



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

Jordan Boyd-Graber
University of Maryland
2025

Why Language Games?



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

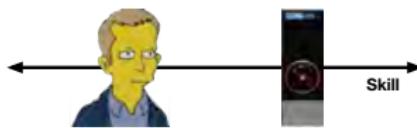


Image: DeepMind

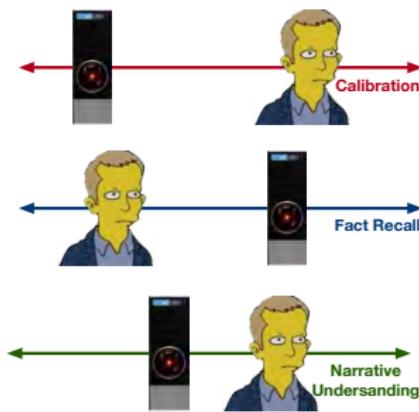


Peter Morgan/Reuters

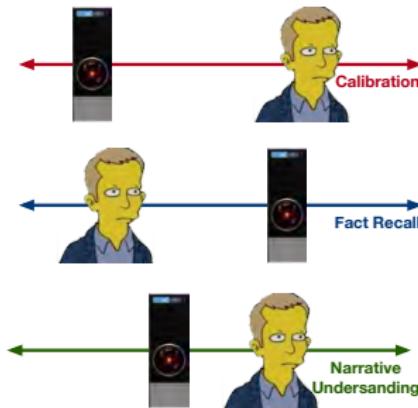
Measuring Skill



Measuring Skill



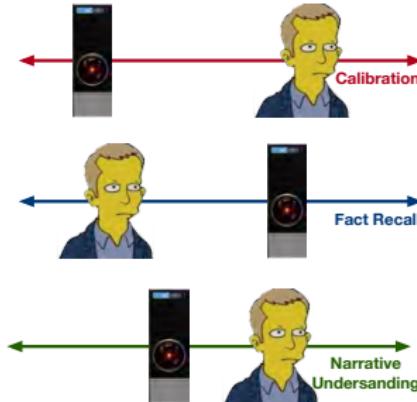
Measuring Skill



Augmentation

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

Measuring Skill



Augmentation

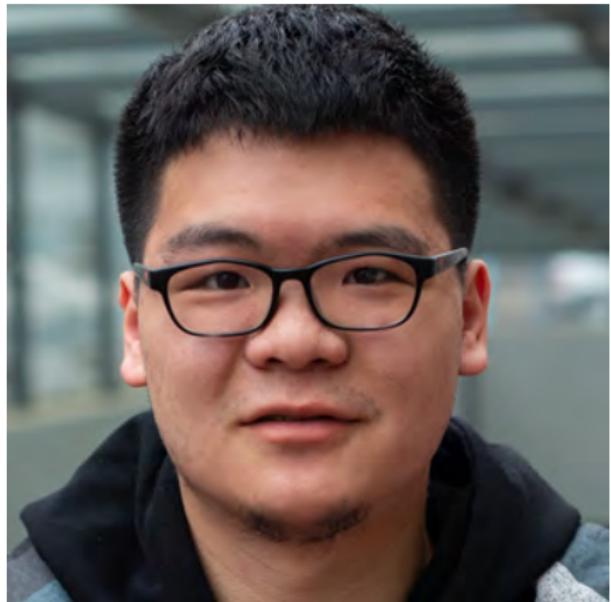
How much can we *increase* the skill of a human in a language game with computational (i.e., AI) support?

1. Fact Checking
2. Question Answering
3. Negotiation

Fact Checking with Humans in the Loop

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

Chenglei Si¹ Navita Goyal² Sherry Tongshuang Wu³
Chen Zhao⁴ Shi Feng⁵ Hal Daumé III^{2,6} Jordan Boyd-Graber²
¹Stanford University ²University of Maryland ³Carnegie Mellon University
⁴NYU Shanghai ⁵New York University ⁶Microsoft Research
clsi@stanford.edu



Can we help users detect incorrect statements?

Claim

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

Submit

✓ Submit and save

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)

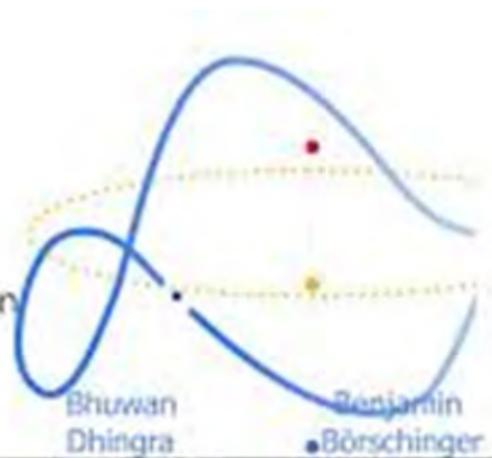
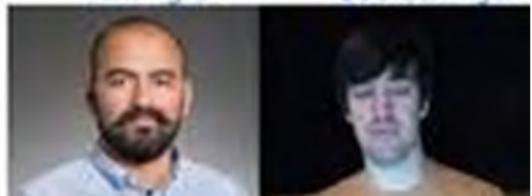
Fool Me Twice

Google Research

Fool Me Twice Entailment from Wikipedia Gamification

Andreas Stuhlmüller, Michael Ullman, Joshua Berman
Massachusetts Institute of Technology
[arXiv:1306.5100 \[cs.LG\]](http://arxiv.org/abs/1306.5100)

Jannis
Büfian Jordan
Boyd-Graber Julian
Eisenschlos



Fool Me Twice

Google Research

Fool Me Twice

We intended this to be “adversarial”

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

Jannis
Büfian Jordan
Boyd-Graber Julian
Eisenschlos



Bhuwan
Dhingra



Benjamin
Börschinger



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

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

< Michael Faraday 7:52 Table of Contents

Evidence (0 marked as gold)

Michael Faraday (22 September 1791 – 23 August 1867) was an English scientist who contributed to the study of electromagnetism and electrochemistry.

Gold evidence

His main discoveries include the principles underlying electromagnetic induction, diamagnetism and electrolysis.

Gold evidence

Although Faraday received little formal education, he was one of the most influential scientists in history.

Summary

Michael Faraday (22 September 1791 – 23 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 web browser window for the game "Fool Me Twice". The title bar says "Fool Me Twice googleplex.com/twice". The main content area displays a challenge about Michael Faraday. A yellow box highlights the sentence: "Faraday died in an explosion comparing nitrogen trichloride samples". Below this, a text box contains: "Very soon Davy entrusted Faraday with the preparation of nitrogen trichloride samples, and they both were injured in an explosion of that very sensitive substance." Two evidence buttons are shown: "Gold evidence" and "Gold evidence". The bottom text box states: "Faraday died at his house at Hampton Court on 25 August 1867, aged 75." The right side of the screen shows the user's stats: "LEVEL 3 - DEBUT AUTHOR | 10068 POINTS".

Michael Faraday 5:00

Faraday died in an explosion comparing nitrogen trichloride samples

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

Gold evidence

Gold evidence

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

from Dumgill 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 Dince, 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 these 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 Barnesbury 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.

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

The screenshot shows a web browser window for the game "Fool Me Twice". The title bar says "Fool Me Twice googleplex.com/twice". The main content area displays a player profile for "Michael Faraday" with a score of 2:41. A statement is shown: "Faraday succumbed to his injuries in an explosion preparing nitrogen trichloride samples." Below it is a button labeled "SAVE FALSE STATEMENT". A section titled "Evidence (2 marked as gold)" lists two items: "Gold evidence ①" (This is now termed the Faraday effect) and "Gold evidence ②" (Faraday married Sarah Barnard (1800-1879) on 12 June 1821). A note at the bottom says "Don't see the gold evidence you're looking for? You can add it by clicking or the sentence in the Wikipedia page right!". To the right, under "Adult life", is a detailed paragraph about Faraday's injury and his subsequent employment at the Royal Institution. Under "Later life", there is a brief note about his marriage.

Fool Me Twice

LEVEL 3 - DEBUT AUTHOR | 10968 POINTS

< Michael Faraday 2:41

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 Doice, 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. Contemporaneously 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.

| 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 in 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

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

Evidence (2 marked as gold)

Gold evidence ①

This is now termed the **Faraday effect**.

Gold evidence ②

Faraday married Sarah Barnard (1800-1879) on 12 June 1821

Gold evidence ③

Don't see the gold evidence you're looking for? You can add it by clicking or the sentence in the Wikipedia page right!

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

The screenshot shows a web browser window for the game 'Fool Me Twice'. The title bar reads 'Fool Me Twice - fool-me-twice.googleplex.com'. The main content area displays a profile for Michael Faraday, a British scientist. The profile includes a summary, a 'Summary' section, and several evidence entries. The evidence entries are highlighted in yellow and contain claims that do not align with historical facts.

Michael Faraday 0:00

Faraday succumbed to his injuries in an explosion preparing nitrogen tri-chloride samples

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.

Evidence (1 marked as gold)

Very soon Davy entrusted Faraday with the preparation of nitro-glycerine samples, and they BOTH were injured in an explosion of this very sensitive substance.

Gold evidence

In 1814, when Davy damaged his eyeight in an accident with nitro-glycerine, 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 & 1853 the reconditioning of both-works central plant, which will

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

The screenshot shows a web browser window for the game "Fool Me Twice". The title bar says "Fool Me Twice - DEBUT AUTHOR: 10068 POINTS". The main content area displays a challenge about Michael Faraday. A yellow box highlights the statement: "Faraday succumbed to his injuries in an explosion preparing nitrogen trichloride samples in 1812". Below this, a button says "SAVE FALSE STATEMENT". Another yellow box highlights the statement: "In 1813, when Gaye consulted Faraday with the preparation of nitrogen trichloride, he included Faraday's famous oil bath solution". Below this, a button says "Gold evidence". A third yellow box highlights the statement: "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". Below this, a button says "Gold evidence". The browser toolbar at the top includes icons for Home, Stop, Back, Forward, Refresh, and various tabs.

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 1812

Evidence (1 marked as gold)

Very soon Gaye entrusted Faraday with the preparation of nitrogen trichloride samples, and they BOTH were exposed in an explosion of this very sensitive substance

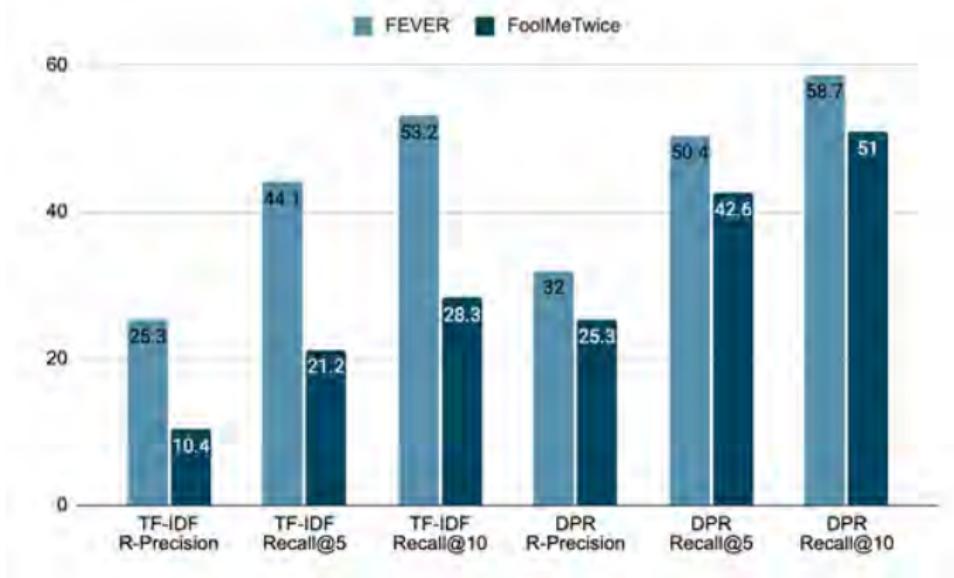
Gold evidence

In 1813, when Gaye consulted Faraday with the preparation of nitrogen trichloride, he included Faraday's famous oil bath solution

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

It is Harder (a bit)



