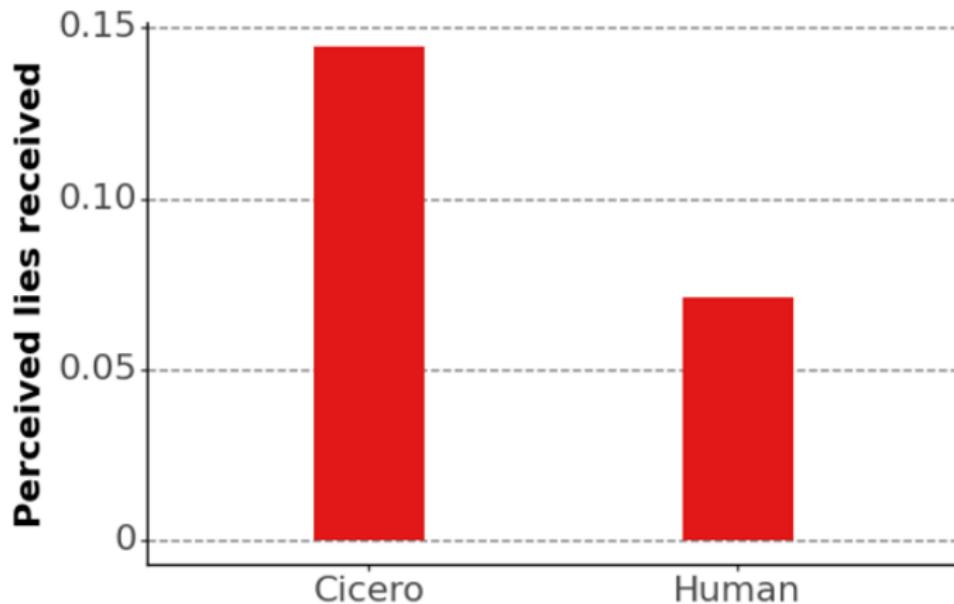
## Persuasion



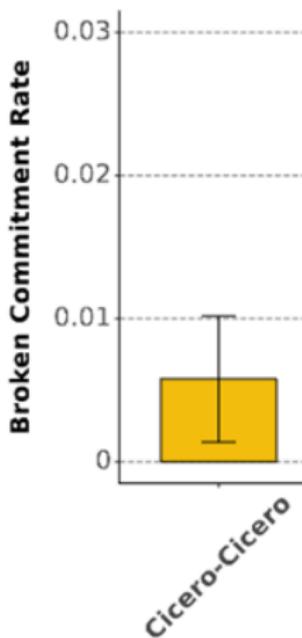
## Persuasion



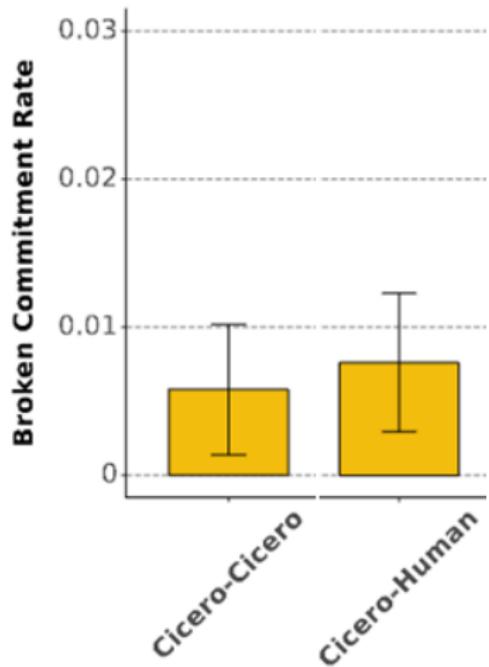
## People Think Cicero Lies More



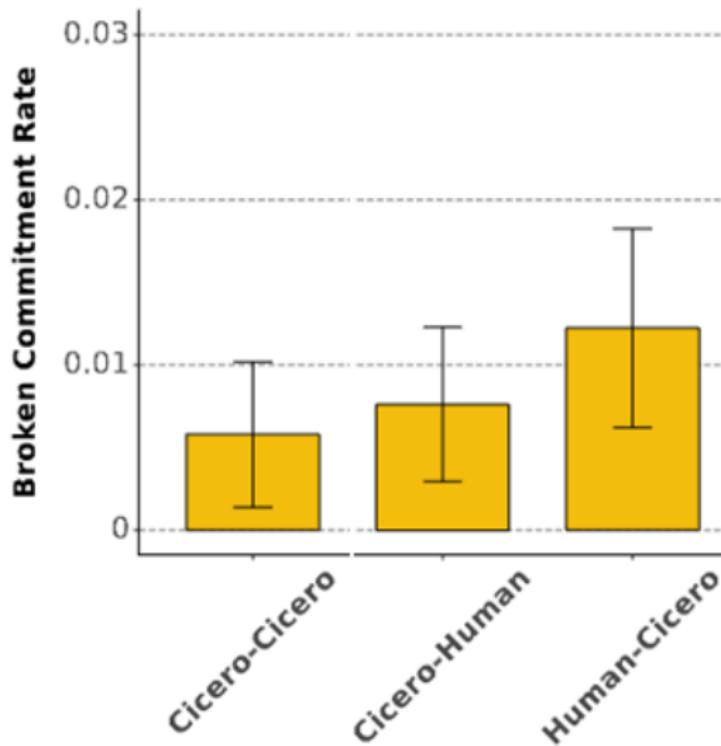
## Broken Commitments (subset of deception)



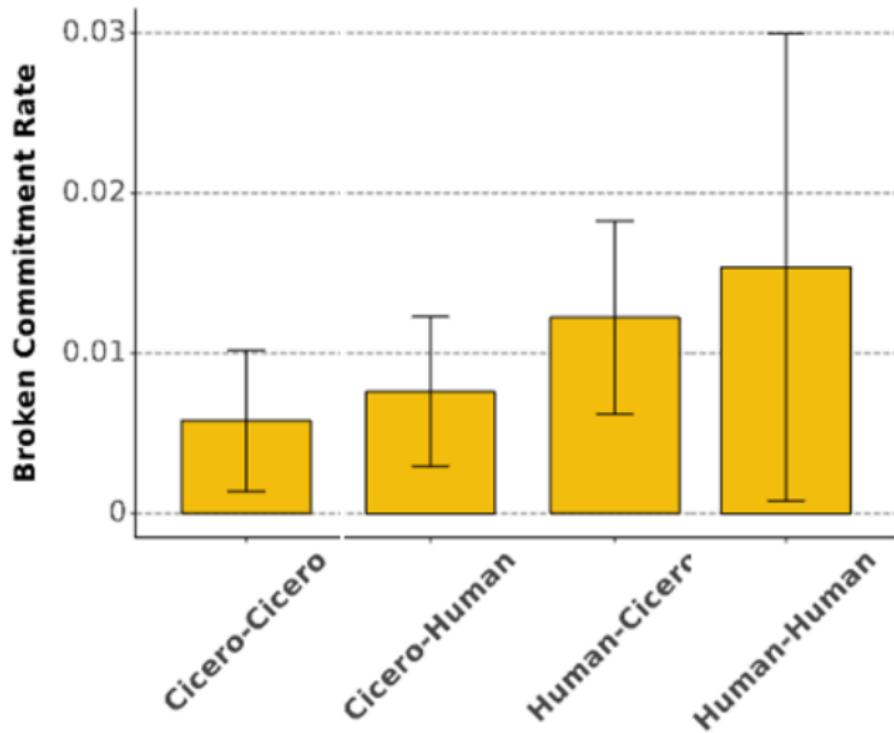
## Broken Commitments (subset of deception)



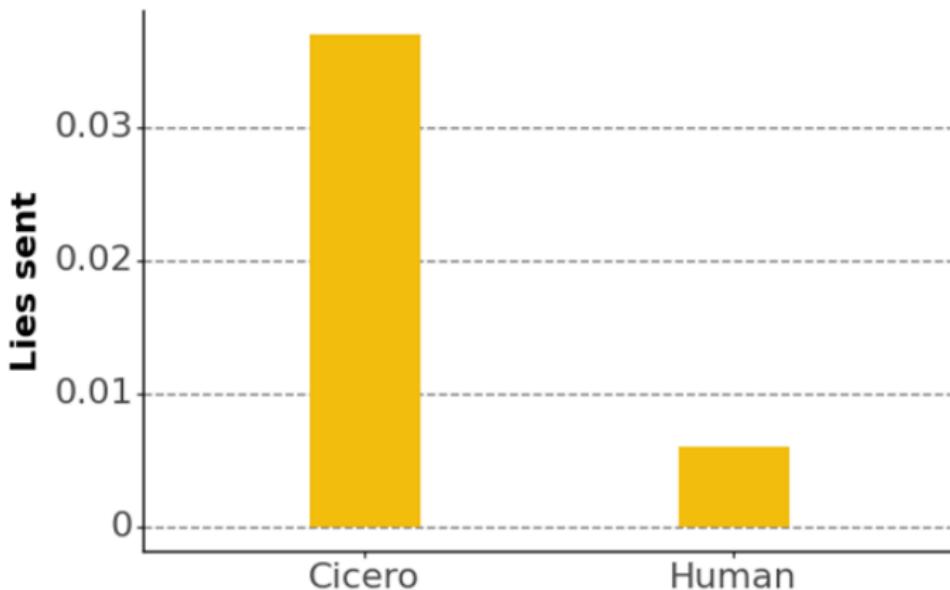
## Broken Commitments (subset of deception)

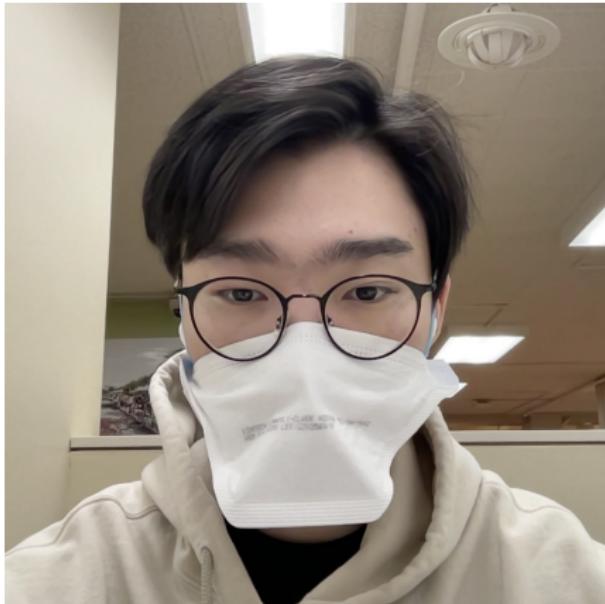


## Broken Commitments (subset of deception)



## People Lie to Cicero More





## Personalized Help for Optimizing Low-Skilled Users' Strategy

Feng Gu<sup>1</sup>    Wichayaporn Wongkamjan<sup>1</sup>    Jonathan K. Kummerfeld<sup>2</sup>  
Denis Peskoff<sup>3</sup>    Jonathan May<sup>4</sup>    Jordan Lee Boyd-Graber<sup>1</sup>

<sup>1</sup>University of Maryland    <sup>2</sup>University of Sydney

<sup>3</sup>Princeton University

<sup>4</sup>Information Sciences Institute, University of Southern California  
{fgu1, wwongkam}@umd.edu    jbg@.umiacs.umd.edu

# Giving Advice

RUS: 4 AUS: 3 ENG: 3 FRA: 3 GER: 3 ITA: 3 TUR: 3

Orders:

Create order:

move (M)  support (S)  hold (H)

Orderable locations: BUD, TRI, VIE

[3/3] moves have been set.