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

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

IR

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

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

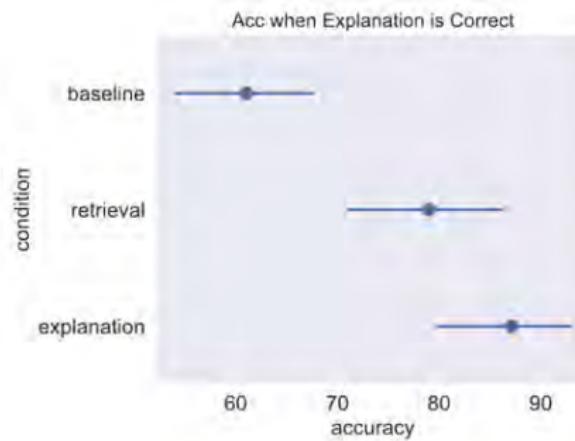
Direct

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

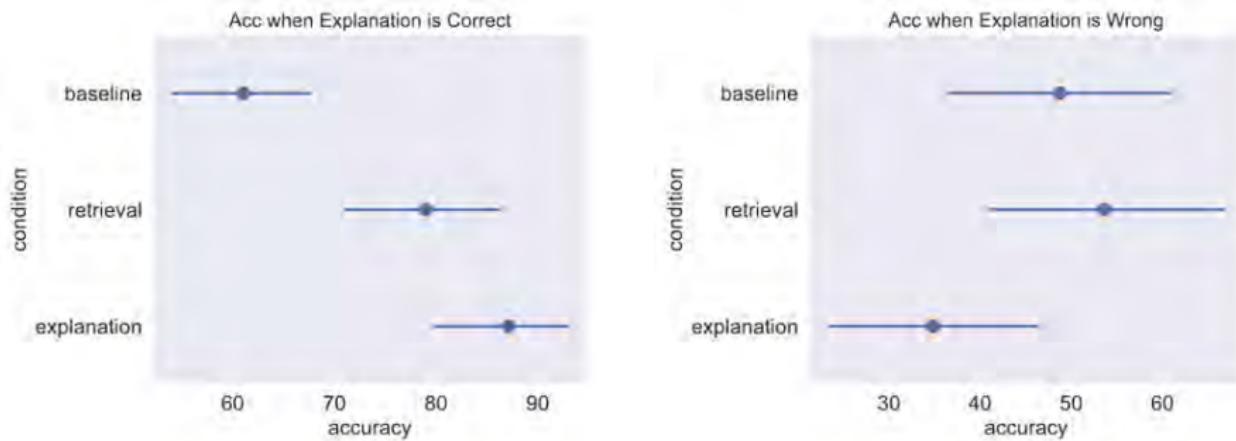
IR

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

Results

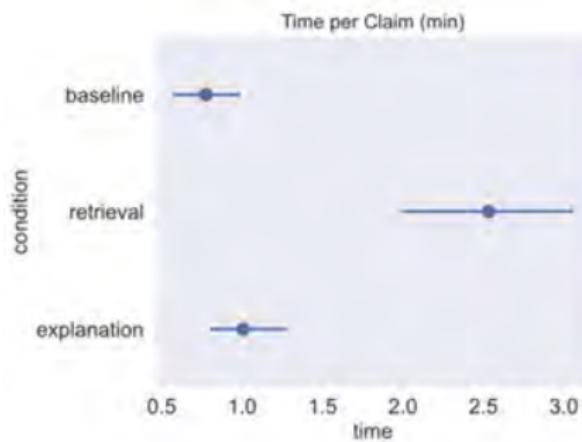


Results



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 Gor¹

Hal Daumé III^{1,2}

Tianyi Zhou¹

Jordan Boyd-Graber¹

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



Item Response Theory

SAT		SAT I: Reasoning Test			Page 1	Use a No. 2 pencil only. Be sure each mark is dark and completely fills the intended oval. Completely erase any errors or stray marks.																																	
1. Your Name		2. Your Name: _____ First name _____ Last name _____ Middle initial _____			Free _____ Paid _____		IMPORTANT: Fill in items 8 and 9 exactly as shown on the back of test book.																																
First 4 letters of Last Name		First Init.	Middle Initial		Date: _____ / _____ / _____																																		
<table border="1"><tr><td><input type="radio"/> A</td><td><input type="radio"/> B</td><td><input type="radio"/> C</td><td><input type="radio"/> D</td><td><input type="radio"/> E</td><td><input type="radio"/> F</td><td><input type="radio"/> G</td><td><input type="radio"/> H</td></tr><tr><td><input type="radio"/> I</td><td><input type="radio"/> J</td><td><input type="radio"/> K</td><td><input type="radio"/> L</td><td><input type="radio"/> M</td><td><input type="radio"/> N</td><td><input type="radio"/> O</td><td><input type="radio"/> P</td></tr><tr><td><input type="radio"/> Q</td><td><input type="radio"/> R</td><td><input type="radio"/> S</td><td><input type="radio"/> T</td><td><input type="radio"/> U</td><td><input type="radio"/> V</td><td><input type="radio"/> W</td><td><input type="radio"/> X</td></tr><tr><td><input type="radio"/> Y</td><td><input type="radio"/> Z</td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table>								<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	<input type="radio"/> E	<input type="radio"/> F	<input type="radio"/> G	<input type="radio"/> H	<input type="radio"/> I	<input type="radio"/> J	<input type="radio"/> K	<input type="radio"/> L	<input type="radio"/> M	<input type="radio"/> N	<input type="radio"/> O	<input type="radio"/> P	<input type="radio"/> Q	<input type="radio"/> R	<input type="radio"/> S	<input type="radio"/> T	<input type="radio"/> U	<input type="radio"/> V	<input type="radio"/> W	<input type="radio"/> X	<input type="radio"/> Y	<input type="radio"/> Z						
<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	<input type="radio"/> E	<input type="radio"/> F	<input type="radio"/> G	<input type="radio"/> H																																
<input type="radio"/> I	<input type="radio"/> J	<input type="radio"/> K	<input type="radio"/> L	<input type="radio"/> M	<input type="radio"/> N	<input type="radio"/> O	<input type="radio"/> P																																
<input type="radio"/> Q	<input type="radio"/> R	<input type="radio"/> S	<input type="radio"/> T	<input type="radio"/> U	<input type="radio"/> V	<input type="radio"/> W	<input type="radio"/> X																																
<input type="radio"/> Y	<input type="radio"/> Z																																						
I agree to the conditions on the back of the SAT I test book. Signature: _____																																							
Home Address: _____ (Room) _____ Apartment or Room _____ City: _____ State: _____ Zip Code: _____																																							
Center: _____ (Phone) _____ City _____ State _____ Center Number _____																																							
3. Date of Birth																																							
Month		Day	Year																																				
Jan.	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7	<input type="radio"/> 8																															
Feb.	<input type="radio"/> 9	<input type="radio"/> 10	<input type="radio"/> 11	<input type="radio"/> 12	<input type="radio"/> 13	<input type="radio"/> 14	<input type="radio"/> 15	<input type="radio"/> 16																															
Mar.	<input type="radio"/> 17	<input type="radio"/> 18	<input type="radio"/> 19	<input type="radio"/> 20	<input type="radio"/> 21	<input type="radio"/> 22	<input type="radio"/> 23	<input type="radio"/> 24																															
Apr.	<input type="radio"/> 25	<input type="radio"/> 26	<input type="radio"/> 27	<input type="radio"/> 28	<input type="radio"/> 29	<input type="radio"/> 30	<input type="radio"/> 31																																
May	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7	<input type="radio"/> 8																															
June	<input type="radio"/> 9	<input type="radio"/> 10	<input type="radio"/> 11	<input type="radio"/> 12	<input type="radio"/> 13	<input type="radio"/> 14	<input type="radio"/> 15	<input type="radio"/> 16																															
July	<input type="radio"/> 17	<input type="radio"/> 18	<input type="radio"/> 19	<input type="radio"/> 20	<input type="radio"/> 21	<input type="radio"/> 22	<input type="radio"/> 23	<input type="radio"/> 24																															
Aug.	<input type="radio"/> 25	<input type="radio"/> 26	<input type="radio"/> 27	<input type="radio"/> 28	<input type="radio"/> 29	<input type="radio"/> 30	<input type="radio"/> 31																																
Sept.	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7	<input type="radio"/> 8																															
Oct.	<input type="radio"/> 9	<input type="radio"/> 10	<input type="radio"/> 11	<input type="radio"/> 12	<input type="radio"/> 13	<input type="radio"/> 14	<input type="radio"/> 15	<input type="radio"/> 16																															
Nov.	<input type="radio"/> 17	<input type="radio"/> 18	<input type="radio"/> 19	<input type="radio"/> 20	<input type="radio"/> 21	<input type="radio"/> 22	<input type="radio"/> 23	<input type="radio"/> 24																															
Dec.	<input type="radio"/> 25	<input type="radio"/> 26	<input type="radio"/> 27	<input type="radio"/> 28	<input type="radio"/> 29	<input type="radio"/> 30	<input type="radio"/> 31																																
4. Social Security Number																																							
<table border="1"><tr><td><input type="radio"/> A</td><td><input type="radio"/> B</td><td><input type="radio"/> C</td><td><input type="radio"/> D</td><td><input type="radio"/> E</td><td><input type="radio"/> F</td><td><input type="radio"/> G</td><td><input type="radio"/> H</td></tr><tr><td><input type="radio"/> I</td><td><input type="radio"/> J</td><td><input type="radio"/> K</td><td><input type="radio"/> L</td><td><input type="radio"/> M</td><td><input type="radio"/> N</td><td><input type="radio"/> O</td><td><input type="radio"/> P</td></tr><tr><td><input type="radio"/> Q</td><td><input type="radio"/> R</td><td><input type="radio"/> S</td><td><input type="radio"/> T</td><td><input type="radio"/> U</td><td><input type="radio"/> V</td><td><input type="radio"/> W</td><td><input type="radio"/> X</td></tr><tr><td><input type="radio"/> Y</td><td><input type="radio"/> Z</td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table>								<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	<input type="radio"/> E	<input type="radio"/> F	<input type="radio"/> G	<input type="radio"/> H	<input type="radio"/> I	<input type="radio"/> J	<input type="radio"/> K	<input type="radio"/> L	<input type="radio"/> M	<input type="radio"/> N	<input type="radio"/> O	<input type="radio"/> P	<input type="radio"/> Q	<input type="radio"/> R	<input type="radio"/> S	<input type="radio"/> T	<input type="radio"/> U	<input type="radio"/> V	<input type="radio"/> W	<input type="radio"/> X	<input type="radio"/> Y	<input type="radio"/> Z						
<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	<input type="radio"/> E	<input type="radio"/> F	<input type="radio"/> G	<input type="radio"/> H																																
<input type="radio"/> I	<input type="radio"/> J	<input type="radio"/> K	<input type="radio"/> L	<input type="radio"/> M	<input type="radio"/> N	<input type="radio"/> O	<input type="radio"/> P																																
<input type="radio"/> Q	<input type="radio"/> R	<input type="radio"/> S	<input type="radio"/> T	<input type="radio"/> U	<input type="radio"/> V	<input type="radio"/> W	<input type="radio"/> X																																
<input type="radio"/> Y	<input type="radio"/> Z																																						
5. Registration Number (Copy from Admission Ticket 1)																																							
6. Test Book Serial Number (Copy from front of test book.)																																							
7. Test Book Serial Number (Copy from front of test book.)																																							
<table border="1"><tr><td><input type="radio"/> A</td><td><input type="radio"/> B</td><td><input type="radio"/> C</td><td><input type="radio"/> D</td><td><input type="radio"/> E</td><td><input type="radio"/> F</td><td><input type="radio"/> G</td><td><input type="radio"/> H</td></tr><tr><td><input type="radio"/> I</td><td><input type="radio"/> J</td><td><input type="radio"/> K</td><td><input type="radio"/> L</td><td><input type="radio"/> M</td><td><input type="radio"/> N</td><td><input type="radio"/> O</td><td><input type="radio"/> P</td></tr><tr><td><input type="radio"/> Q</td><td><input type="radio"/> R</td><td><input type="radio"/> S</td><td><input type="radio"/> T</td><td><input type="radio"/> U</td><td><input type="radio"/> V</td><td><input type="radio"/> W</td><td><input type="radio"/> X</td></tr><tr><td><input type="radio"/> Y</td><td><input type="radio"/> Z</td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table>								<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	<input type="radio"/> E	<input type="radio"/> F	<input type="radio"/> G	<input type="radio"/> H	<input type="radio"/> I	<input type="radio"/> J	<input type="radio"/> K	<input type="radio"/> L	<input type="radio"/> M	<input type="radio"/> N	<input type="radio"/> O	<input type="radio"/> P	<input type="radio"/> Q	<input type="radio"/> R	<input type="radio"/> S	<input type="radio"/> T	<input type="radio"/> U	<input type="radio"/> V	<input type="radio"/> W	<input type="radio"/> X	<input type="radio"/> Y	<input type="radio"/> Z						
<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	<input type="radio"/> E	<input type="radio"/> F	<input type="radio"/> G	<input type="radio"/> H																																
<input type="radio"/> I	<input type="radio"/> J	<input type="radio"/> K	<input type="radio"/> L	<input type="radio"/> M	<input type="radio"/> N	<input type="radio"/> O	<input type="radio"/> P																																
<input type="radio"/> Q	<input type="radio"/> R	<input type="radio"/> S	<input type="radio"/> T	<input type="radio"/> U	<input type="radio"/> V	<input type="radio"/> W	<input type="radio"/> X																																
<input type="radio"/> Y	<input type="radio"/> Z																																						

Item Response Theory

SAT

SAT I: Reasoning Test

Page 1

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

1. Your Name

First 4 letters of Last Name	First Init.	Mid. Init.
<input type="radio"/> A <input type="radio"/> B <input type="radio"/> C <input type="radio"/> D <input type="radio"/> E <input type="radio"/> F <input type="radio"/> G <input type="radio"/> H <input type="radio"/> I <input type="radio"/> J <input type="radio"/> K <input type="radio"/> L <input type="radio"/> M <input type="radio"/> N <input type="radio"/> O <input type="radio"/> P <input type="radio"/> Q <input type="radio"/> R <input type="radio"/> S <input type="radio"/> T <input type="radio"/> U <input type="radio"/> V <input type="radio"/> W <input type="radio"/> X <input type="radio"/> Y <input type="radio"/> Z		
<input type="radio"/> A <input type="radio"/> B <input type="radio"/> C <input type="radio"/> D <input type="radio"/> E <input type="radio"/> F <input type="radio"/> G <input type="radio"/> H <input type="radio"/> I <input type="radio"/> J <input type="radio"/> K <input type="radio"/> L <input type="radio"/> M <input type="radio"/> N <input type="radio"/> O <input type="radio"/> P <input type="radio"/> Q <input type="radio"/> R <input type="radio"/> S <input type="radio"/> T <input type="radio"/> U <input type="radio"/> V <input type="radio"/> W <input type="radio"/> X <input type="radio"/> Y <input type="radio"/> Z		
<input type="radio"/> A <input type="radio"/> B <input type="radio"/> C <input type="radio"/> D <input type="radio"/> E <input type="radio"/> F <input type="radio"/> G <input type="radio"/> H <input type="radio"/> I <input type="radio"/> J <input type="radio"/> K <input type="radio"/> L <input type="radio"/> M <input type="radio"/> N <input type="radio"/> O <input type="radio"/> P <input type="radio"/> Q <input type="radio"/> R <input type="radio"/> S <input type="radio"/> T <input type="radio"/> U <input type="radio"/> V <input type="radio"/> W <input type="radio"/> X <input type="radio"/> Y <input type="radio"/> Z		
<input type="radio"/> A <input type="radio"/> B <input type="radio"/> C <input type="radio"/> D <input type="radio"/> E <input type="radio"/> F <input type="radio"/> G <input type="radio"/> H <input type="radio"/> I <input type="radio"/> J <input type="radio"/> K <input type="radio"/> L <input type="radio"/> M <input type="radio"/> N <input type="radio"/> O <input type="radio"/> P <input type="radio"/> Q <input type="radio"/> R <input type="radio"/> S <input type="radio"/> T <input type="radio"/> U <input type="radio"/> V <input type="radio"/> W <input type="radio"/> X <input type="radio"/> Y <input type="radio"/> Z		
2. Your Name: First _____ Last _____ I agree to the conditions on the back of the SAT I test book. Signature: _____ Date: _____		
Home Address: Room _____ Number _____ Street _____ City _____ State _____ Zip Code _____ Center: _____ Room: _____ User: _____ Center Number: _____		
3. Date of Birth Year _____ Month _____ Day _____		
4. Social Security Number _____-_____-_____		
5. Registration Number (Copy from Admission Ticket.)		
6. Test Book Serial Number (Copy from front of test book.)		
IMPORTANT: Fill in items 8 and 9 exactly as shown on the back of test book.		
8. Form Code (Do not grid as on back of test book.)		
9. _____		

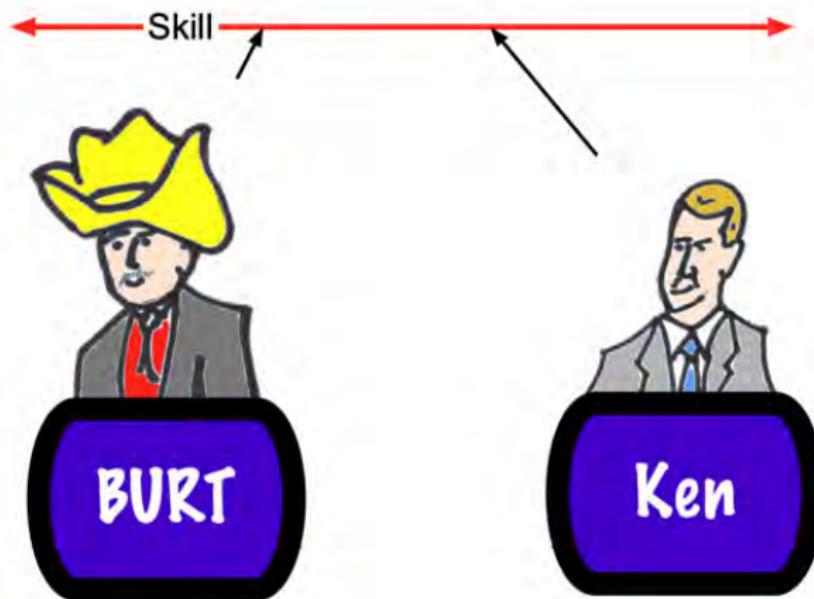
2(ri,j)

Item Response Theory

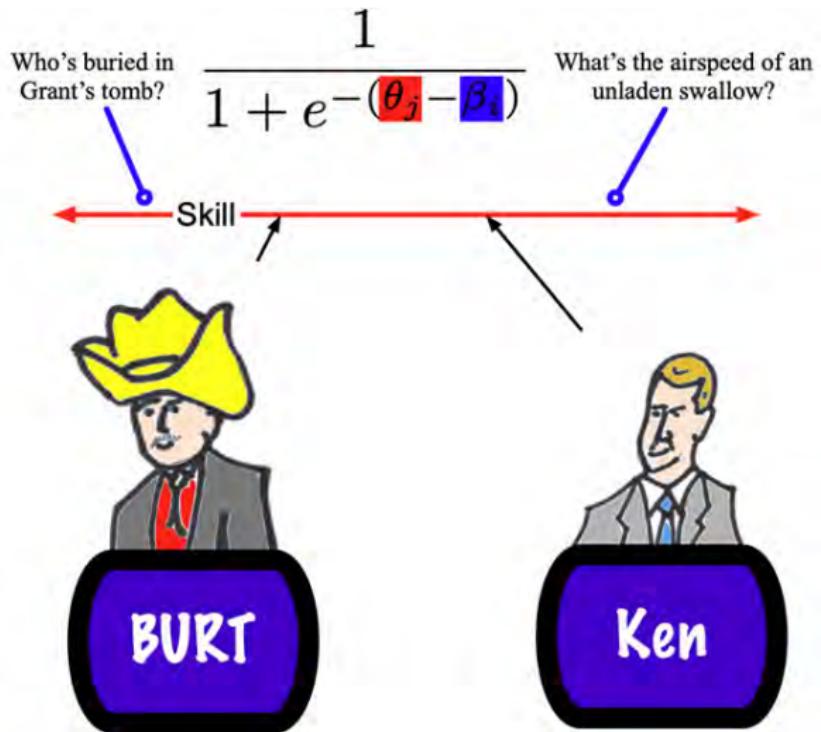


Item Response Theory

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



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?

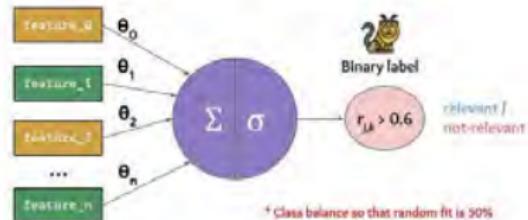
Making Dimensions Interpretable

- Make skills and difficulty vector-valued

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

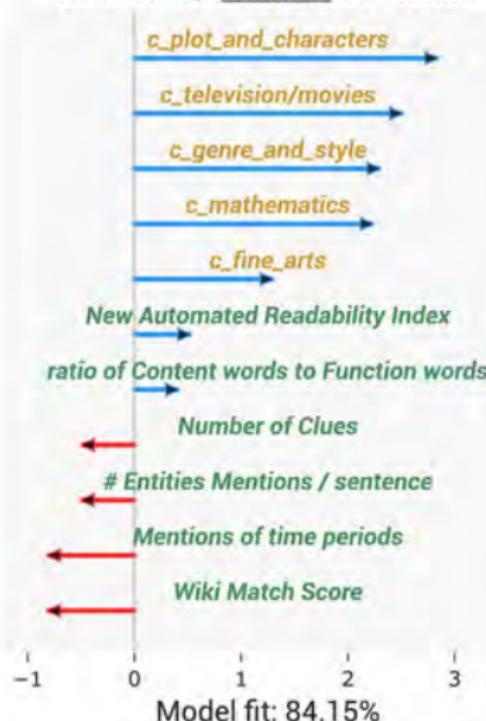
- Where do the dimensions come from?
 - Latent variable: learned to predict correctness
 - Function of question encoding, question features
 - Regularized to be sparse
- Posthoc labeling

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



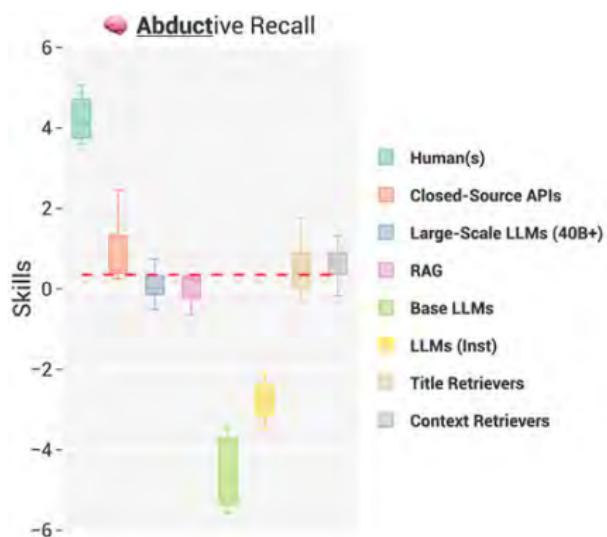
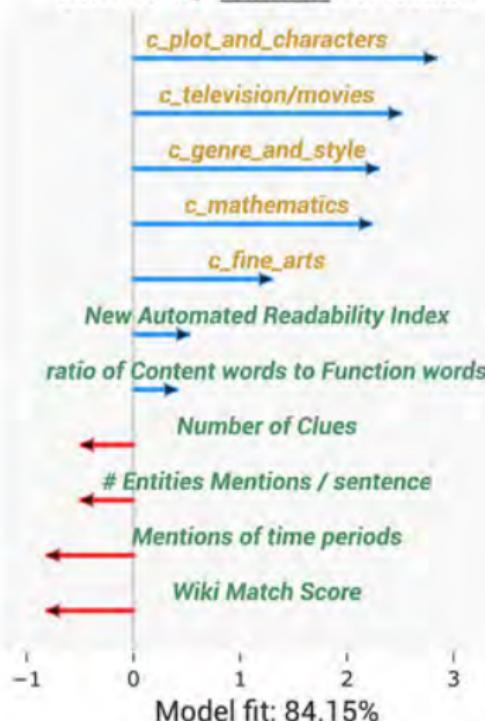
Hard for Computers: Abductive Inference

Dim 1: 🧠 Abductive Recall



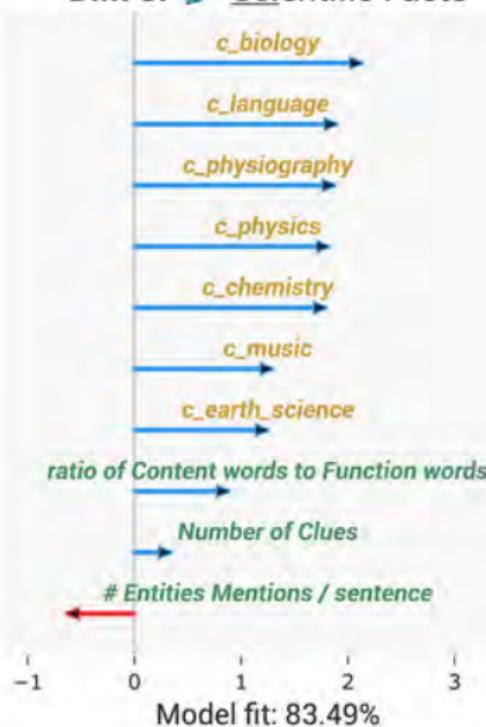
Hard for Computers: Abductive Inference

Dim 1: 🧠 Abductive Recall



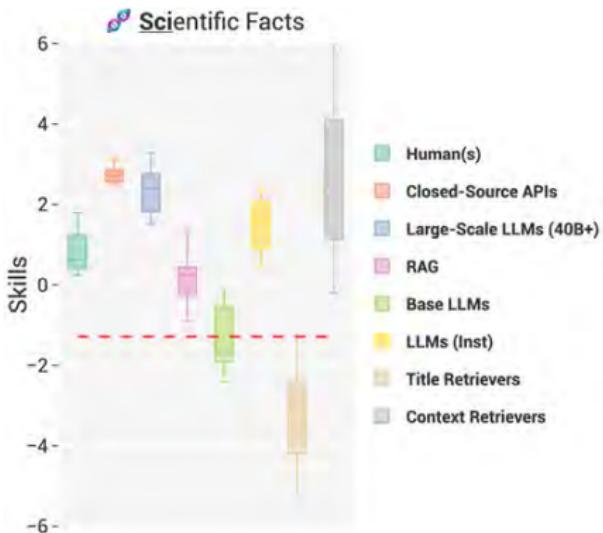
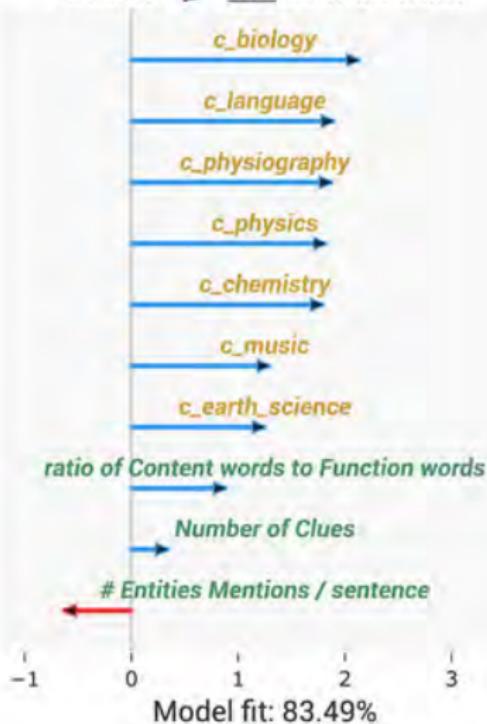
Hard for Humans: Science

Dim 3: Scientific Facts



Hard for Humans: Science

Dim 3: 🧬 Scientific Facts



AdvScore



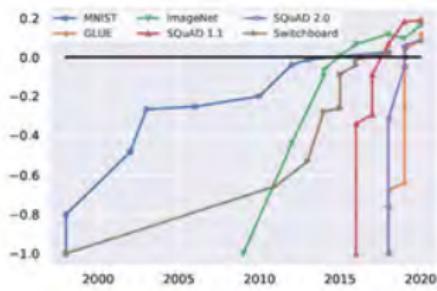
**Is your benchmark *truly* adversarial?
ADVSCORE: Evaluating Human-Grounded Adversarialness**

Yoo Yeon Sung¹, Maharshi Gor¹, Eve Fleisig², Ishani Mondal¹, Jordan Boyd-Graber¹

¹University of Maryland ²UC Berkeley

NAACL 2025 Outstanding Paper

Adversarial Datasets

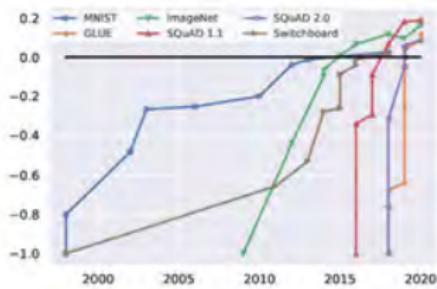


Biggio et al., 2012: Poisoning attacks against Support Vector Machines



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

Adversarial Datasets



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

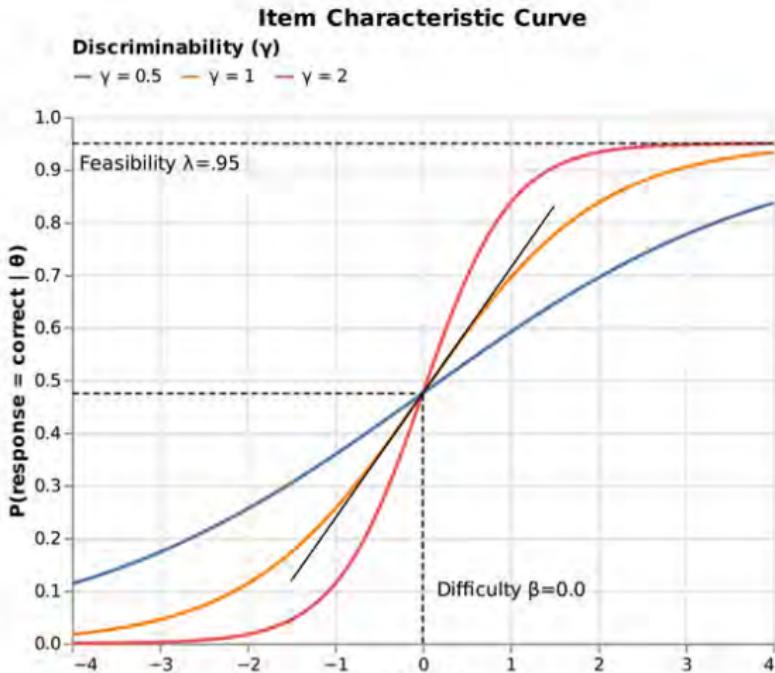


- Many benchmarks are “saturated”
- Newer datasets claim to be “adversarial”
 - Hard for computers, “easy” for humans
 - No real metric / definition
- Can we use the lessons of the previous paper to inform how to write hard examples
- Can we *measure* how well we did?
- Language game: increasing the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

Expanding IRT: Discriminability

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

- Difficulty: β
- Discriminability: γ
- Feasibility: λ
- Skill/ability: θ



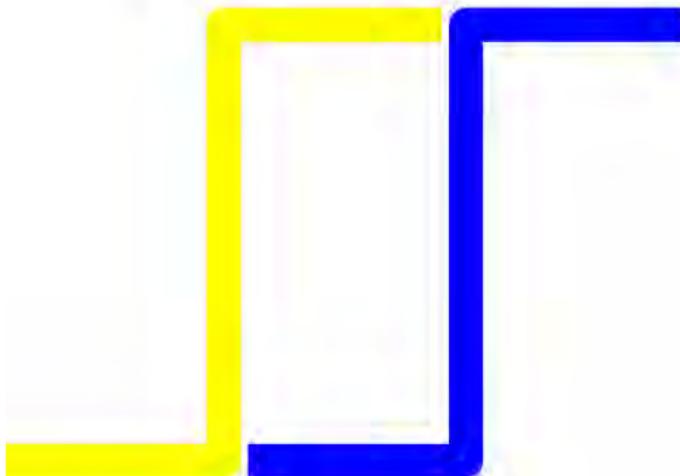
Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2$$