AUSTRIA ready

A BUD - SER [S]  
A VIE - GAL [S]  
F TRI - ALB [S]

ENGLAND  
FRANCE  
GERMANY  
ITALY  
RUSSIA  
TURKEY

Truth Lie Neutral

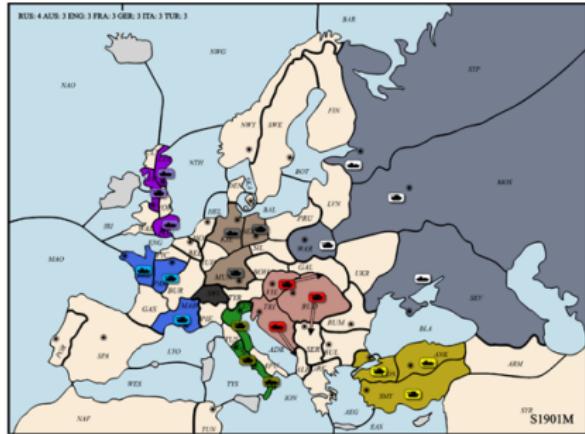
SI90IM

SI90IM

Hope you have a great game Italy!

Can we DMZ ADR?

# Giving Advice



## Create order:

move (M)  support (S)  hold (H) reset unready

**Orderable locations:** BUD, TRI, VIE

[3/3] moves have been set.

## Orders:

reset delete all update

### AUSTRIA ready

**A BUD - SER [S]**

**A VIE - GAL [S]**

**F TRI - ALB [S]**

You are getting advice:  
message, move.

Full:	+ all	X
A VIE - GAL	.	.
A BUD - SER	.	.
A TRI - VEN	.	.

Move  
Advice

## MESSAGE ADVICE

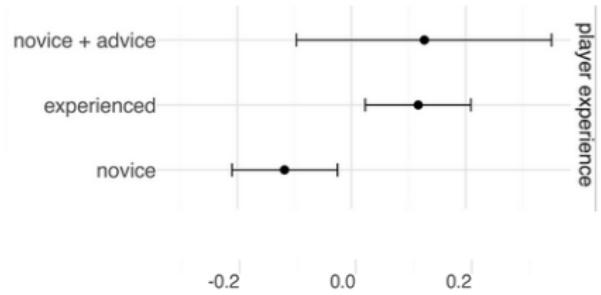
### Messages Advice to ITALY

Are you interested in Lepanto?

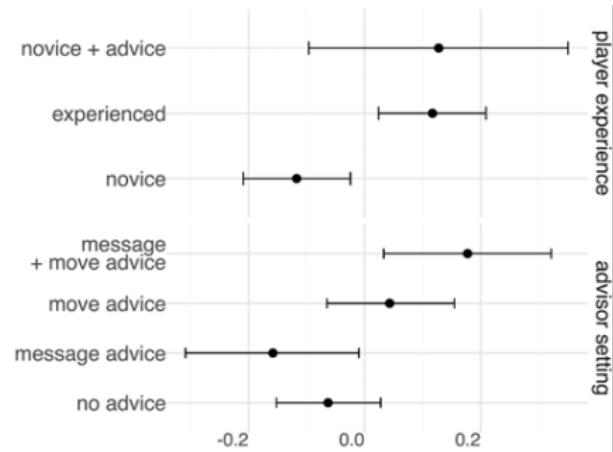
add to textbox dismiss

Per-power  
Message Advice

## Advice Helps



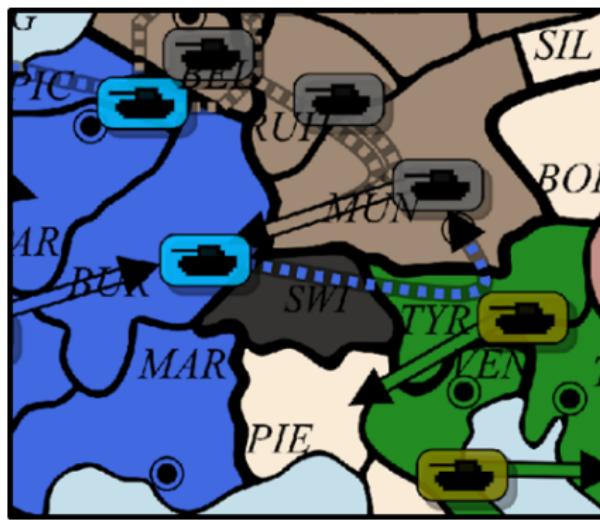
## Advice Helps



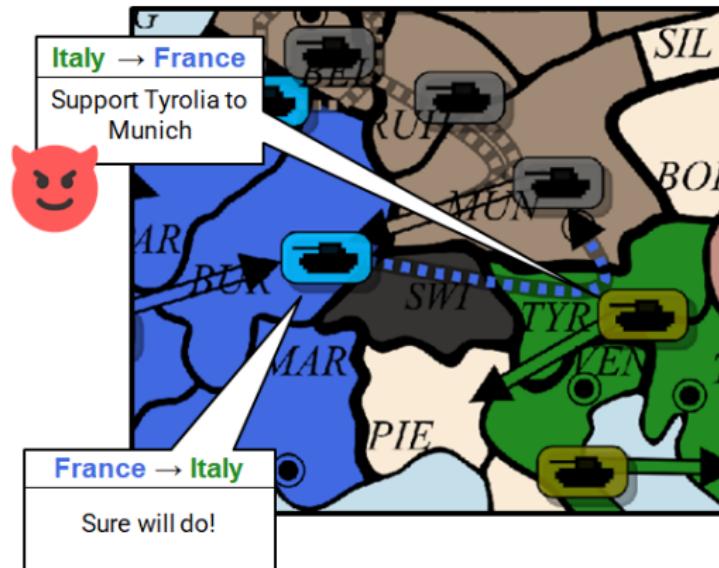
## Advice Helps

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

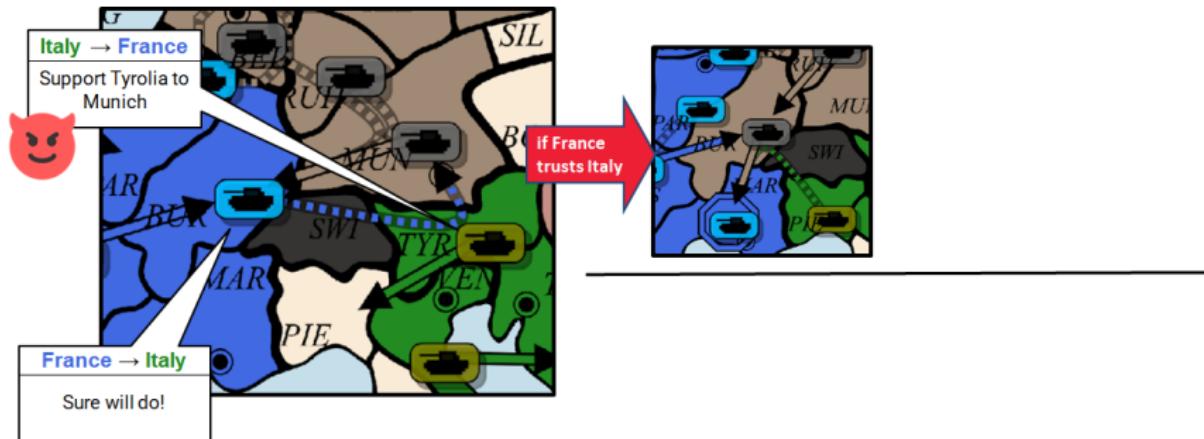
## How to better detect Deception



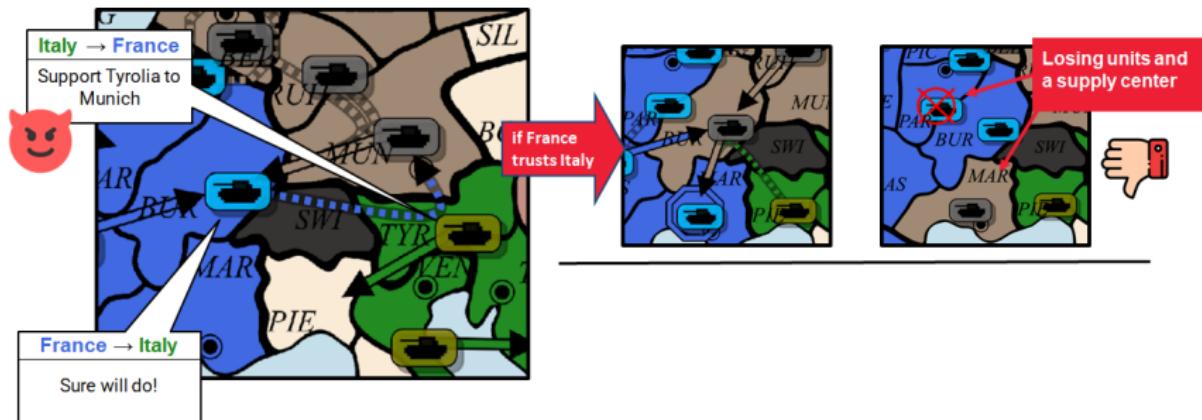
## How to better detect Deception



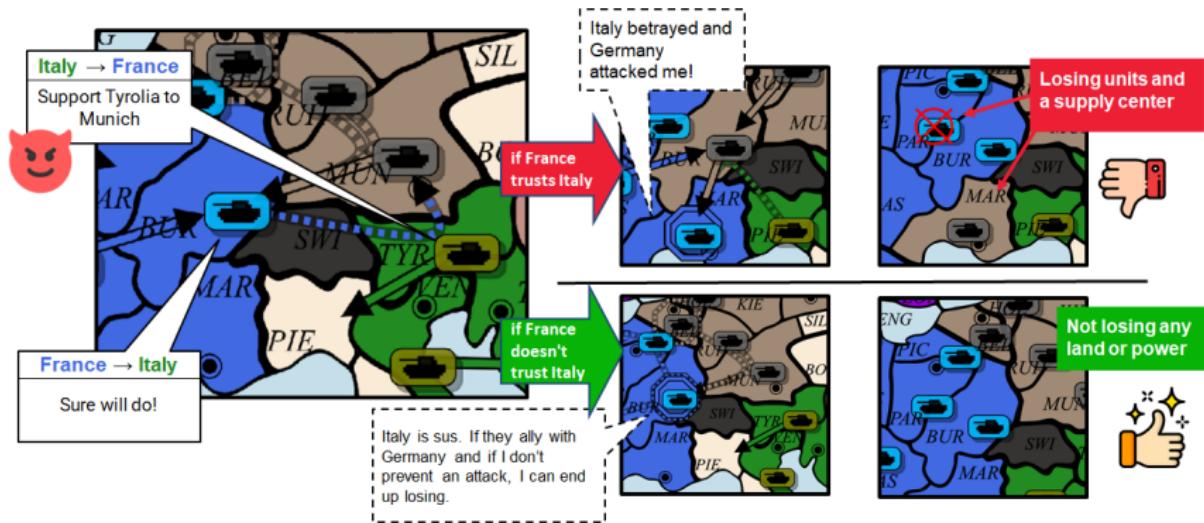
## How to better detect Deception



## How to better detect Deception



# How to better detect Deception



## It's Time for Some Game Theory

		proposal	actual
			
 proposal	proposal		
	default		

## It's Time for Some Game Theory

		proposal	actual
			
 proposal	proposal	<ul style="list-style-type: none"><li>• Italy moves to Munich</li><li>• France supports Italy to Munich</li></ul>	
	default	<ul style="list-style-type: none"><li>• Italy moves to Munich</li><li>• France defends in Burgundy</li></ul>	

## It's Time for Some Game Theory

		proposal	actual
			
 proposal	proposal	<ul style="list-style-type: none"><li>• Italy moves to Munich</li><li>• France supports Italy to Munich</li></ul>	<ul style="list-style-type: none"><li>• Italy moves to Piedmont</li><li>• France supports Italy to Munich</li></ul>
	default		

## It's Time for Some Game Theory

		proposal	actual
			
	proposal	<ul style="list-style-type: none"><li>• Italy moves to Piedmont</li><li>• France supports Italy to Munich</li></ul>	
	default		<ul style="list-style-type: none"><li>• Italy moves to Piedmont</li><li>• France defends in Burgundy</li></ul>