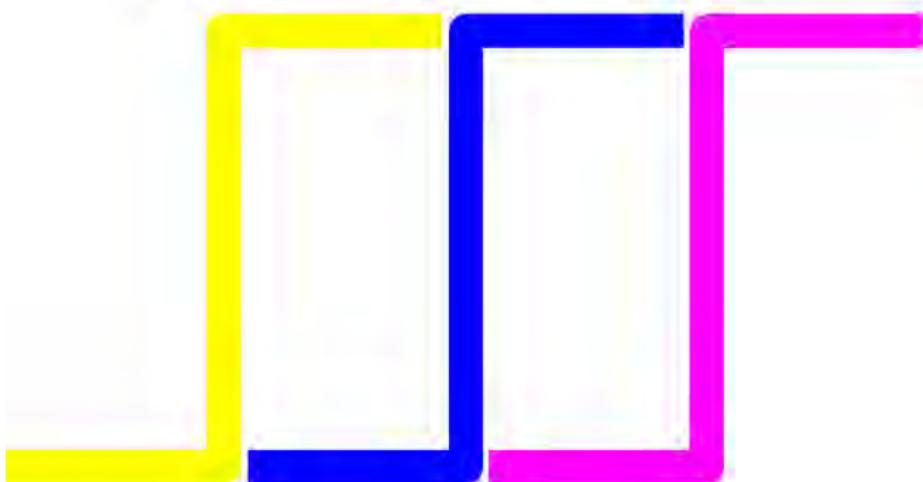
Finding Skill of Subject is Like Binary Search

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



Finding Skill of Subject is Like Binary Search

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



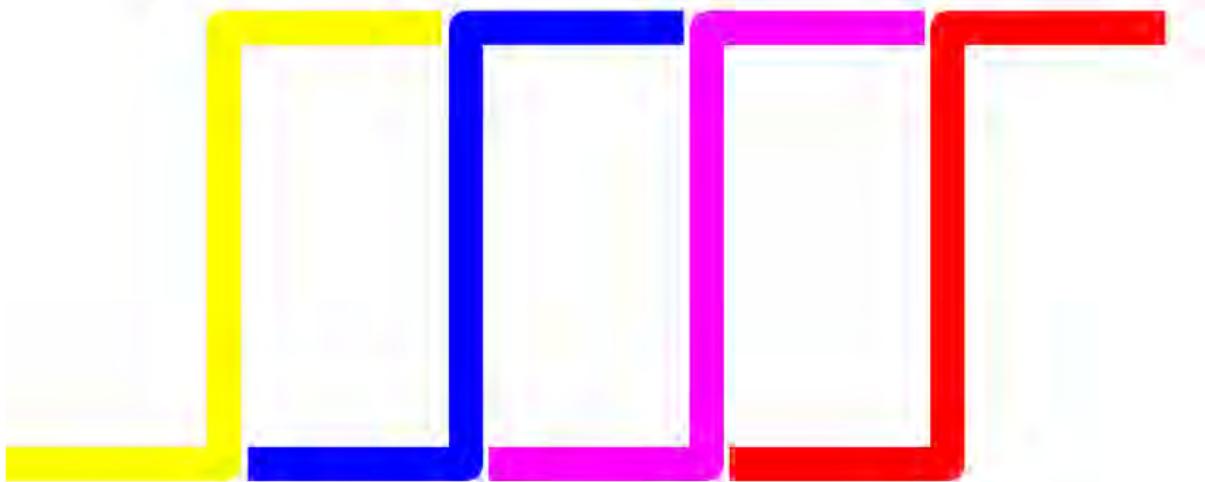
Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2$$

$$\beta_2 = .4$$

$$\beta_3 = .6$$

$$\beta_4 = .8$$



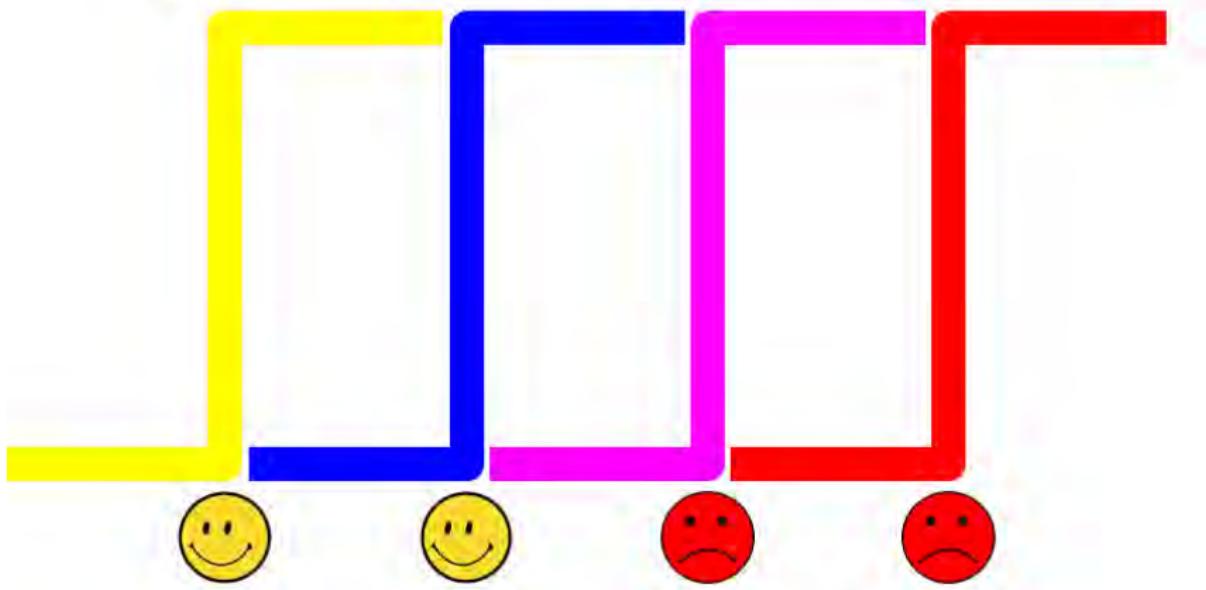
Finding Skill of Subject is Like Binary Search

$$\beta_1 = .2$$

$$\beta_2 = .4$$

$$\beta_3 = .6$$

$$\beta_4 = .8$$



Adversarial Score

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

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

Adversarial Score

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

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

Why not use raw accuracy?

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

Adversarial Score

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

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

- Skilled humans should agree on the answer

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

Adversarial Score

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

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

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

- Skilled humans should agree on the answer

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

Adversarial Score

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

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

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

- Skilled humans should agree on the answer

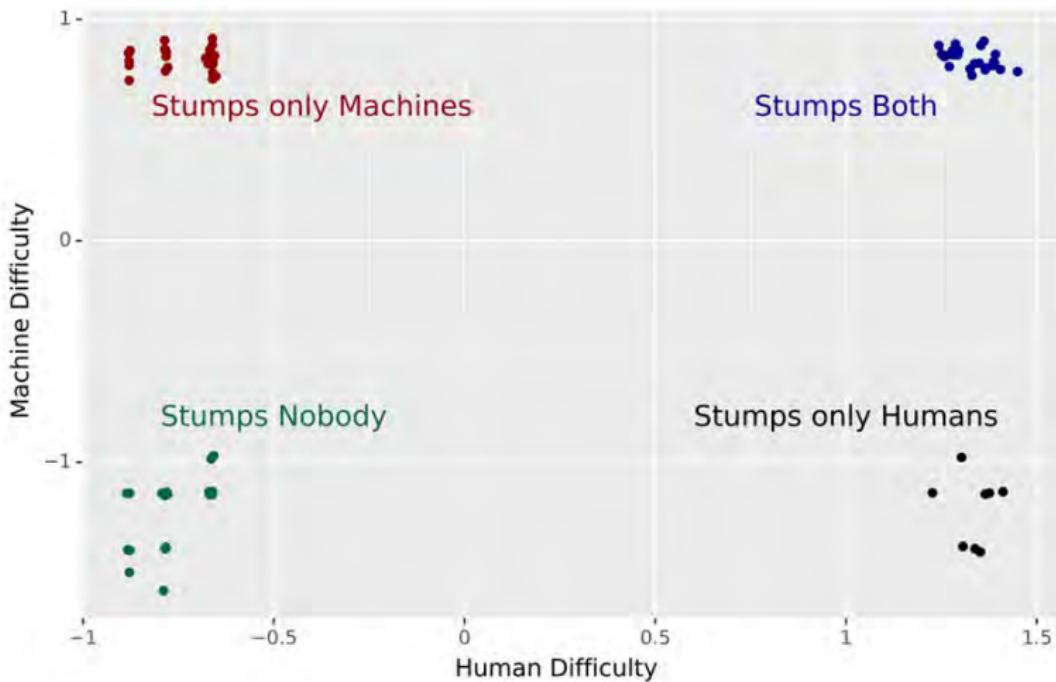
$$\delta_j = \sum_{i \sim H_{(1)}} \left[\frac{1}{1 + \exp \left\{ -\gamma_j \left[\beta_i^{H_{(1)}} - \theta_j \right] \right\}} - \overline{p_{H_{(1)}}}(r_{i,j}) \right] / |H_{(1)}| \quad (4)$$

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

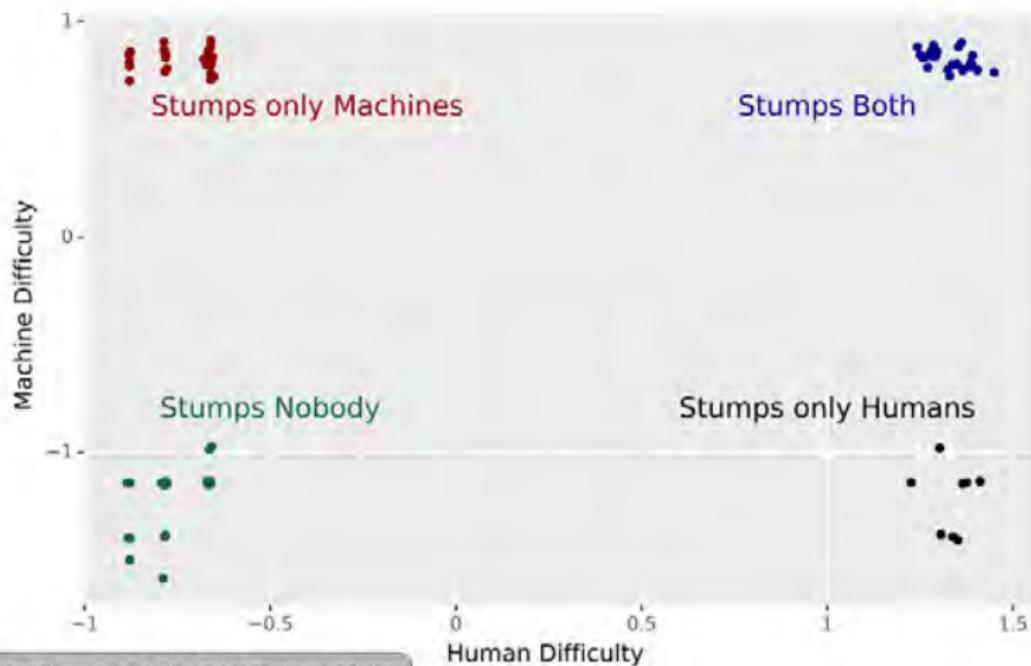
AdvQA: Is this a viable incentive structure?

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

What makes for Adversarial Example

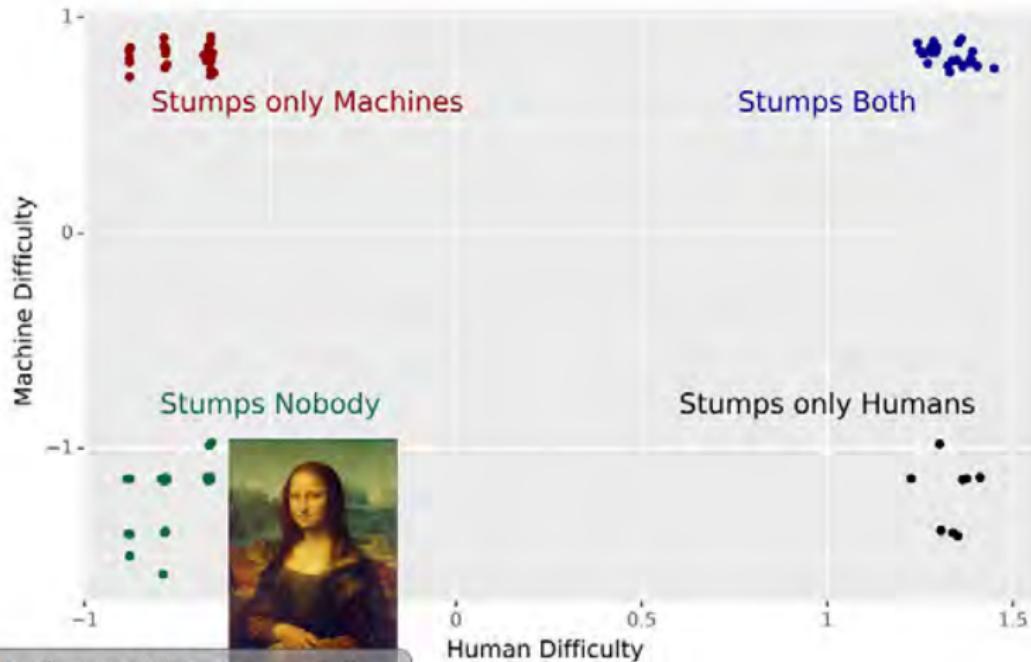


What makes for Adversarial Example



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

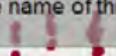
What makes for Adversarial Example



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

Stumps Nobody



0

0.5

1

1.5

Human Difficulty

Stumps only Humans

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

Machine Difficulty

0



Stumps Both

Stumps Nobody



-1

-1

-1

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

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 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

Human Difficulty

0

0.5

1

1.5

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

Machine Difficulty

0 -

Stumps Nobody

-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 Poomaalai and is considered to have one of the best bowling sides?



Stumps Both



Stumps only Humans



0.5
0

Human Difficulty

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

Adversarial Strategies



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

Adversarial Strategies



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

Brad Pitt / Will Smith

Adversarial Strategies



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

Adversarial Strategies



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

Snowpiercer / Train to Busan

Adversarial Strategies



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

Adversarial Strategies



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

Omega / Lululemon

Adversarial Strategies



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

Adversarial Strategies



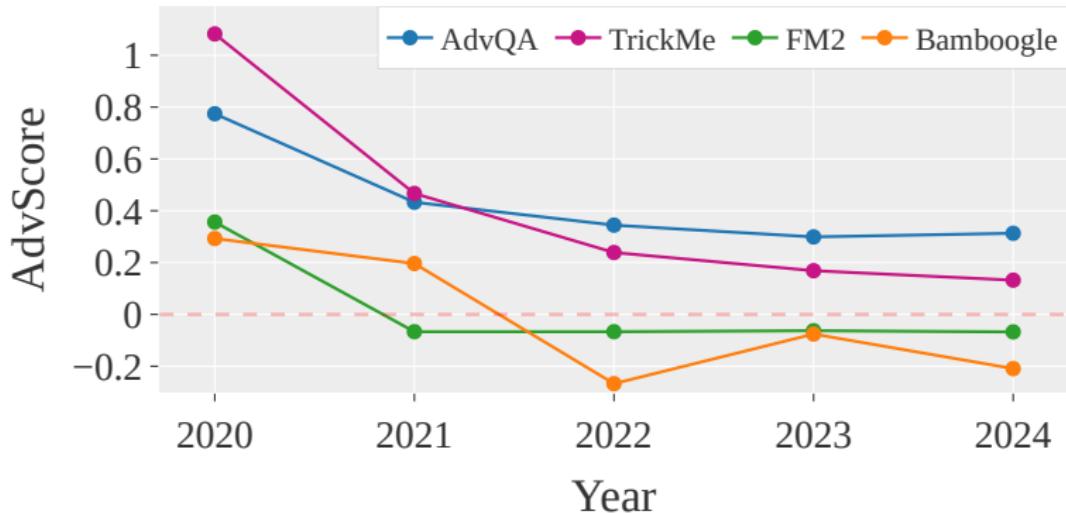
A character in one story by
this author opens *Crime and
Punishment* to discover that it has
turned into *The Brothers
Karamazov*

Adversarial Strategies



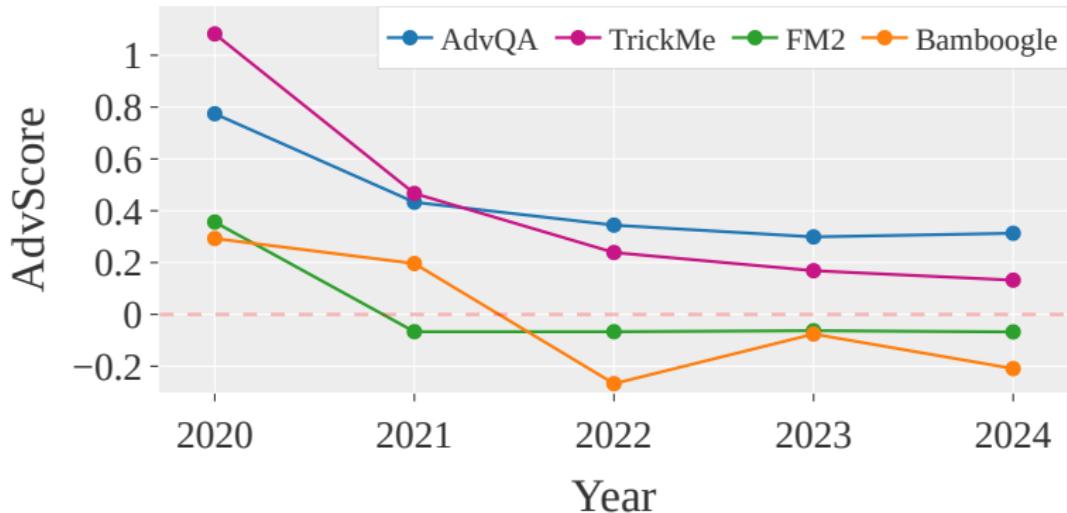
A character in one story by
this author opens Crime and
Punishment to discover that it has
turned into The Brothers
Karamazov
Dostoyevski / Akutagawa

Which Datasets are Adversarial?



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

Which Datasets are Adversarial?



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

початок об 11.00

Готель Radisson Blu
м. Київ, проспект Свободи 22

Олена
БОЙЧУН



Centaur Chess

Buzz

0:30

Guesses

#	Guess	Score
1	Congo River	0.1987
2	Zambezi	0.1121
3	Yukon River	0.0956
4	Irrawaddy River	0.0904
5	Amazon River	0.0864

Question

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

Settings

 Guesses

 Highlights

 Evidence

Pause

Sign Out

Evidence

for Congo River

the Lualaba and the Chambeshi Rivers. It is navigable downstream from Kisangani, except for the area

Falls lies on this river, and after it reaches Kisangani, it is no longer called the Lualaba. This

Instructions

- Press space to buzz
- Press enter to submit
- Use autocomplete to

Players

1 active

#	Score	Name	Country
1	-15	Summer Dew	1/5
2	475	munizmildtown	54
3	285	Cottman	80
4	200	Elton John	20

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

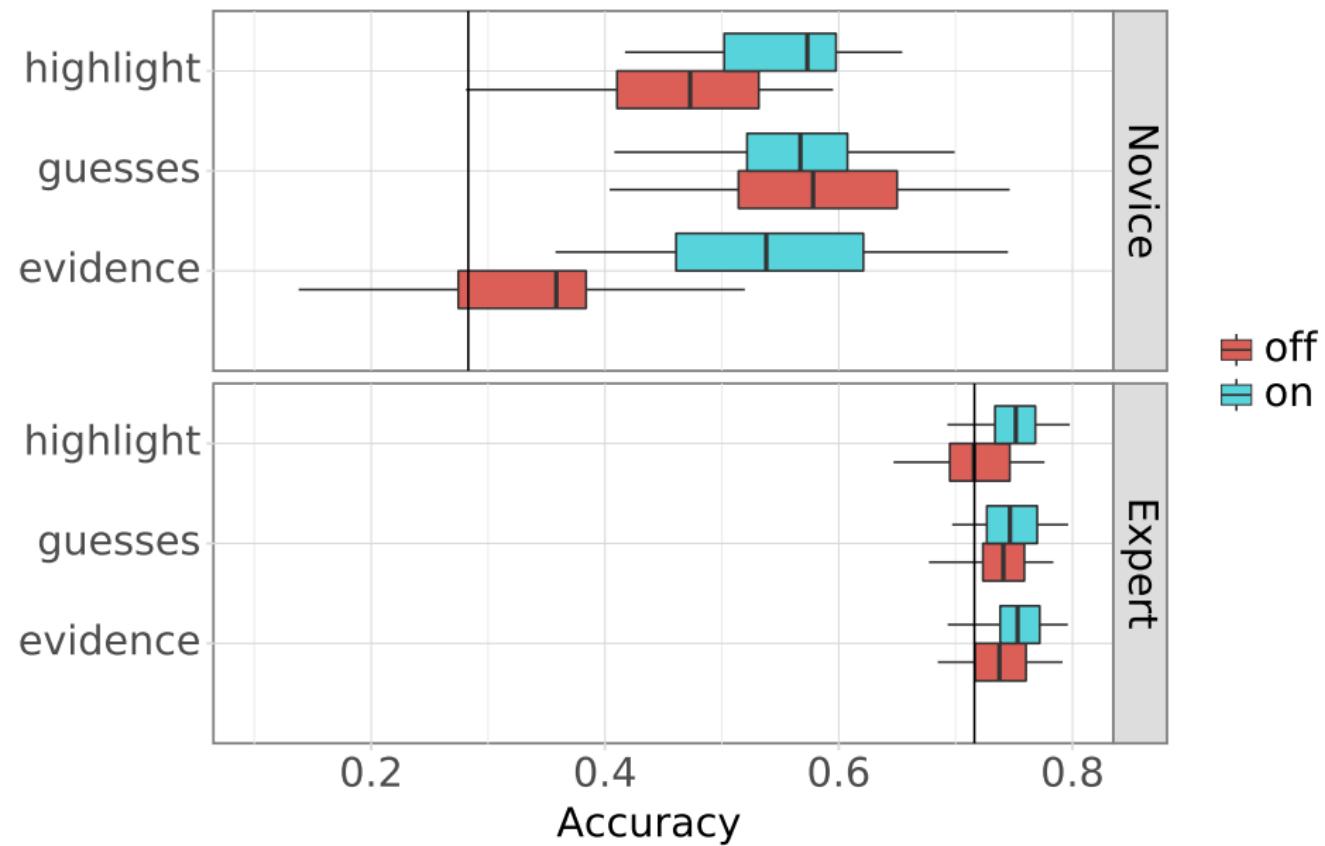
Experts vs. Novices

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

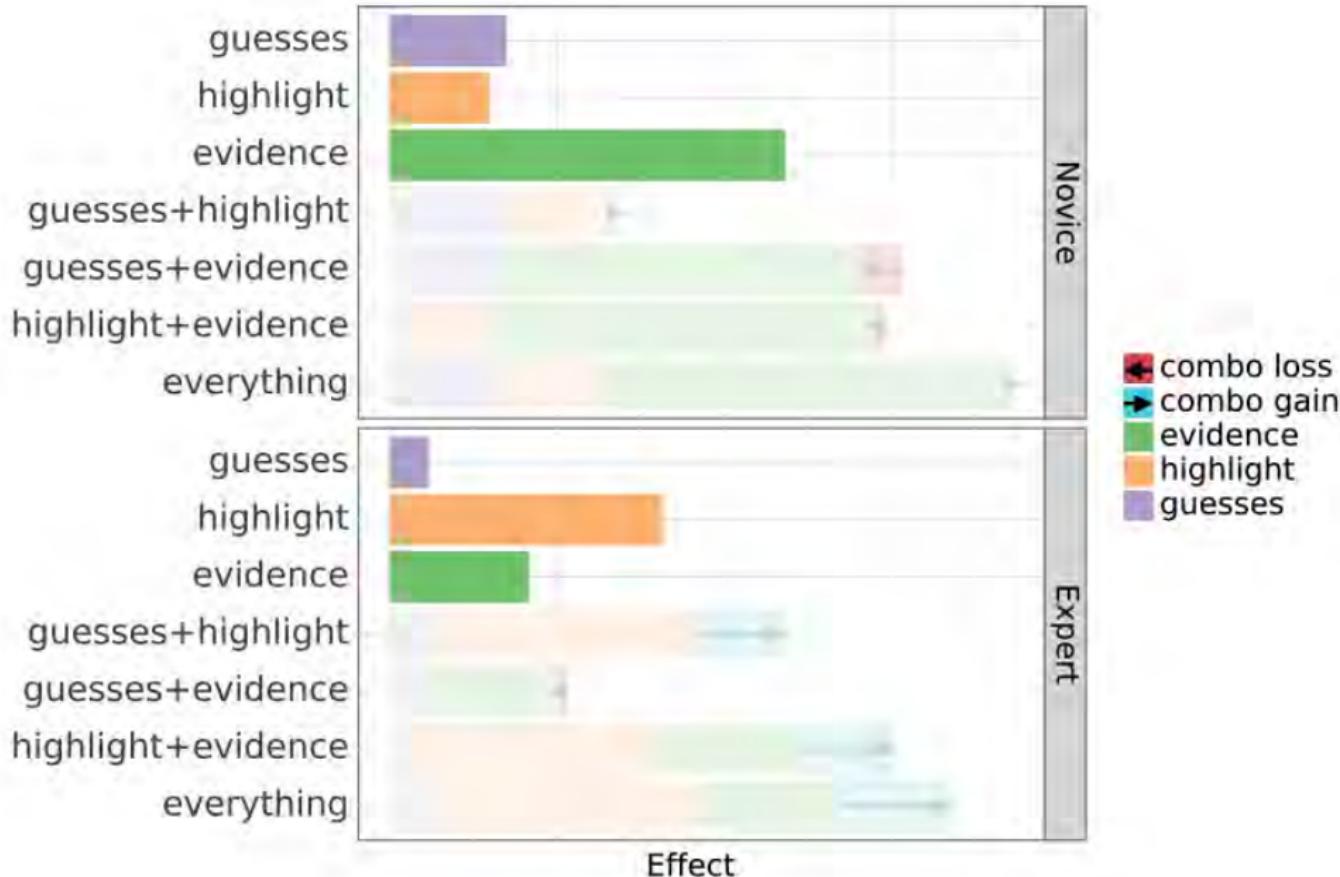
Rerun IRT analysis: rather than contrast human vs. computer skill, we see how much interface boosts (or diminishes) user skill.

Skill Boost

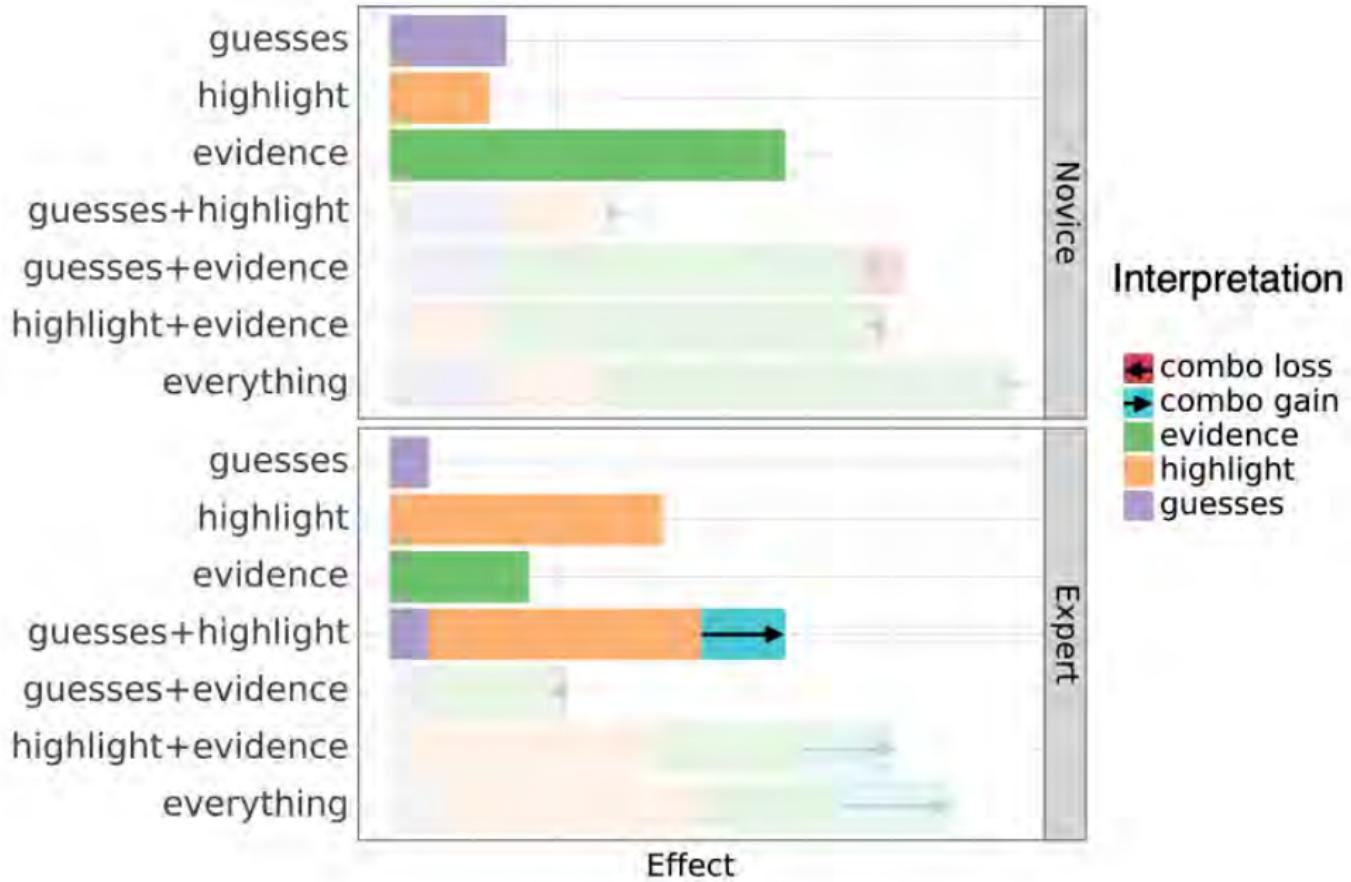
Rerun IRT analysis: rather than contrast human vs. computer skill, we see how much interface boosts (or diminishes) user skill.