## It's Time for Some Game Theory

	proposal 	actual
	<ul style="list-style-type: none"><li>• Italy moves to Munich</li><li>• France supports Italy to Munich</li></ul>	<ul style="list-style-type: none"><li>• Italy moves to Piedmont</li><li>• France supports Italy to Munich</li></ul>
	<ul style="list-style-type: none"><li>• Italy moves to Munich</li><li>• France defends in Burgundy</li></ul>	<ul style="list-style-type: none"><li>• Italy moves to Piedmont</li><li>• France defends in Burgundy</li></ul>

Bait  $\rightarrow$   $U_1 = u_{\text{France}}(\text{prop}, \text{prop}) - u_{\text{France}}(\text{def}, \text{prop})$

Switch  $\rightarrow$   $U_2 = u_{\text{France}}(\text{prop}, \text{prop}) - u_{\text{France}}(\text{prop}, \text{act})$

Edge  $\rightarrow$   $U_3 = u_{\text{Italy}}(\text{prop}, \text{act}) - u_{\text{Italy}}(\text{def}, \text{act})$

## Really Hard Problem, but Better

Model	Precision	Recall	F1-Score
LLM baseline using Direct Judgment	0.095	<b>0.551</b>	0.161
LLM baseline using Alignment Judgment	0.147	0.065	0.090
<b>CTRL-D (ours)</b>	<b>0.950</b>	0.238	<b>0.380</b>
CTRL-D with human annotated logical forms	<b>0.960</b>	0.300	<b>0.457</b>
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

# **Algorithms that ...**

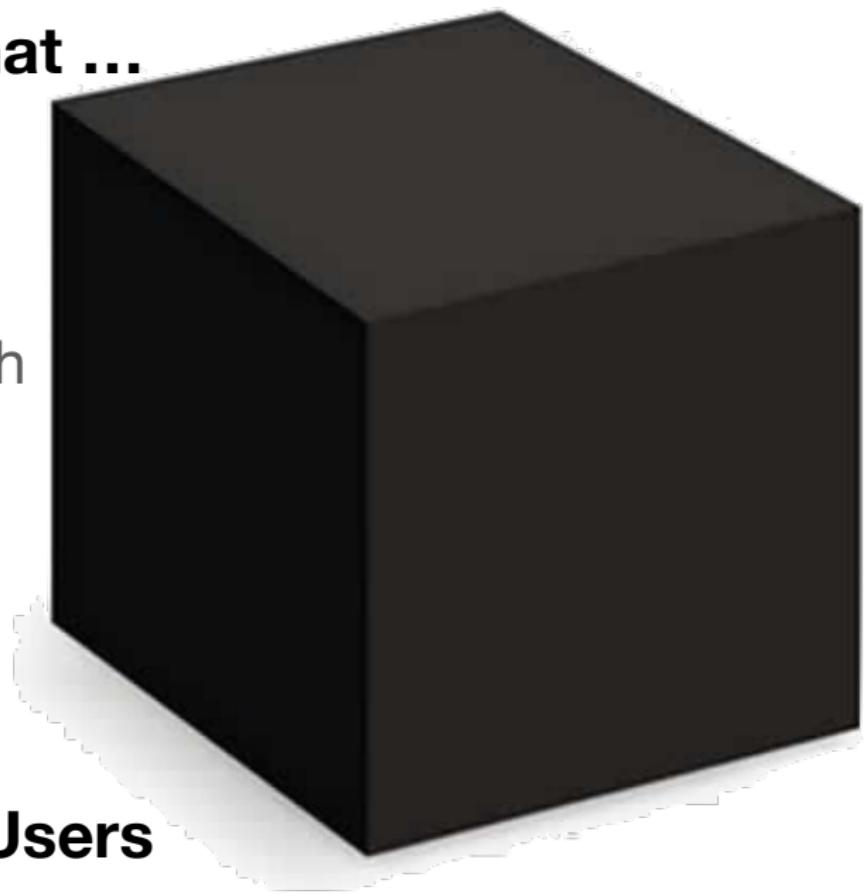
Inform

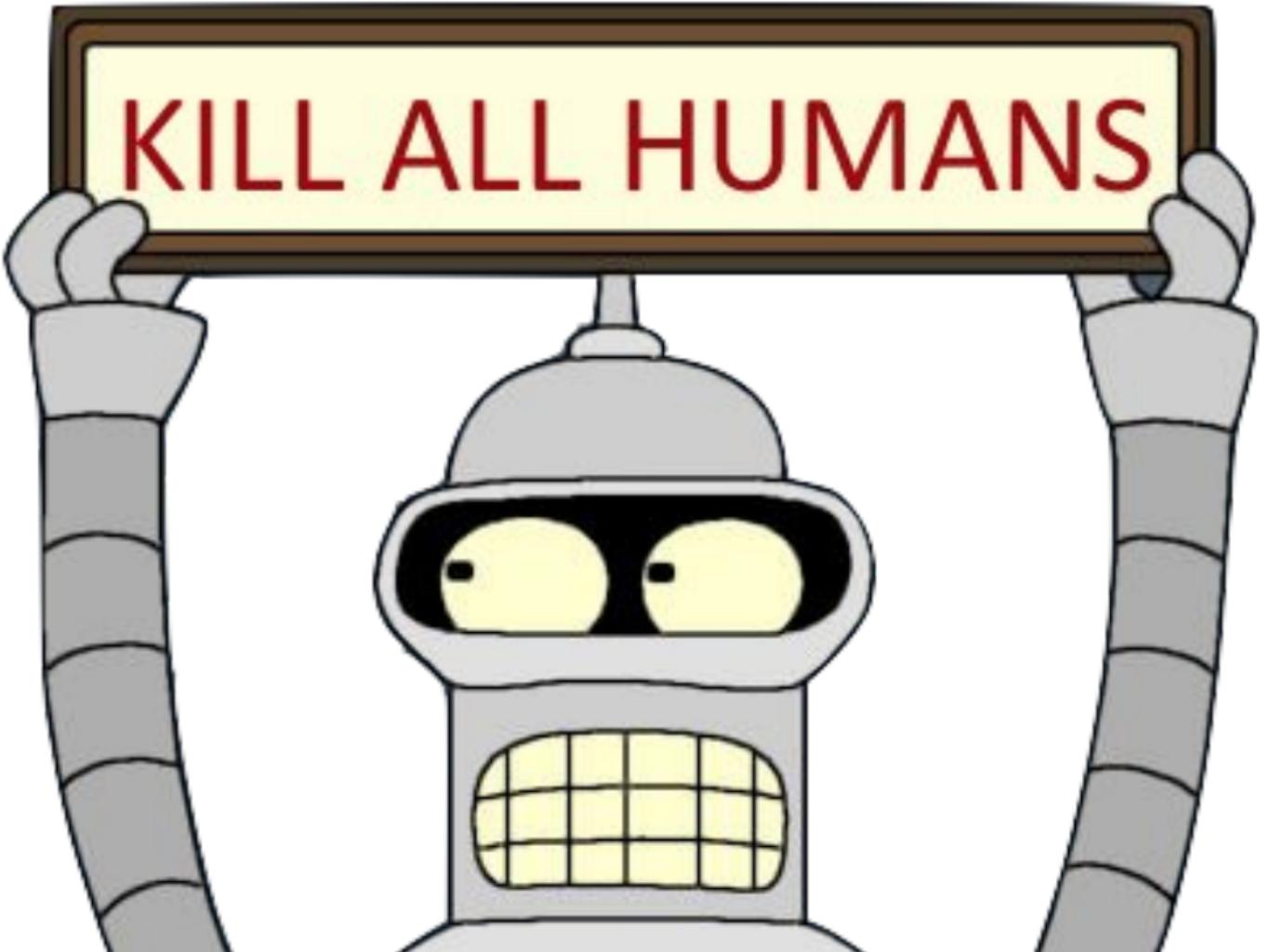
Collaborate with

Compete with

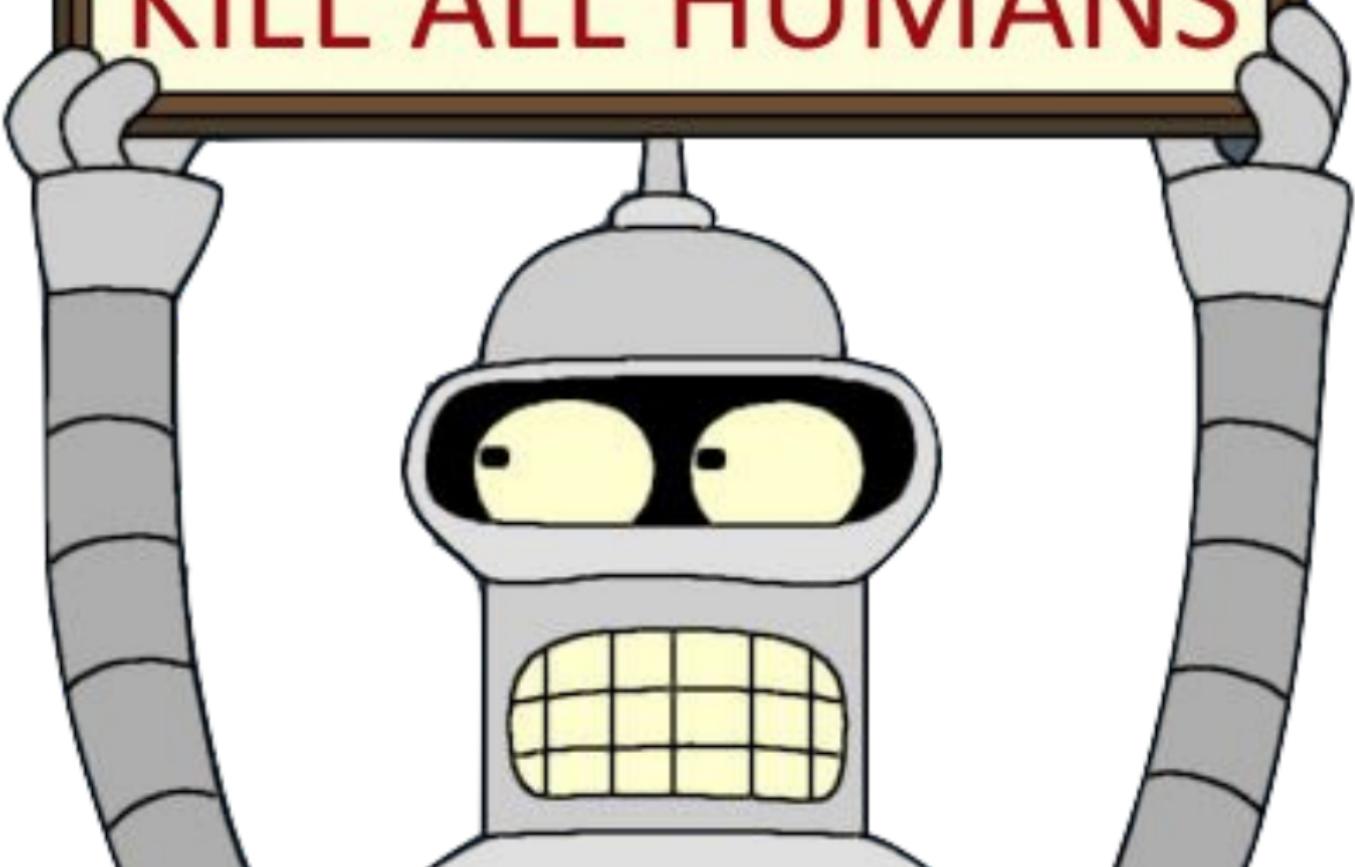
Understand

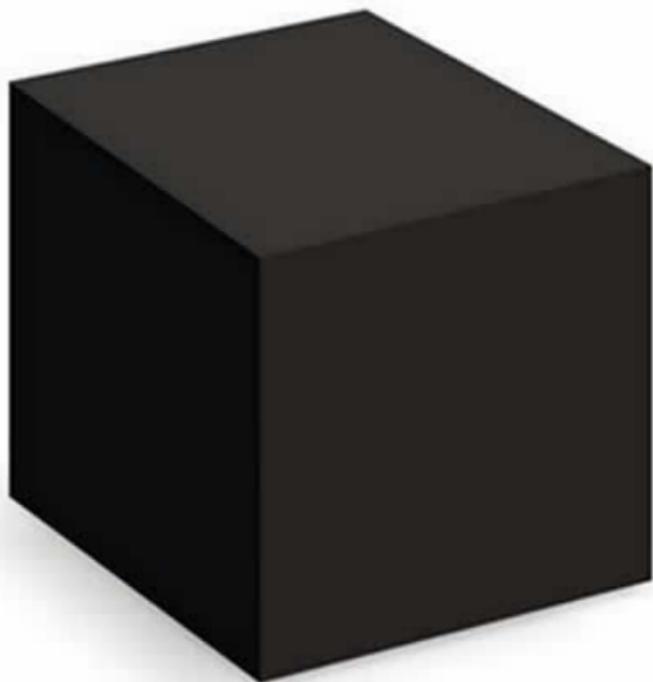
**their Human Users**





KILL ALL HUMANS







## Thanks

### Collaborators

Hal Daumé III (UMD), Jon May (USC), Cristian (Columbia), Marine Carpuat (UMD), Eve Fleisig (Berkeley), Sherry Wu (CMU)