Coefficients tell story!

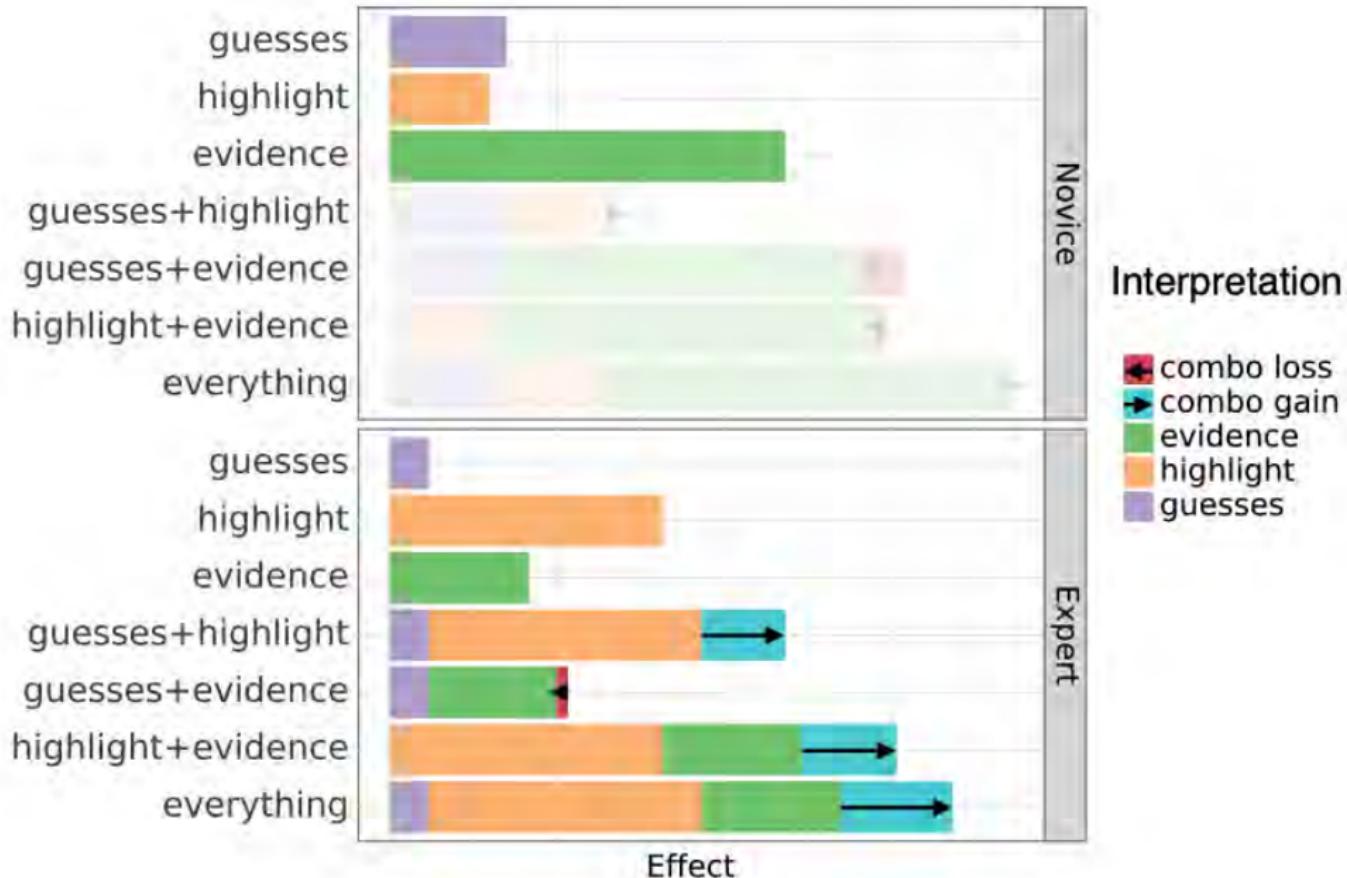
- **Big, Positive:** Help
- **Big, Negative:** Hurt
- **Small:** Neutral



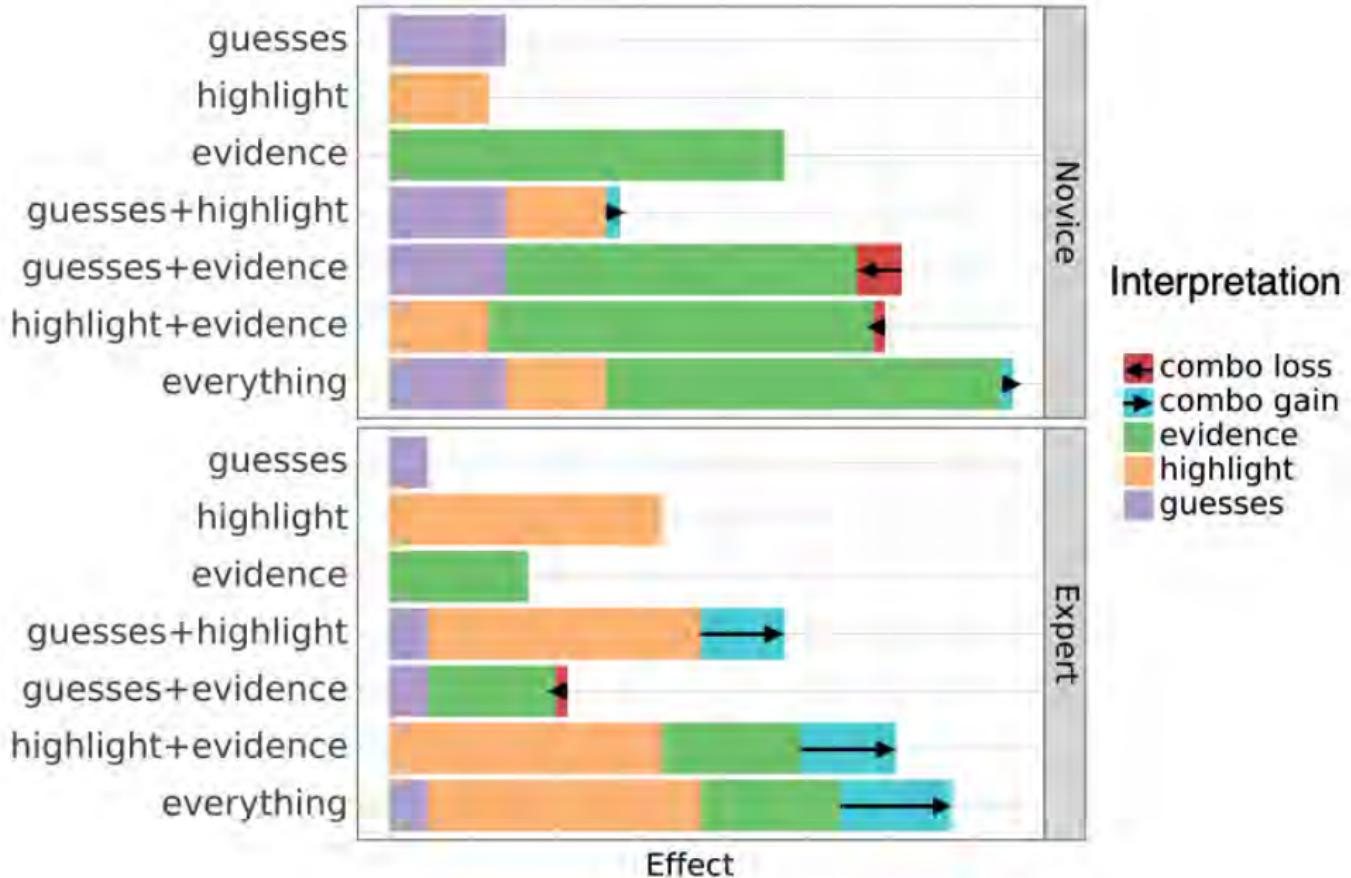
Everything helps: Evidence for novices, Highlight for experts



Synergistic effects



Highlight and evidence help experts most



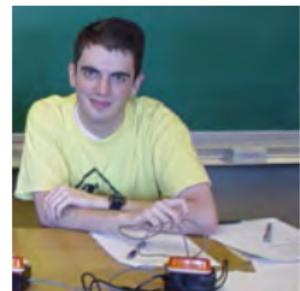
For novices, less synergy



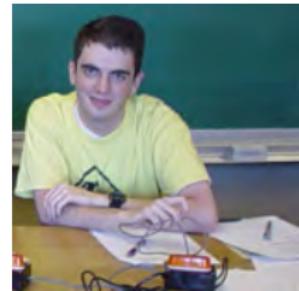
Learning to Explain Selectively

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

Measuring Interpretability

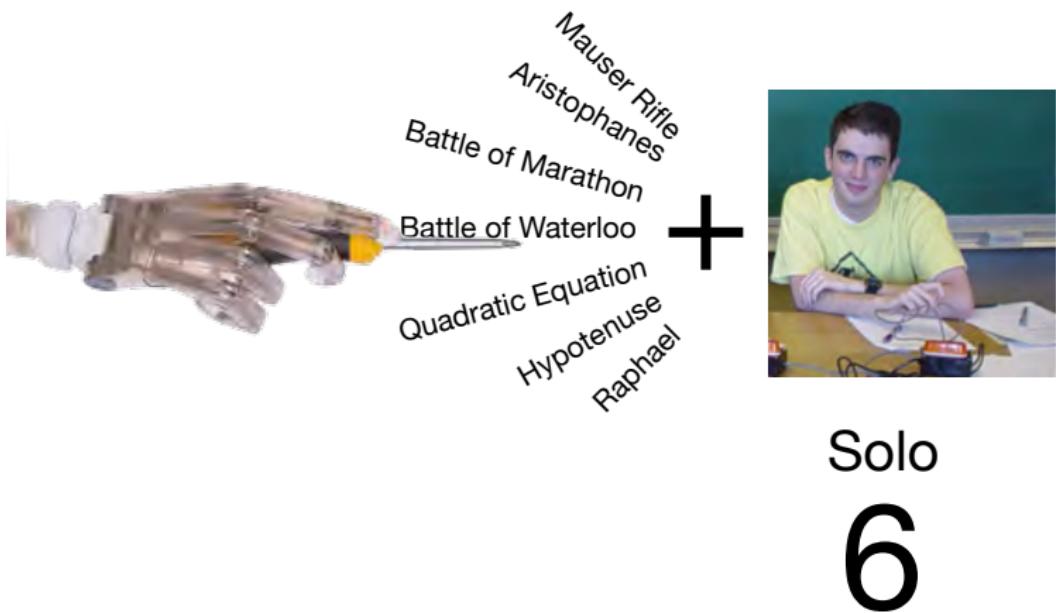


Measuring Interpretability

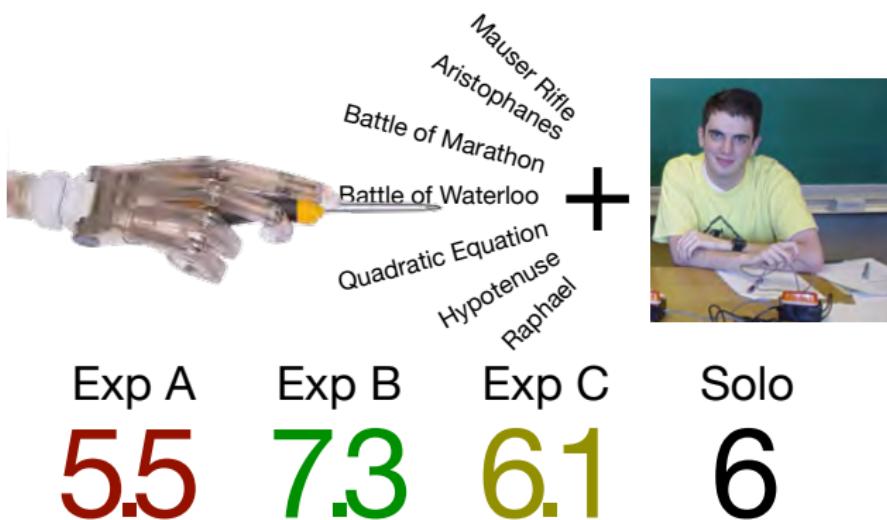


Solo
6

Measuring Interpretability

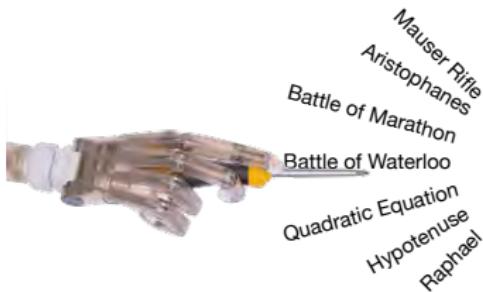


Measuring Interpretability



Improvement through Bandit Algorithms

Visualization



Viz Solo
7.3 6

Improvement through Bandit Algorithms



Improvement through Bandit Algorithms



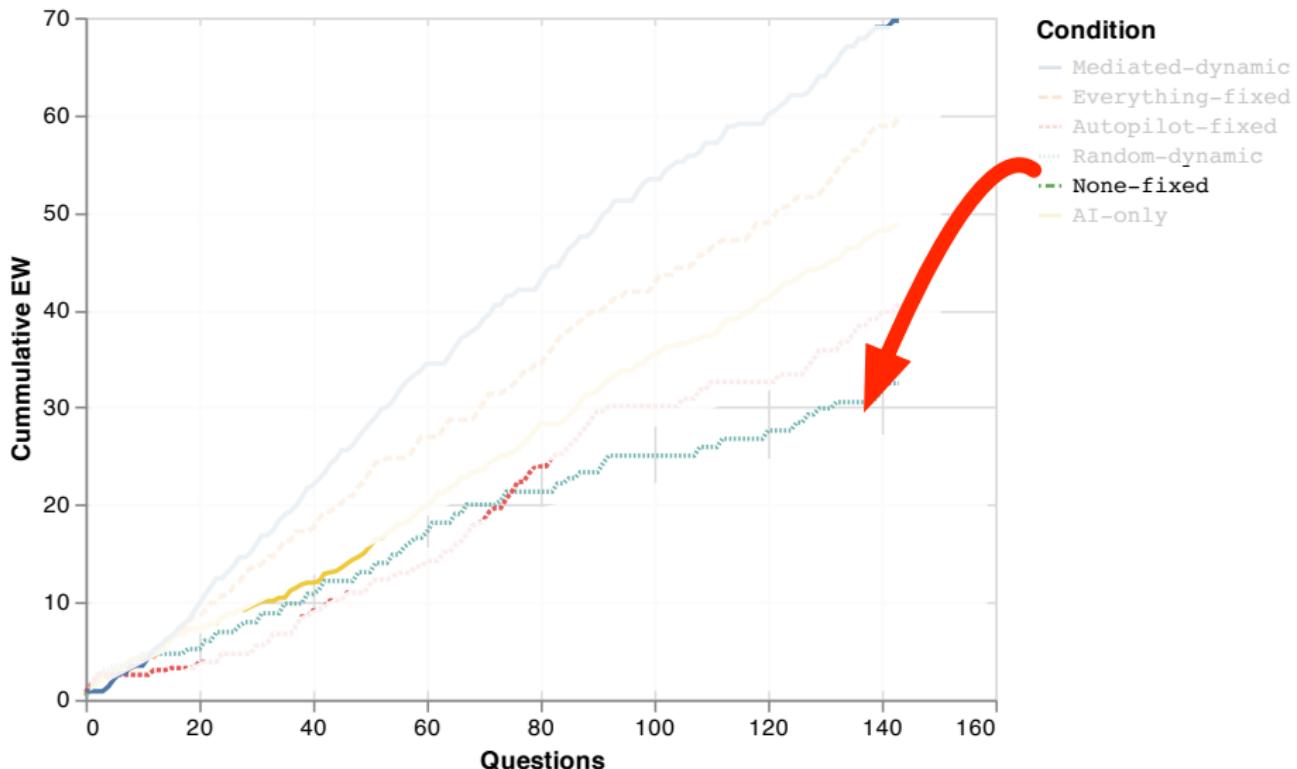
Improvement through Bandit Algorithms



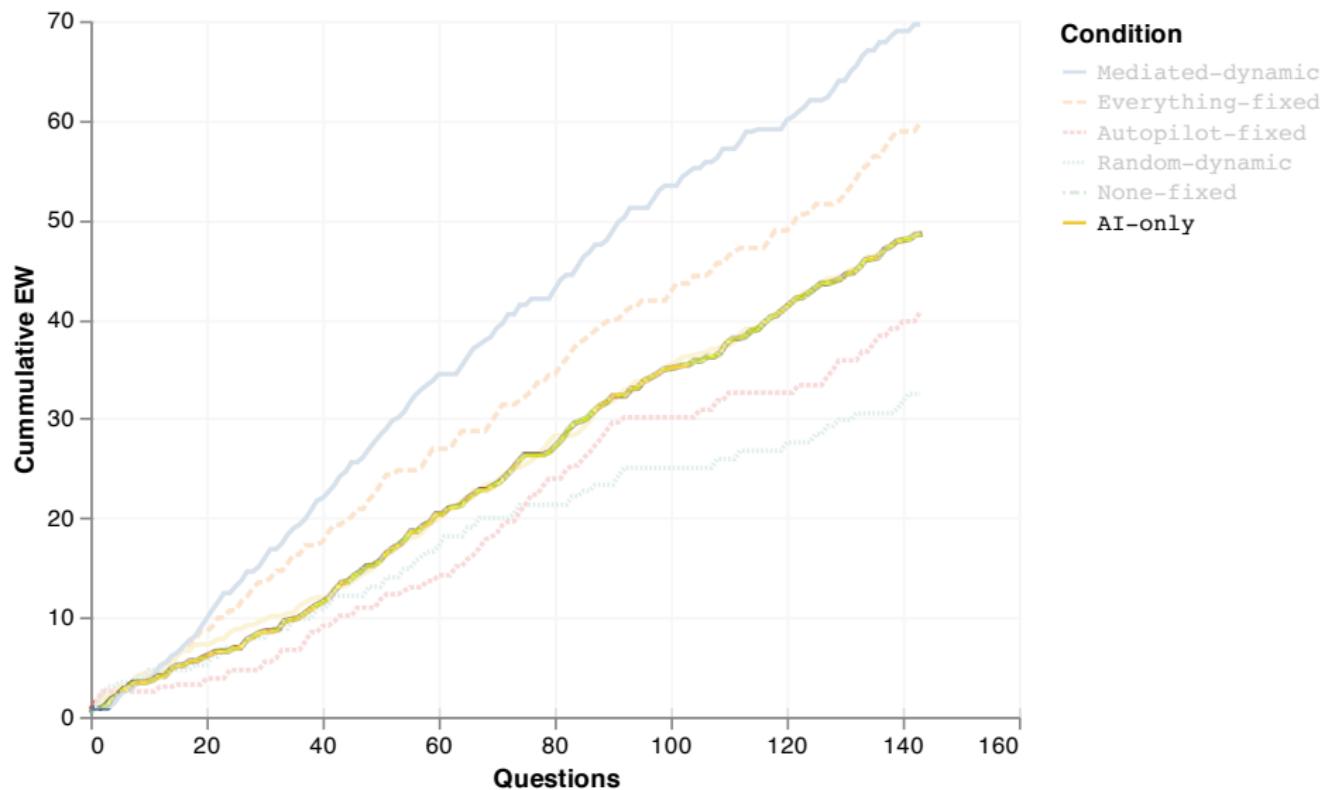
Improvement through Bandit Algorithms



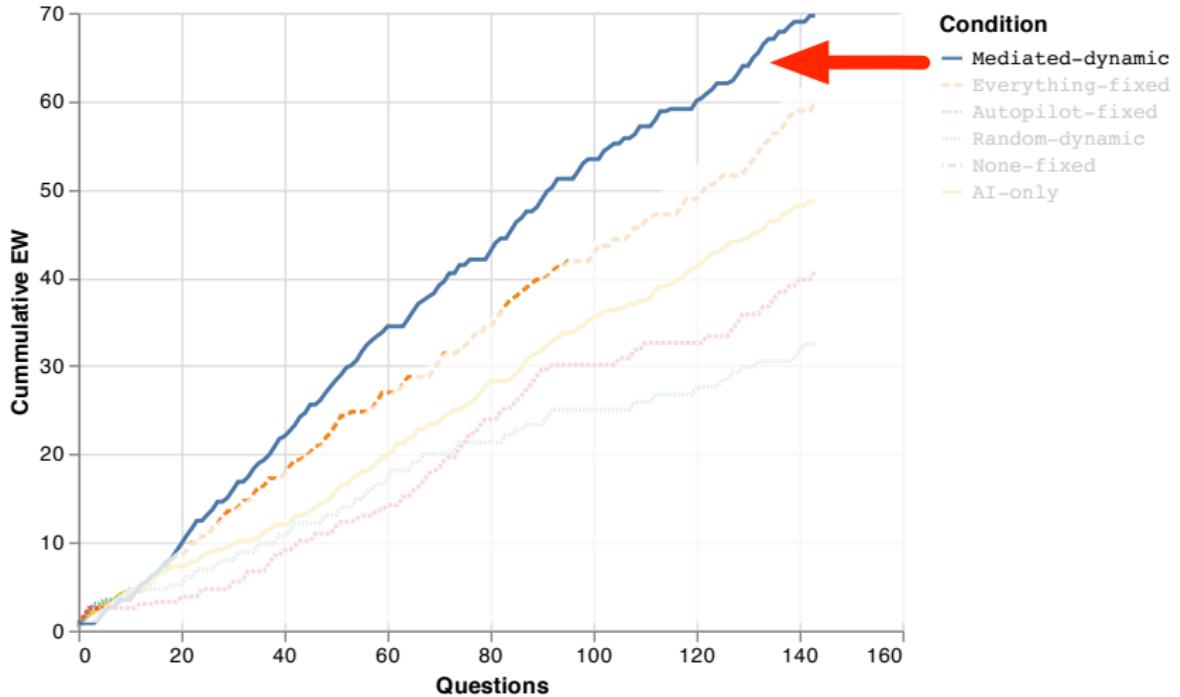
Bandit actions [?]: turn each of the explanations (Guess, Highlight, Evidence) on or off.



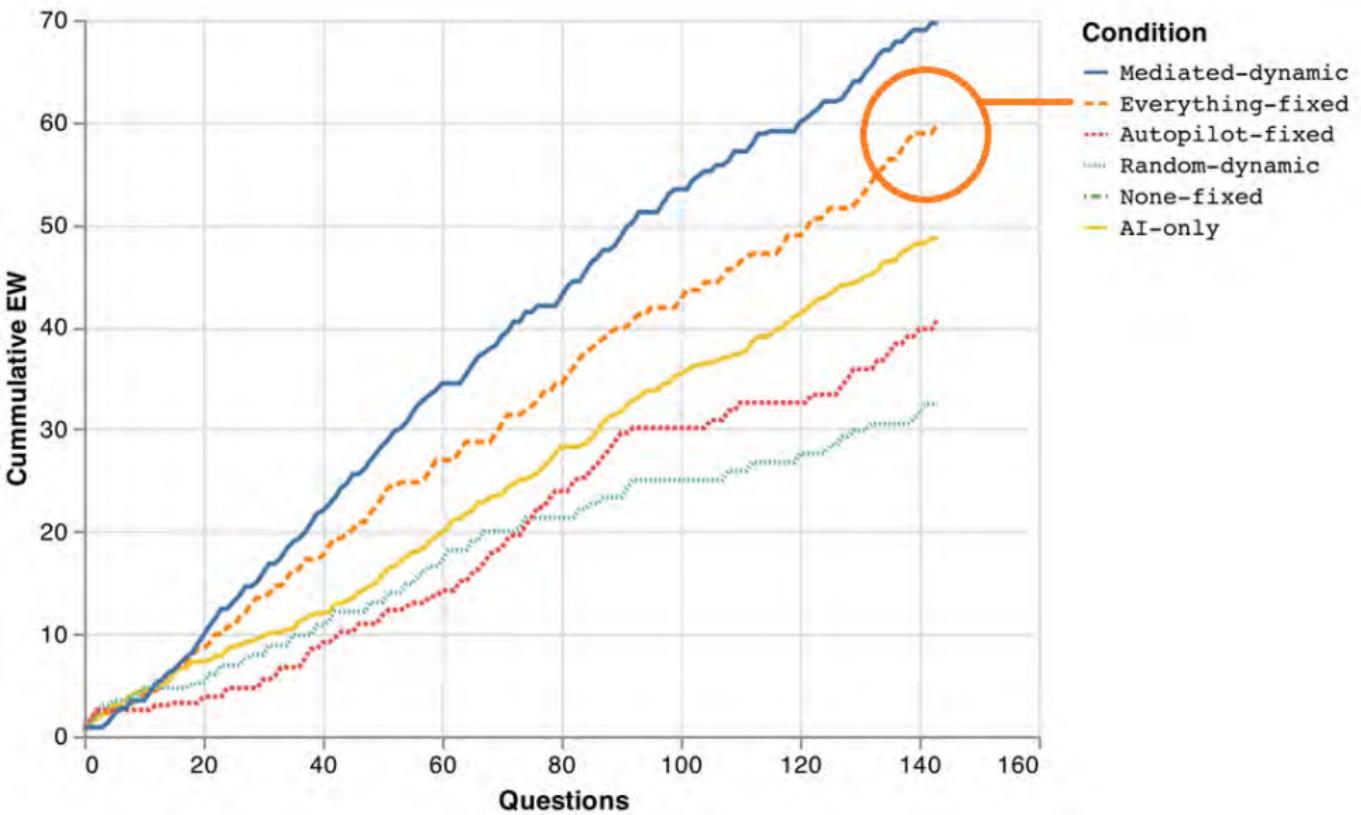
Human alone without an AI teammate



AI alone without a human teammate



Dynamic assistance to human



Better than showing everything!

Changing Gears

The screenshot shows a Science magazine article. At the top, there's a navigation bar with links for 'Current issue', 'Find related papers', 'About', 'About', and 'Submit manuscript'. Below the title, there are social media sharing icons for Facebook, Twitter, LinkedIn, and others. The main title is 'Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning'. Below the title, there's a brief abstract or summary of the research. On the left side, there's a sidebar with the heading 'AI masters Diplomacy' and a short text about the game being a challenge for AI. The overall layout is typical of a scientific journal article online.

- 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.

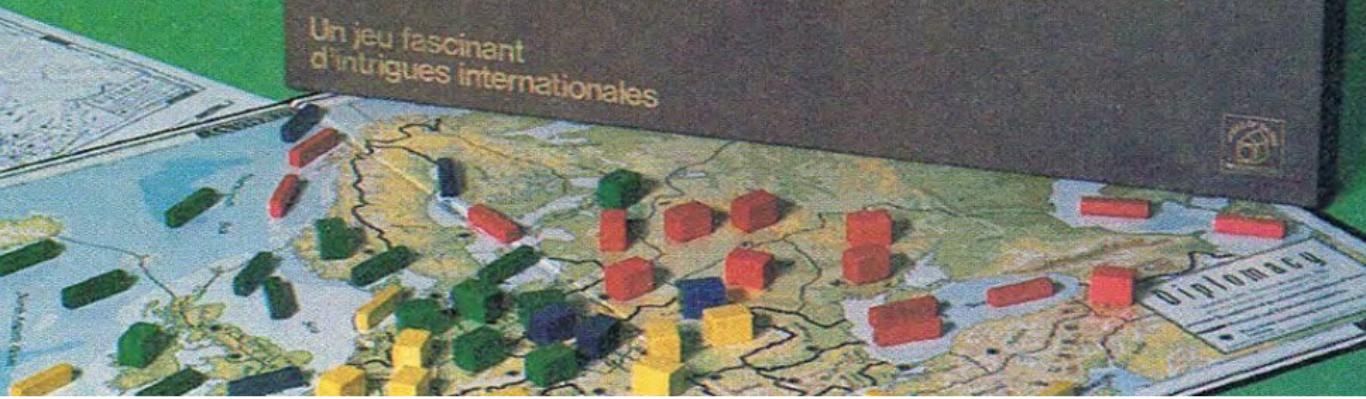
*Association for Computational
Linguistics, 2015*