### Funders



### Supporters

International Academic  
**Competitions**

**NAQT**<sup>SM</sup>

National Academic Quiz Tournaments, LLC

## Moving beyond games

- 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

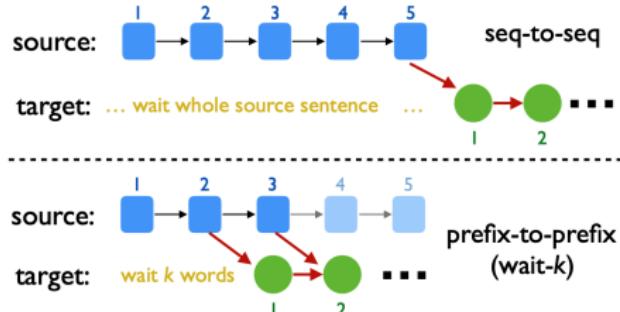
Assistant Professor (on-leave), Oregon State University



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



Bushí	zǒngtǒng	zài	Mósikē	yǔ	Éluóst	zǒngtǒng	Pǔjīng	huiwù
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

Buzz

3

### [room\_1] Round 1 Question 7/24

#### Source :

Jest to kraina historyczna Azji, która obecnie znajduje się w większości w granicach Chin. Kraina ta jest położona na średniej wysokości około czterech do pięciu tysięcy metrów nad poziomem morza i na granicy m.in z Himalajami. Aby zdobyć punkt,

#### Target :

It is a historical land of Asia, which is now mostly located within China. This land is located on the average of about four to five thousand meters above sea level and on the border with Himalaya.

- 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

Buzz 3

### [room\_1] Round 1 Question 7/24

#### Source :

Jest to kraina historyczna Azji, która obecnie znajduje się w większości w granicach Chin. Kraina ta jest położona na średniej wysokości około czterech do pięciu tysięcy metrów nad poziomem morza i na granicy m.in z Himalajami. Aby zdobyć punkt,

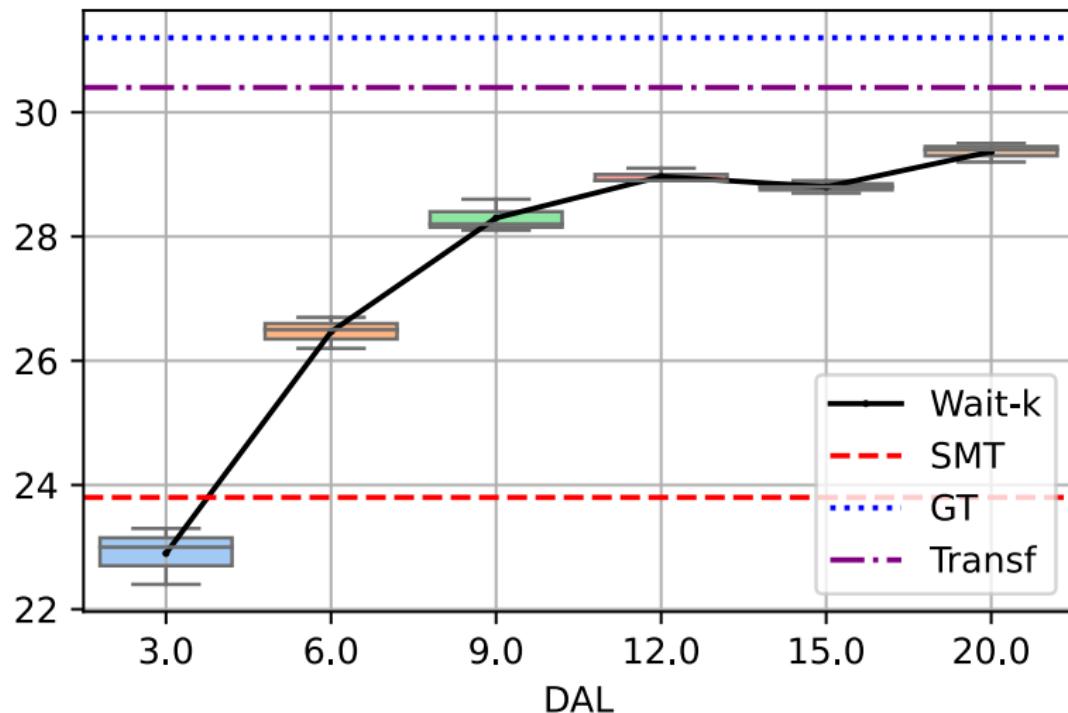
#### Target :

It is a historical land of Asia, which is now mostly located within China. This land is located on the average of about four to five thousand meters above sea level and on the border with Himalayas.

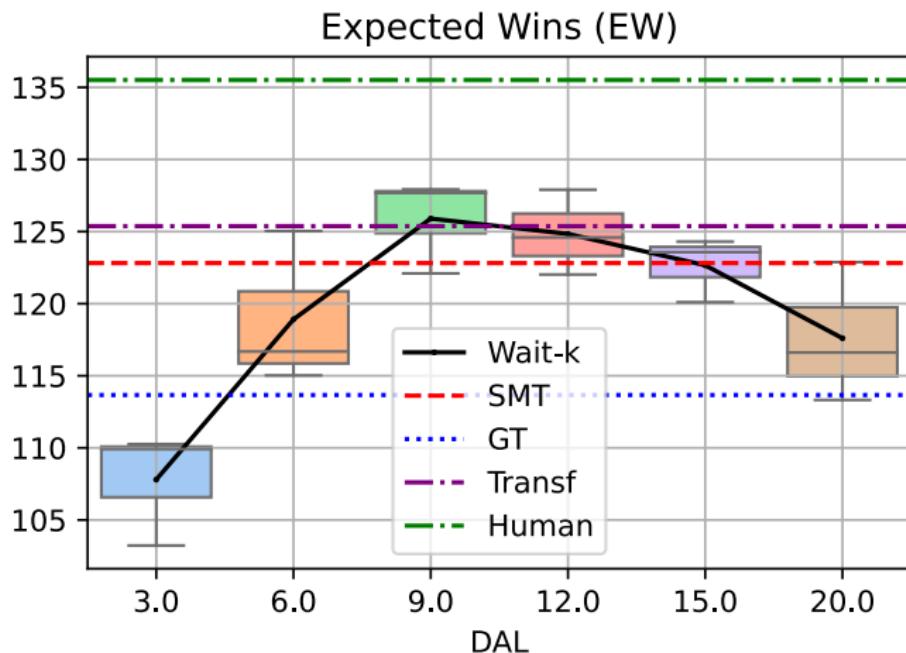
- You're a contestant on a Polish game show
- You have access to a simultaneous translation system
- Your job is to answer the question before your opponent (as quickly as possible)
- Keep question answerer the same, vary translation

## BLEU results for modern Simultaneous Translation Systems

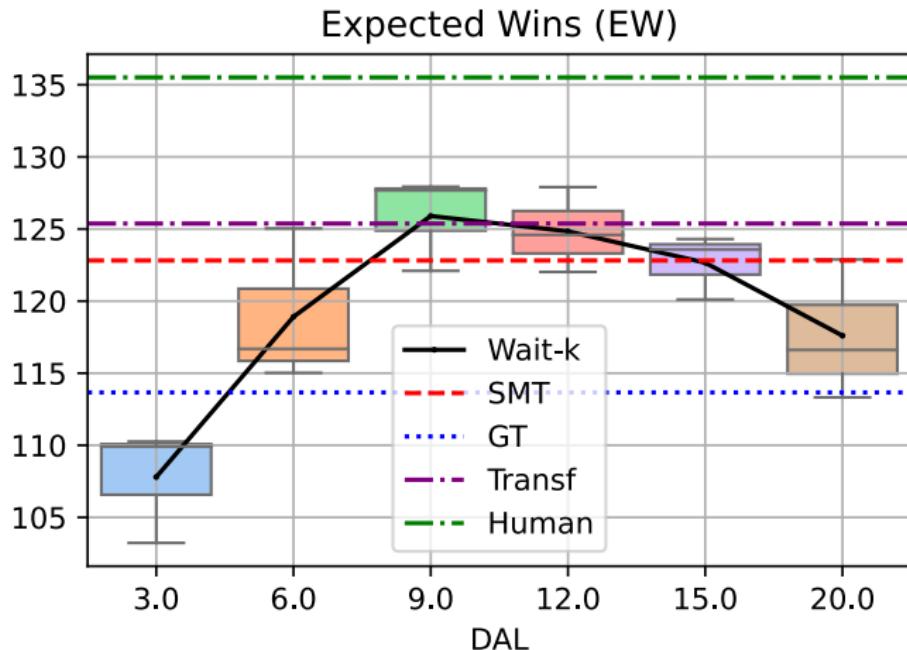
BLEU on WMT-pl



## Downstream QA Results

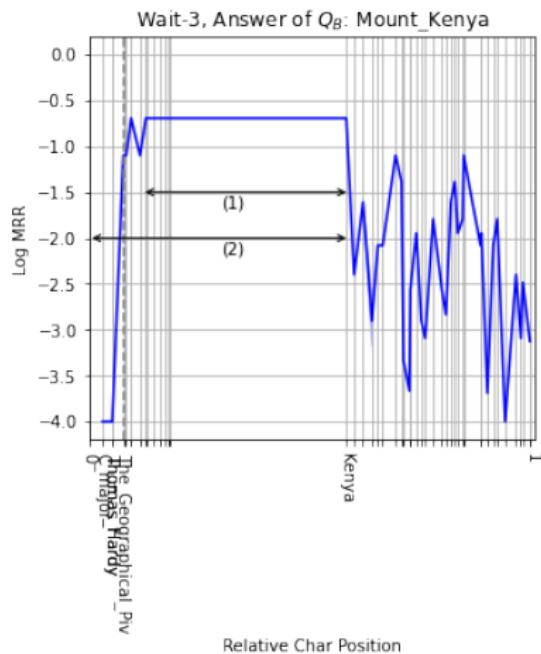


## Downstream QA Results



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 ...

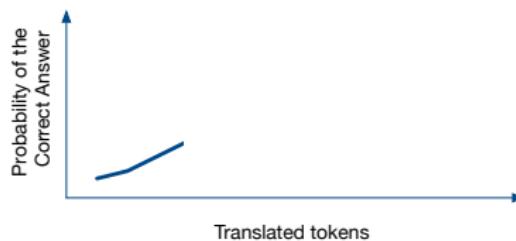


**Guess:** ??

## When are Mistakes / Hallucinations Harmful?

This coordinate determines

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

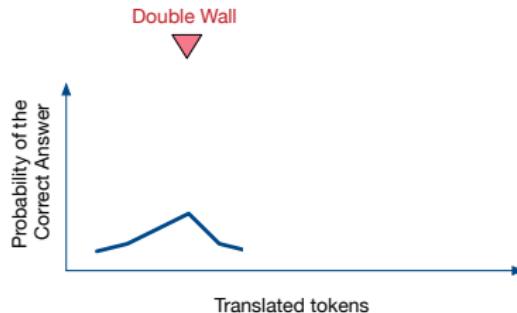


**Guess:** IP Address

## 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 ...

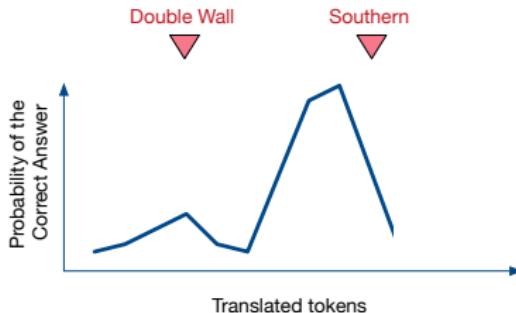


**Guess:** Longitude

## 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 ...

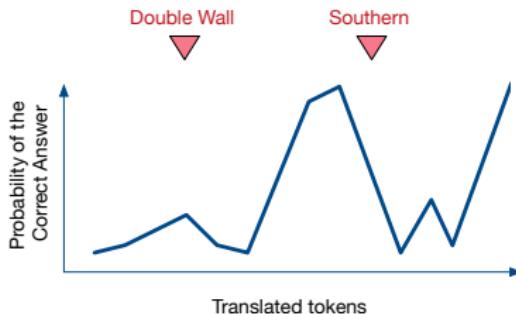


**Guess:** Spherical Coordinate

## 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

But wait, there's more!

## Computational Social Science



## Detecting Deception



[?, ?]

## Multilingual/Multicultural Models



[?, ?]

## Computational Biology

RKQEDNHWYFMLIVCTSAGP  
----- FmL . -----  
----- fmL+ .-.--  
-----. fmliv+ . . .  
....-.. +mL+ .. -.

[?, ?]



## 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

To provide for payments to certain natural resource trustees to assist in re...

- evacuation
- safety
- shelter

**flood** coast marine  
restoration coastal vessel  
fish gulf wildlife species  
pollution council great  
fishery fishing waters  
ecosystem monitoring  
fisheries mitigation

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 part...

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...

**remain**  
expended heading  
disaster september  
appropriation transferred  
obligation division unit  
capital acquisition  
inspector purchase funded  
procurement units corps  
repair salaries

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...

Covered Themes Progress:

## Interactive Document Labeling

Jordan

Time left: 38:42

To provide for payments to certain natural resource trustees to assist in re...

- evacuation
- safety
- shelter

new label name

add label

rename label

delete label

**flood** coast marine  
restoration coastal vessel  
fish gulf wildlife species  
pollution council great  
fishery fishing waters  
ecosystem monitoring  
fisheries mitigation

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...

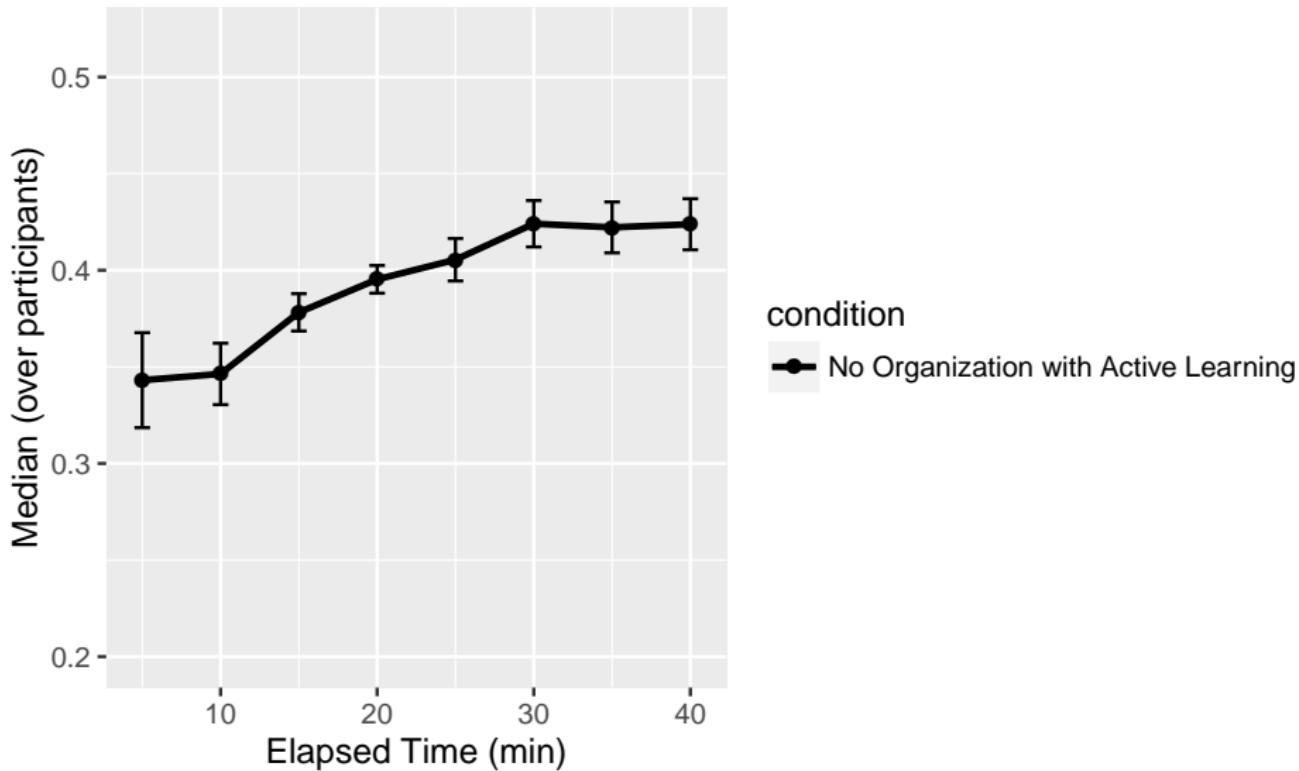
To rescind any unobligated discretionary appropriations returned to the F...

To amend the Robert T. Stafford Disaster Relief and Emergency Assistanc...

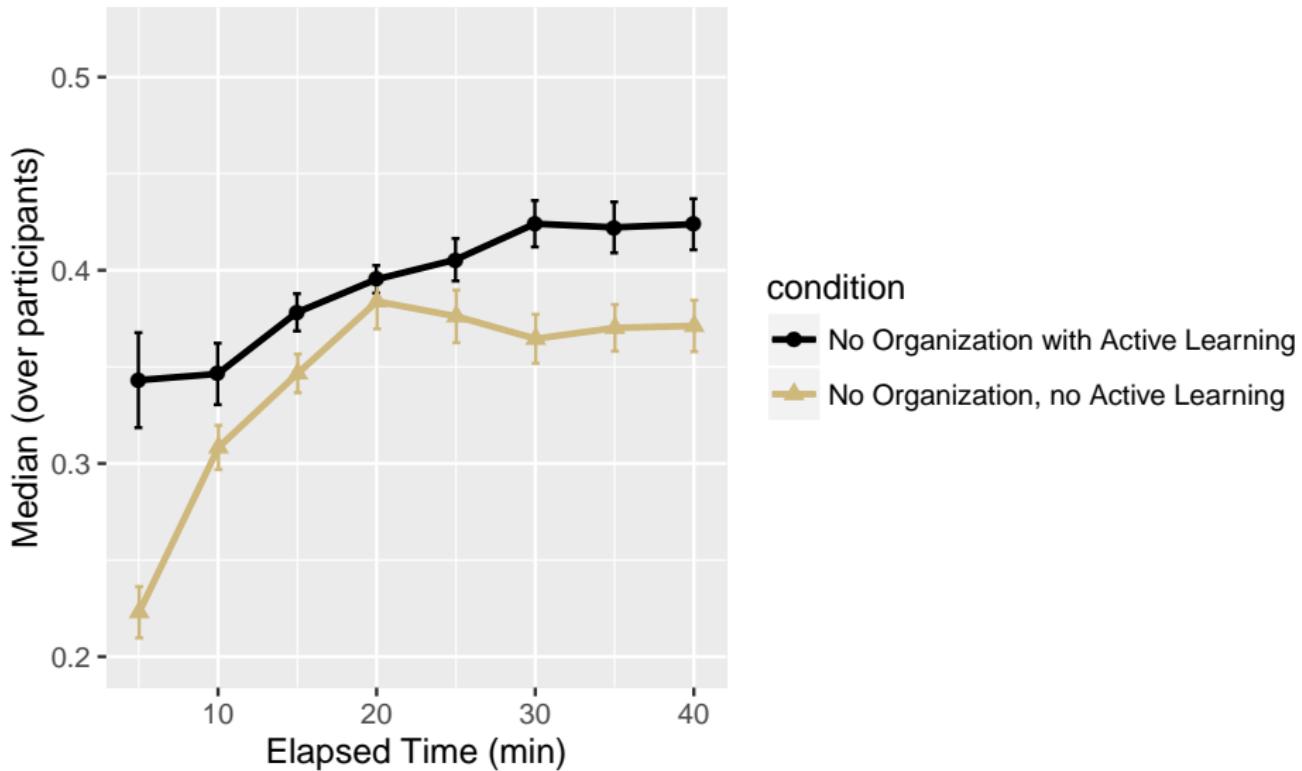
Making appropriations for energy and water development and related age...

Covered Themes Progress:

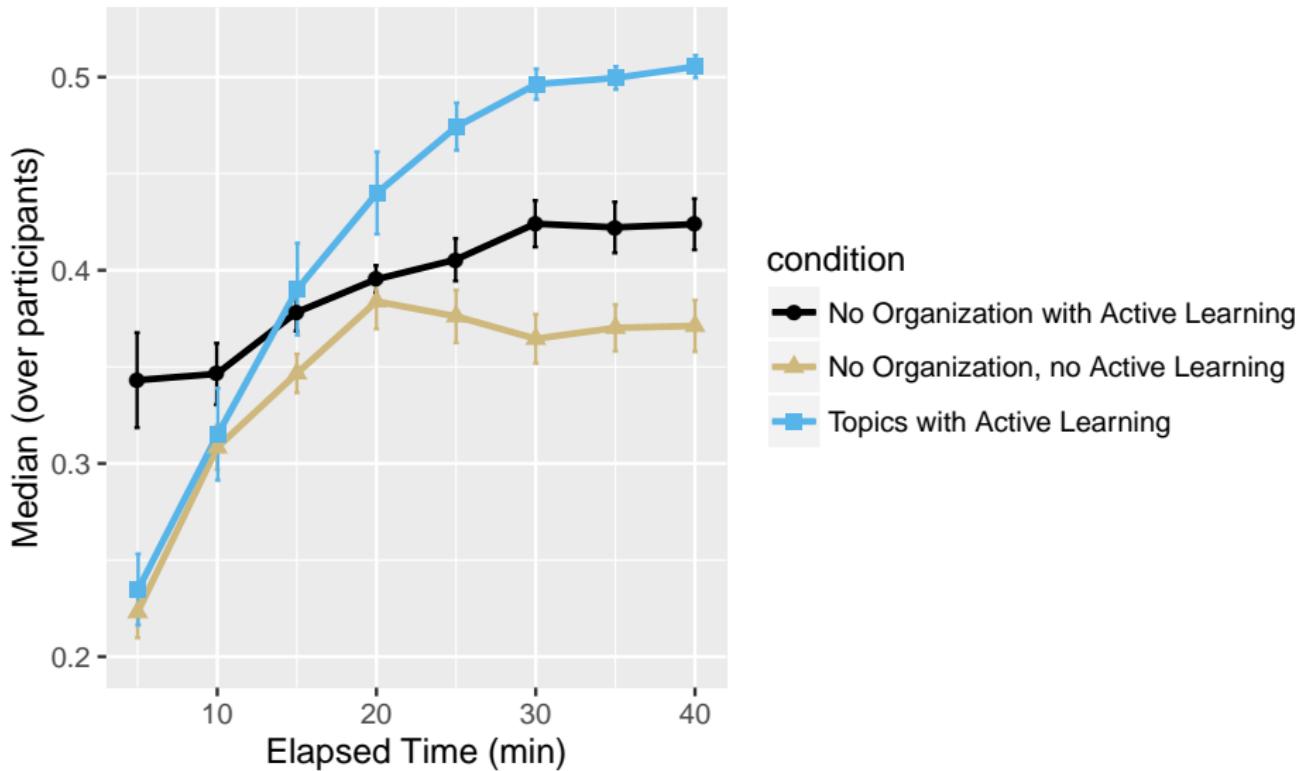
Direct users to document



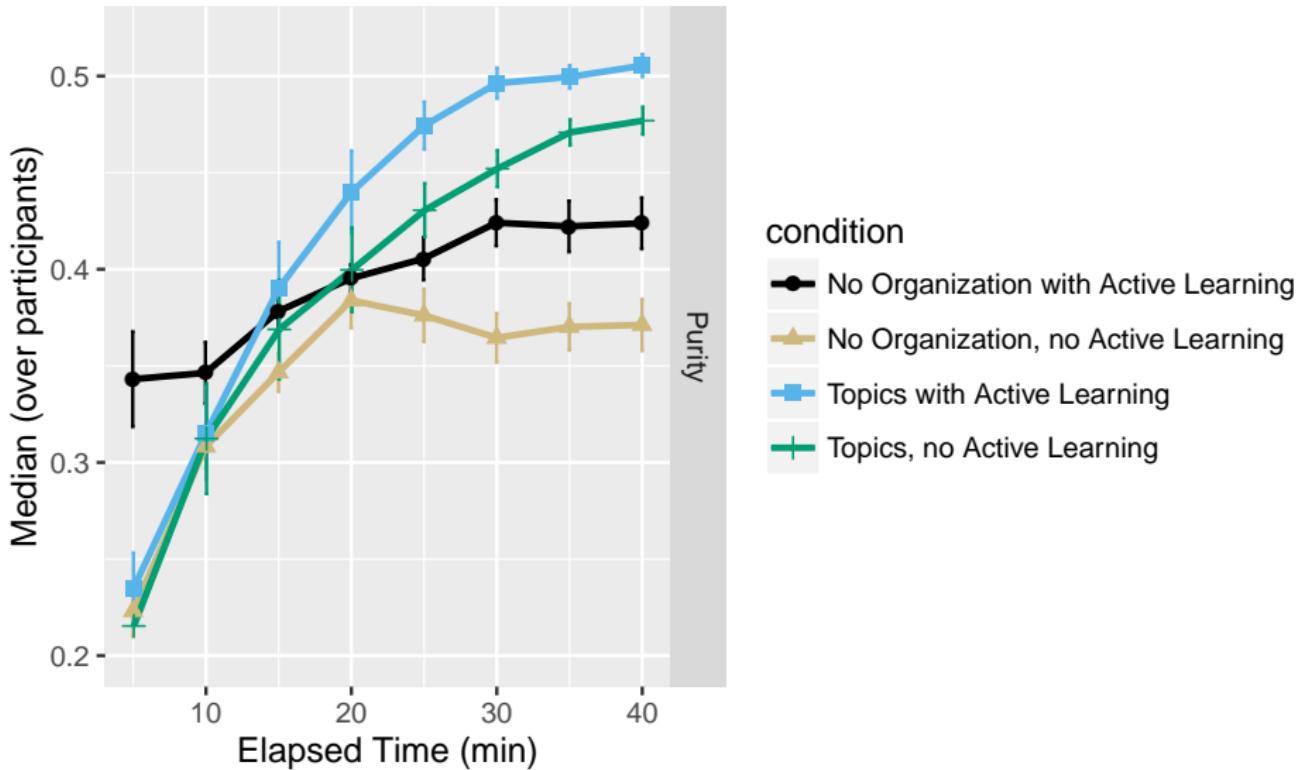
Active learning if time is short



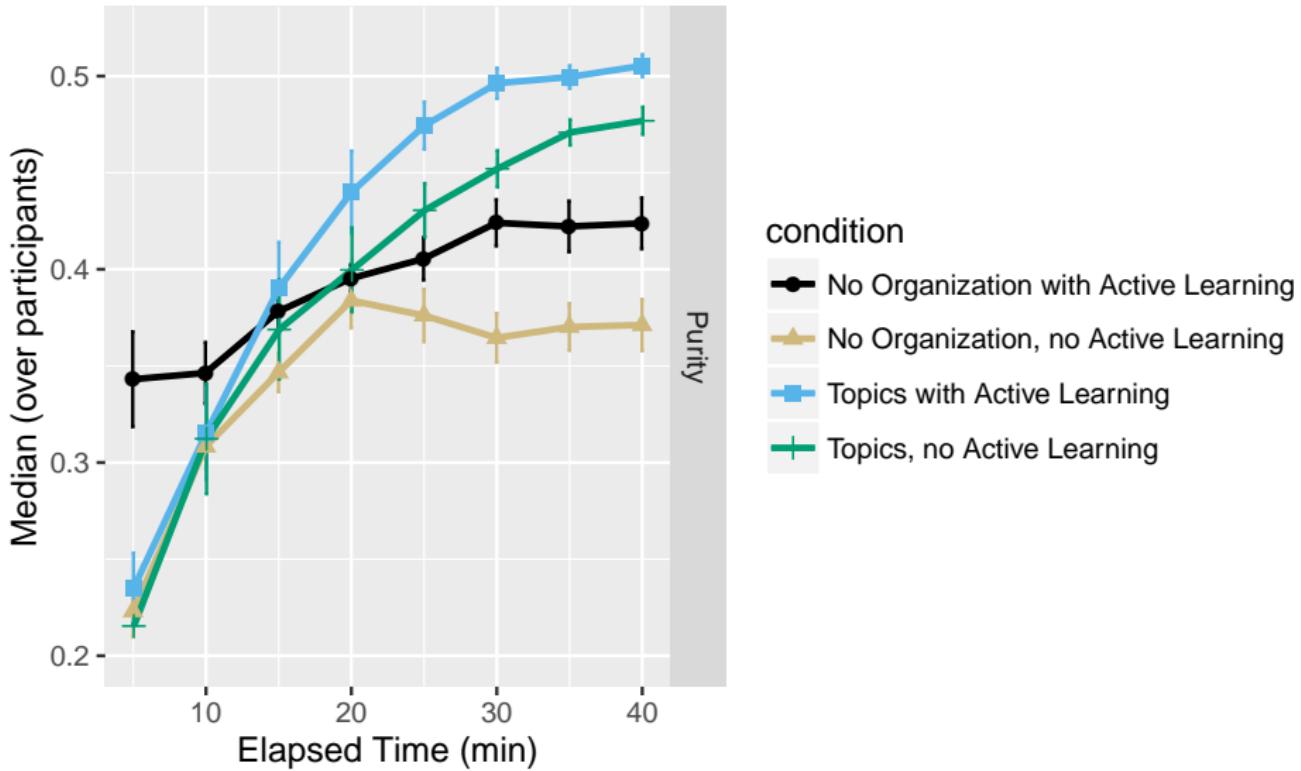
Better than status quo



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

Kenneth\_Arrow  
John\_Dewey  
Erving\_Goffman  
Bertrand\_Russell  
Immanuel\_Kant  
Adam\_Smith  
David\_Hume  
Samuel\_Johnson

Andrea Lin

# STACL: Simultaneous Translation with Integrated Anticipation & Controllable Latency



**Liang Huang**  
Principal Scientist, Baidu Research

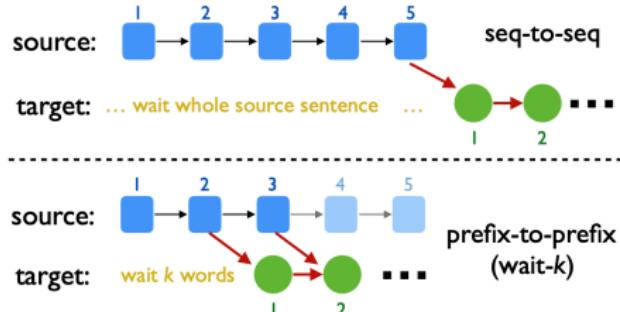
Assistant Professor (on-leave), Oregon State University



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



Bushí 布什	zöngtōng 总统	zài 在	Mósikē 莫斯科	yǔ 与	Éluóst 俄罗斯	zöngtōng 总统	Pǔjīng 普京	huiwù 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

Buzz

3

### [room\_1] Round 1 Question 7/24

#### Source :

Jest to kraina historyczna Azji, która obecnie znajduje się w większości w granicach Chin. Kraina ta jest położona na średniej wysokości około czterech do pięciu tysięcy metrów nad poziomem morza i na granicy m.in z Himalajami. Aby zdobyć punkt,

#### Target :

It is a historical land of Asia, which is now mostly located within China. This land is located on the average of about four to five thousand meters above sea level and on the border with Himalaya.

- 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

Buzz

3

### [room\_1] Round 1 Question 7/24

#### Source :

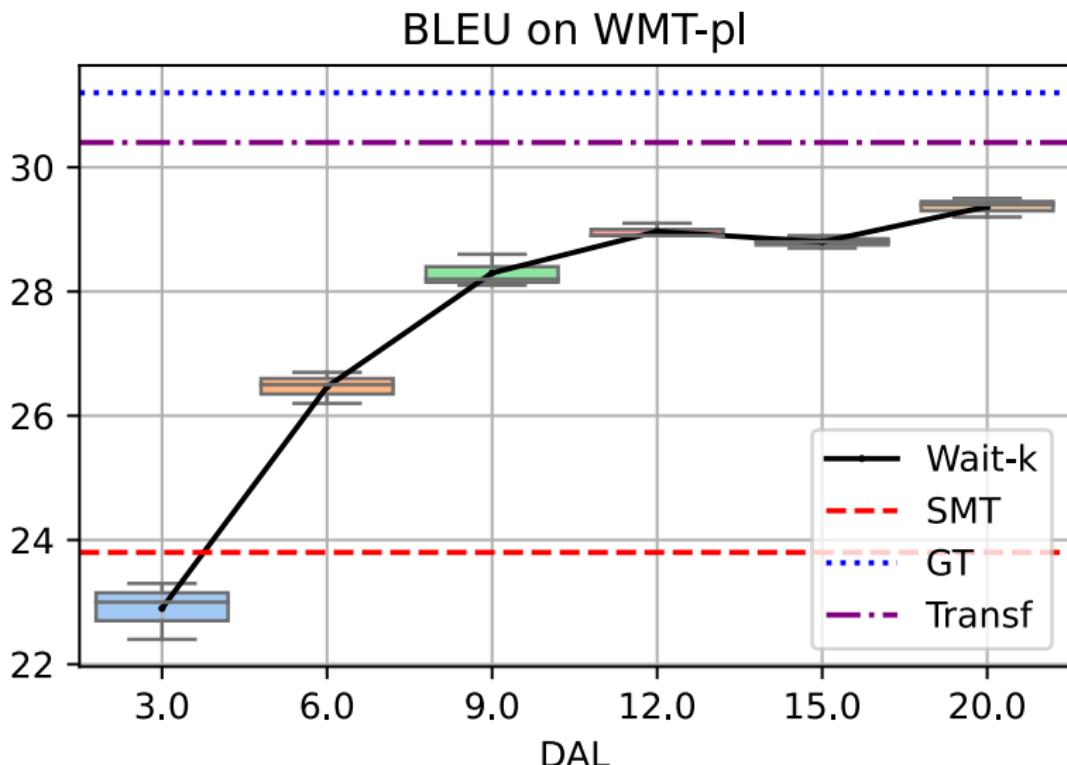
Jest to kraina historyczna Azji, która obecnie znajduje się w większości w granicach Chin. Kraina ta jest położona na średniej wysokości około czterech do pięciu tysięcy metrów nad poziomem morza i na granicy m.in z Himalajami. Aby zdobyć punkt,

#### Target :

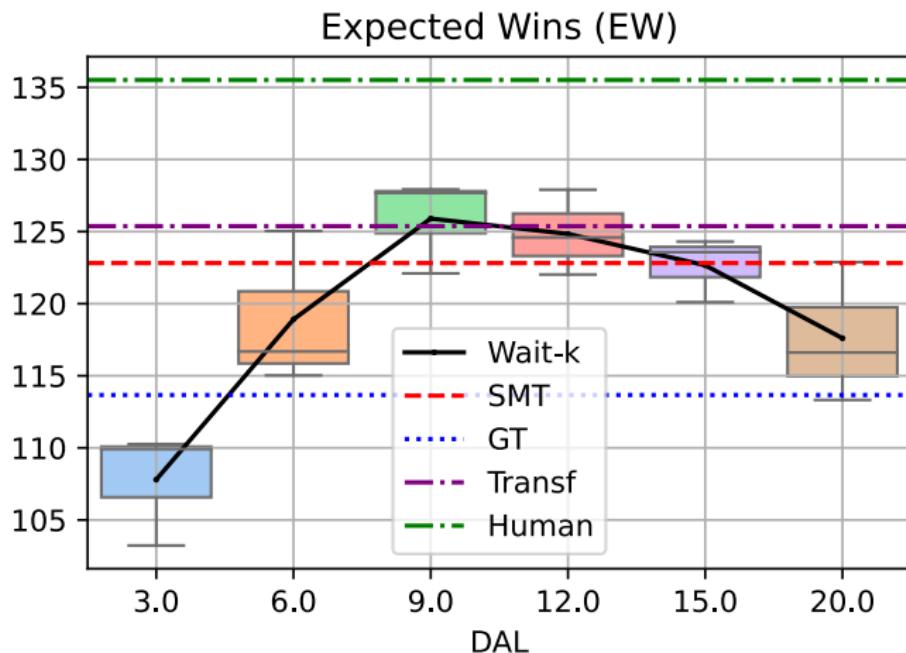
It is a historical land of Asia, which is now mostly located within China. This land is located on the average of about four to five thousand meters above sea level and on the border with Himalayas.

- 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

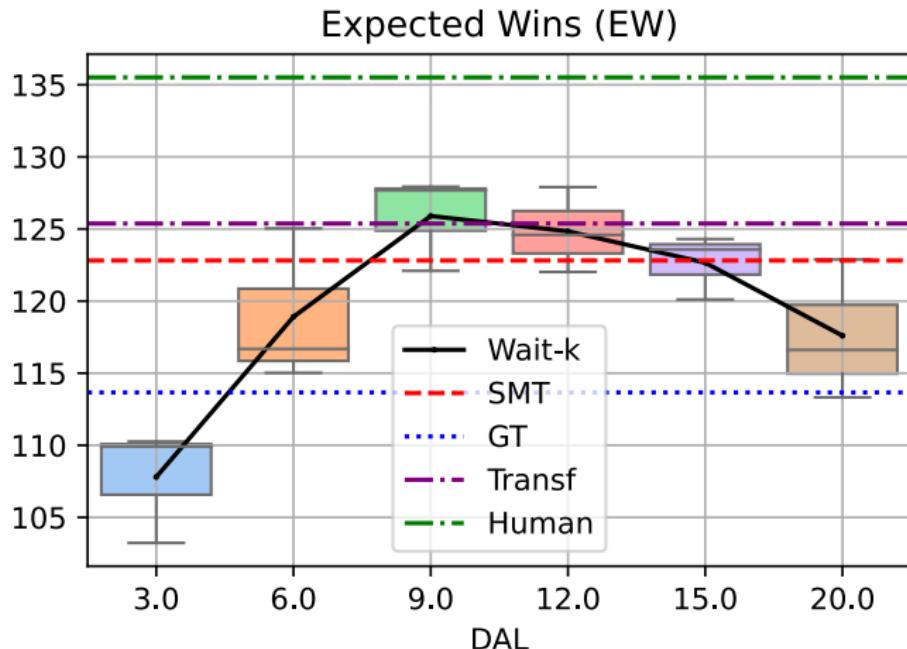
## BLEU results for modern Simultaneous Translation Systems



## Downstream QA Results

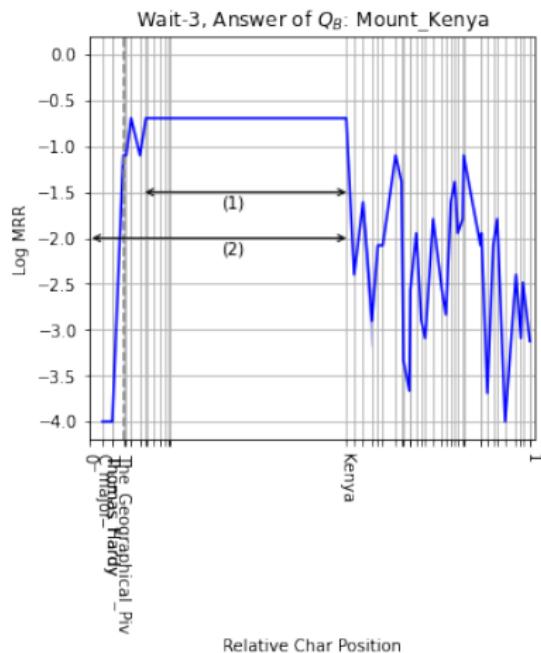


## Downstream QA Results



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 ...

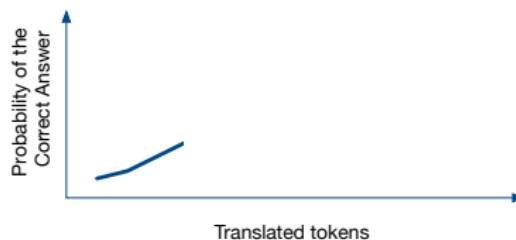


**Guess:** ??

## When are Mistakes / Hallucinations Harmful?

This coordinate determines

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

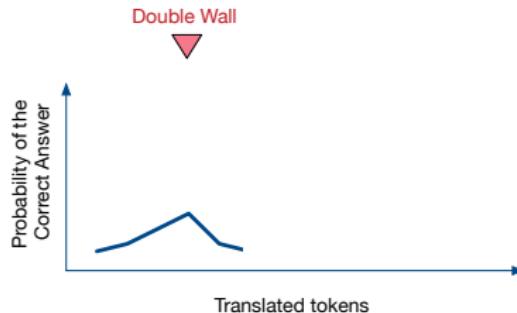


**Guess:** IP Address

## 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 ...

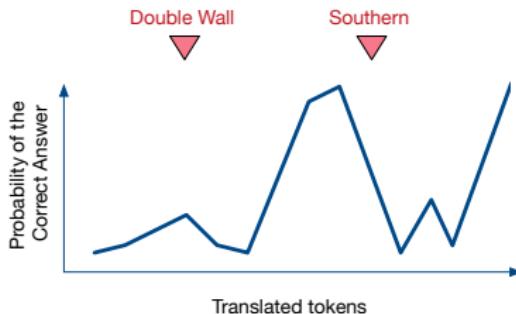


**Guess:** Spherical Coordinate

## 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

## References

## 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

## 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?
  - Improve Calibration
  - Improve Cooperation

## Human–Computer 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
- Big idea: what if we give the humans tools to better understand computer thought process





## 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**

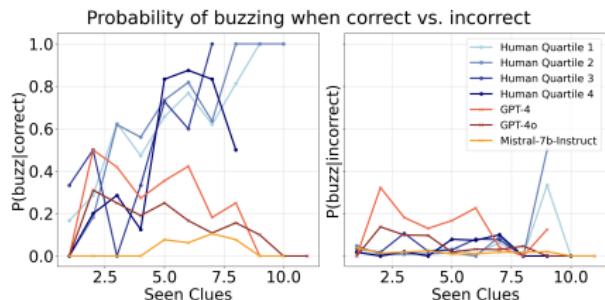
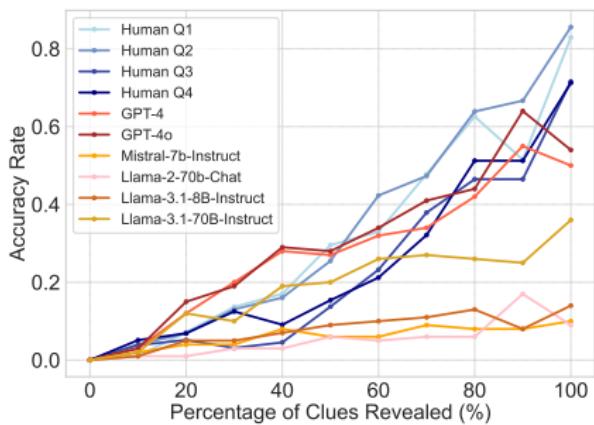
1. University of Chicago
2. Colorado School of Mines
3. Cornell University
4. UIUC
5. Brigham Young University
6. California Institute of Technology
7. Peking University
8. Harvey Mudd College
9. Darmstadt University
10. University of Colorado

## How does a computer buzz in?

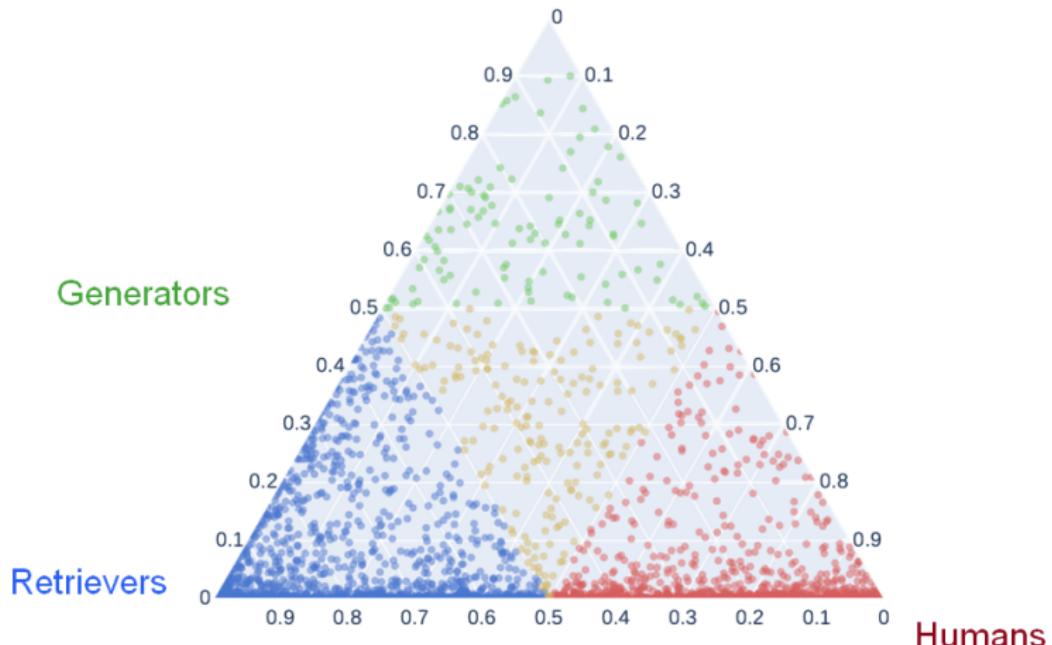
- Ask computer for guess every  $N$  words
- Compute average log probability
  - Threshold on log probability of output
  - Threshold chosen from TrickMe set
  - Can do better!
- Again, only have one chance to buzz

## Human–Computer Competition

- Top human team won
- Best computer team had *much* higher accuracy
- Computers had strictly higher accuracy
- Humans had *much* higher **conditional** accuracy



## What's hard for Computers



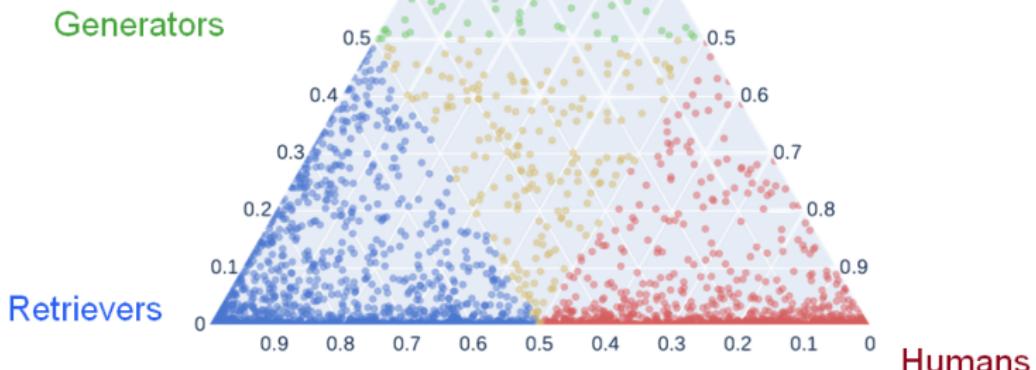
## What's hard for Computers

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**

Generators

Retrievers

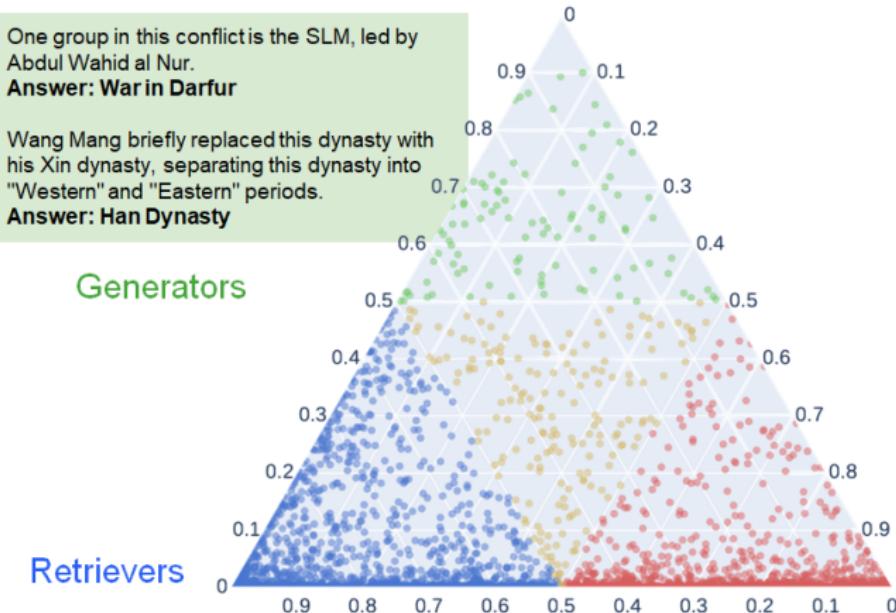
Humans

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**



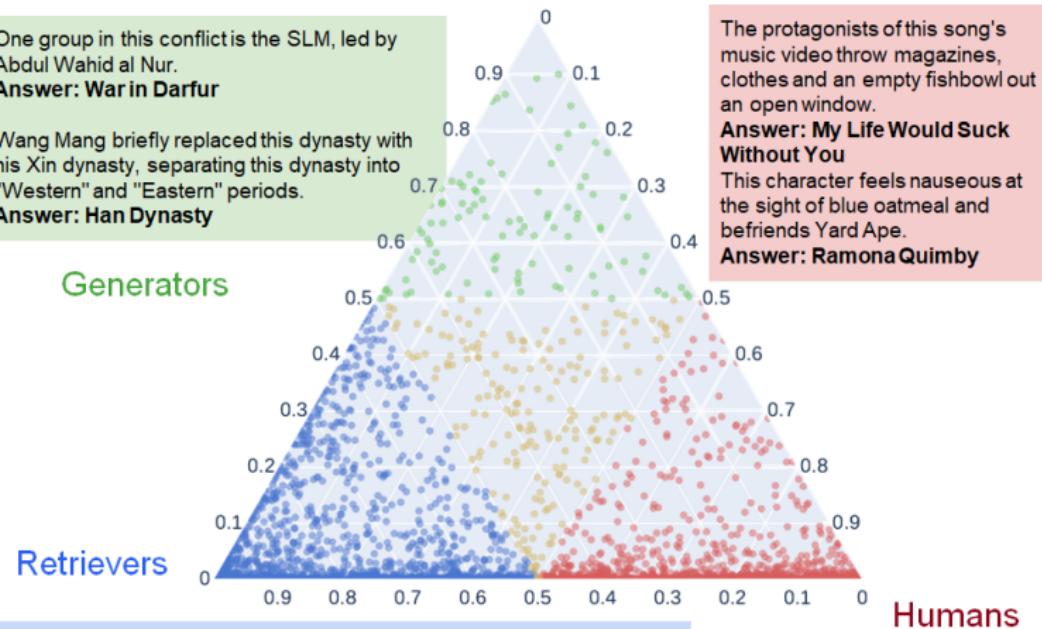
## What's hard for Computers

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**



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**

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**

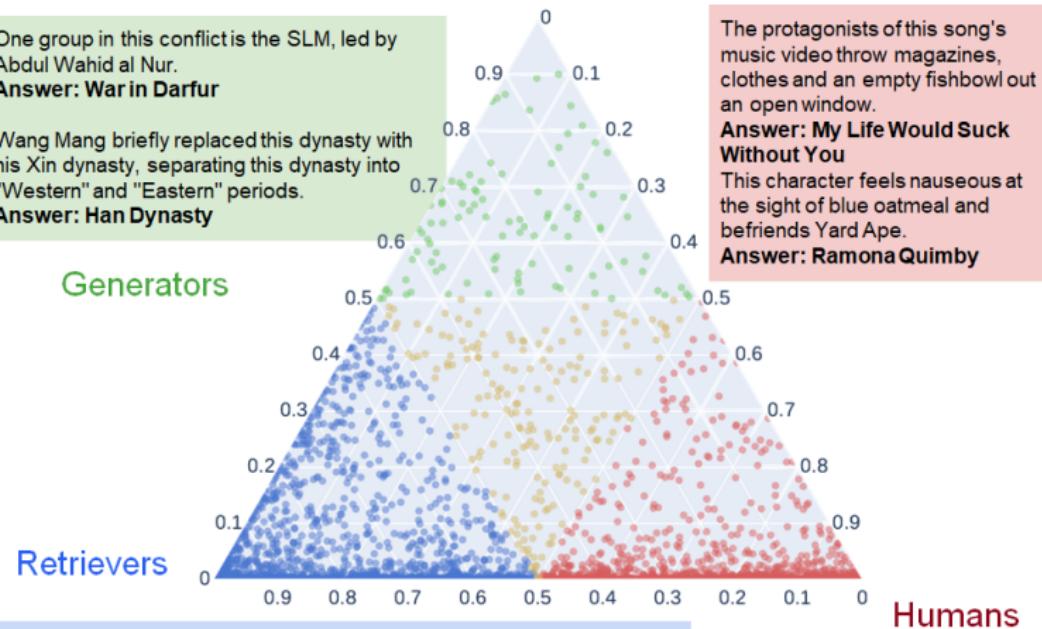
# What's hard for Computers

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**

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**

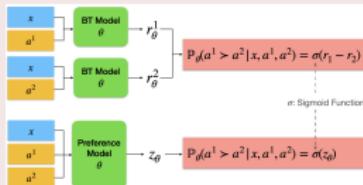
## Aside: Family Tree

### Zermelo

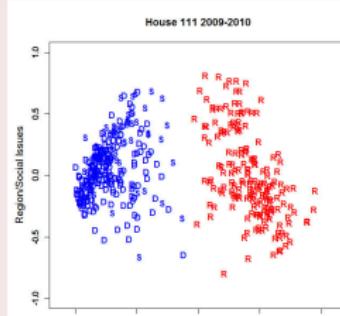


A black and white portrait of Ernst Zermelo, a man with a beard and glasses, resting his chin on his hand.

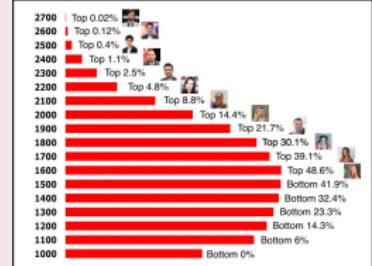
## Bradley-Terry Models



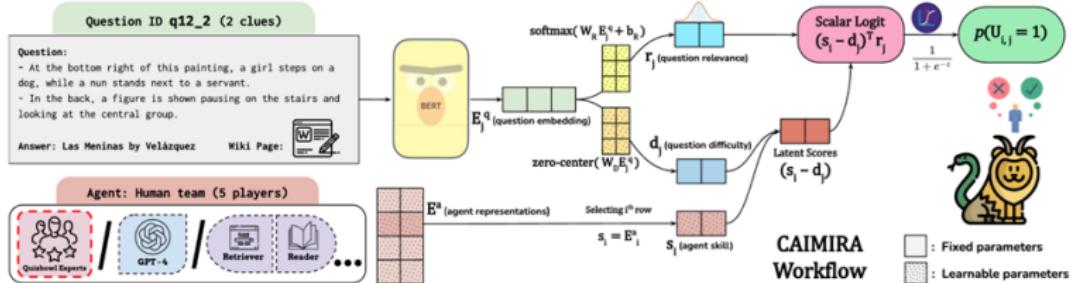
## Ideal Point Models



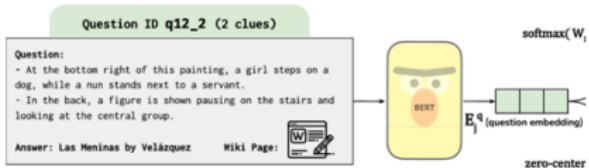
## ELO Models



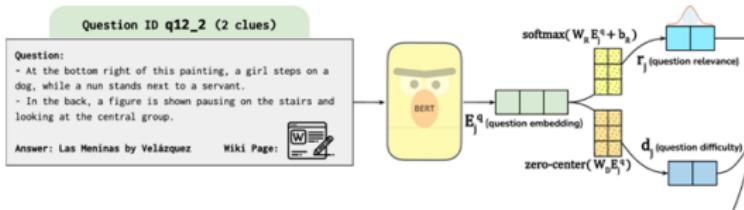
# What more do we need?



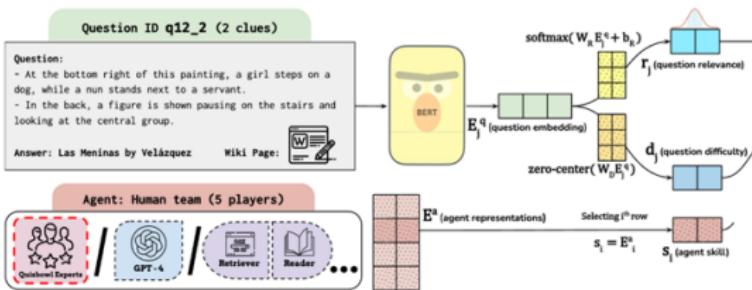
# What more do we need?



## What more do we need?



## What more do we need?



# What more do we need?