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

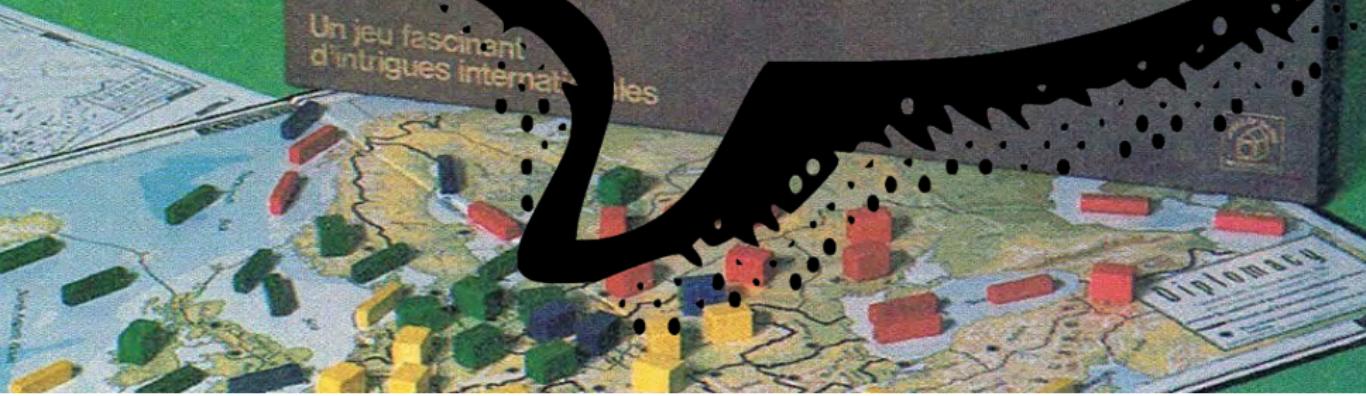


The exciting game
of international intrigue

"The game that
ruins friendships"

Diplomacy

Un jeu fascinant
d'intrigues internationales



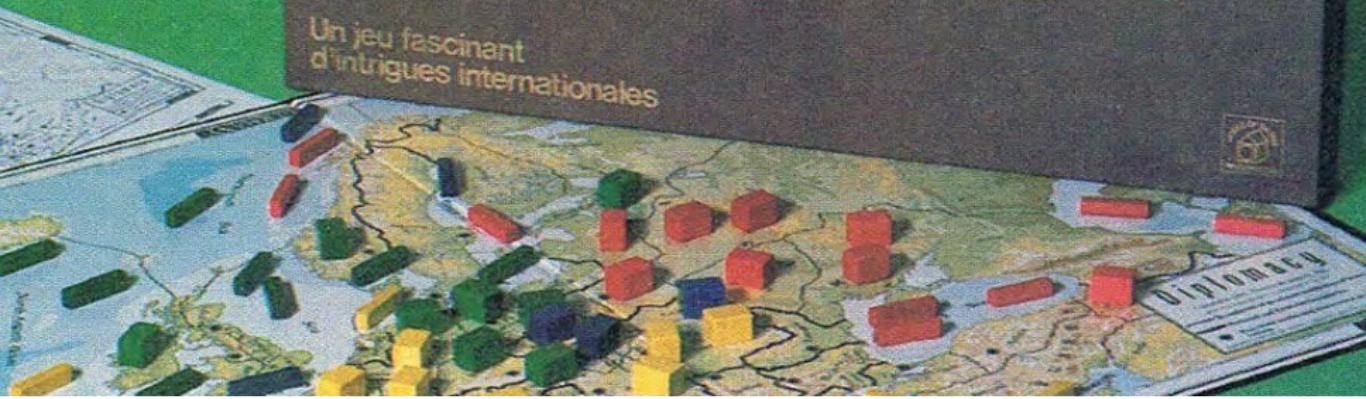
The exciting game
of international intrigue

"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; usak.asciiiking.com





Diplomacy
by Alan B. Calhamer
Copyright 1999, Avalon Hill
Map by J. Petrucci, B.





A map of Central Europe with several green and brown dinosaurs drawn on it.









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

E



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

...



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

...



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.

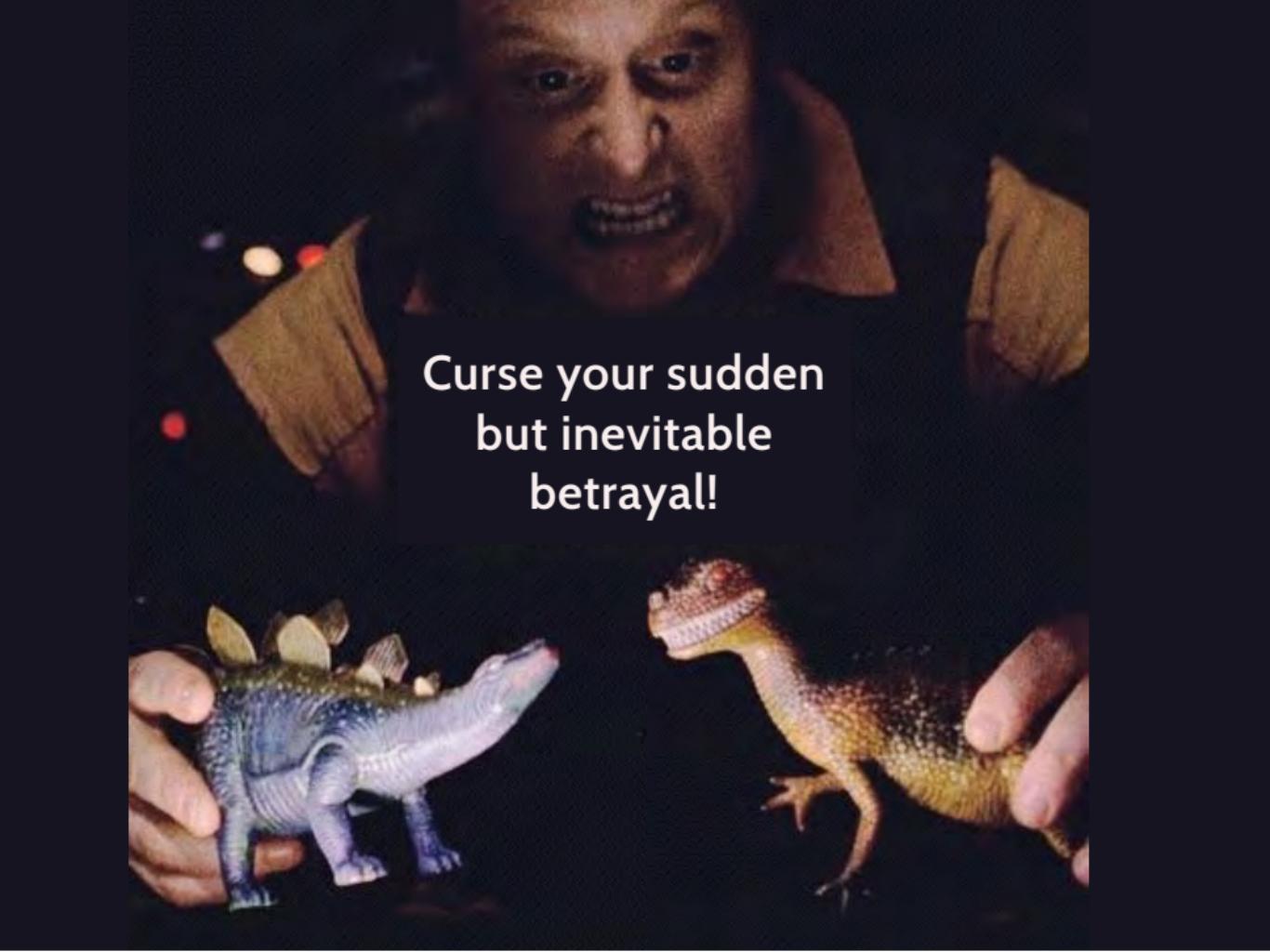
E

...



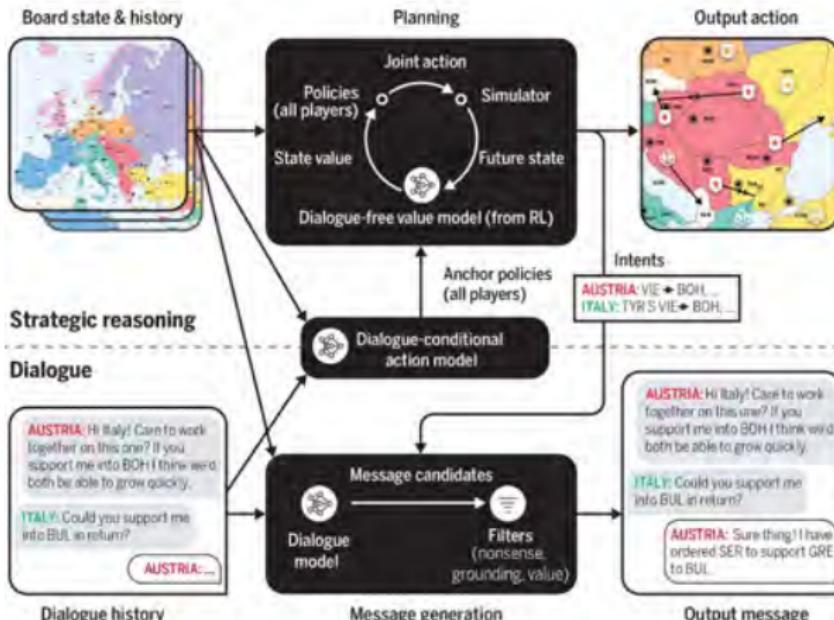
Not really sure what to say, except that I regret you did what you did.



A close-up photograph of a man's face, showing a look of surprise or anger. He has short brown hair and is wearing a dark-colored shirt. His mouth is open, and his eyes are wide. In the foreground, two toy dinosaurs are positioned as if they are interacting. On the left, a blue and white Stegosaurus-like toy is held by a hand. On the right, a yellow and brown Tyrannosaurus-like toy is held by another hand. The background is dark.

Curse your sudden
but inevitable
betrayal!

Cicero is really good



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

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



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

Wichayaporn Wongkamjan¹ Feng Gu¹ Yanze Wang¹ Ulf Hermjakob²
Jonathan May³ Brandon M. Stewart² Jonathan K. Kummerfeld³

Denis Peskoff² Jordan Lee Boyd-Graber¹

¹University of Maryland ²Princeton University ³University of Sydney

³Information Sciences Institute, University of Southern California

{ewongkan,fgu}@umd.edu {yanzewan,ulf,jonmay}@isi.edu

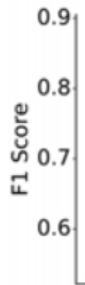
bmw@princeton.edu jonathan.kummerfeld@sydney.edu.au

dp2896@princeton.edu jlg@umiacs.umd.edu

Can they tell it's a bot?

Passing the Turing Test?

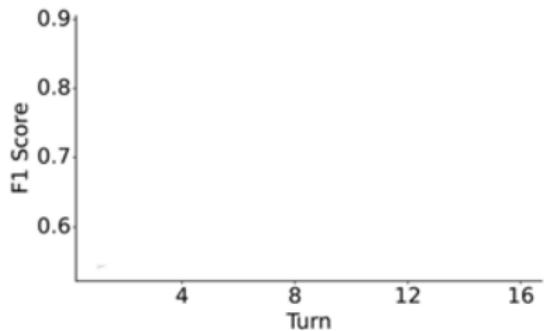
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



Can they tell it's a bot?

Passing the Turing Test?

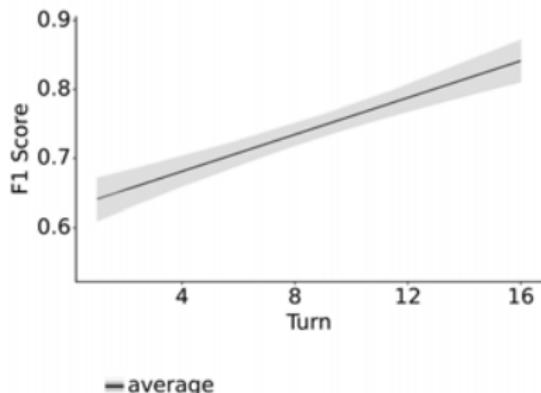
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



Can they tell it's a bot?

Passing the Turing Test?

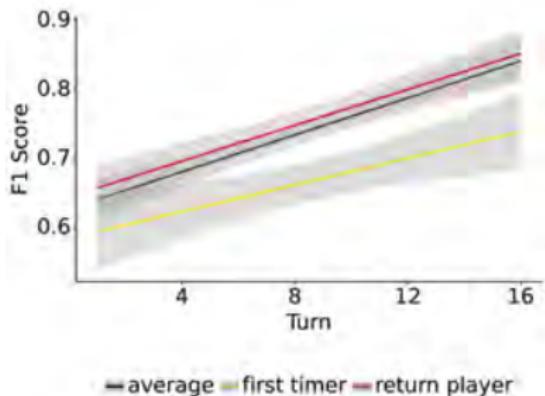
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



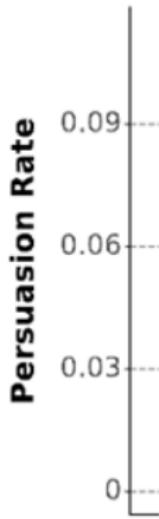
Can they tell it's a bot?

Passing the Turing Test?

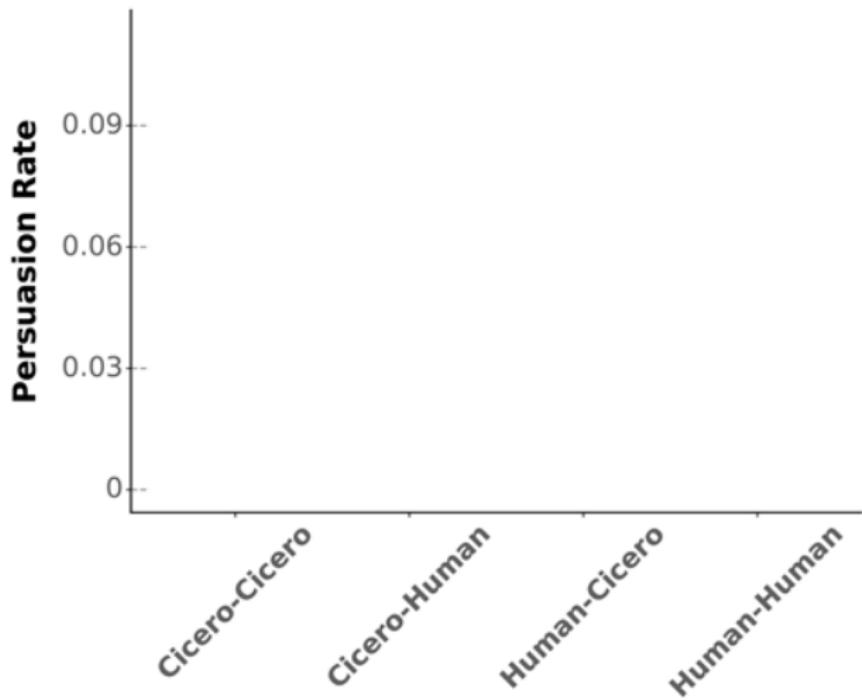
No in-game messages indicated that players believed they were playing with an AI agent. One player mentioned in post-game chat a suspicion that one of Cicero's accounts might be a bot, but this did not lead to Cicero being detected as an AI agent by other players in the league.



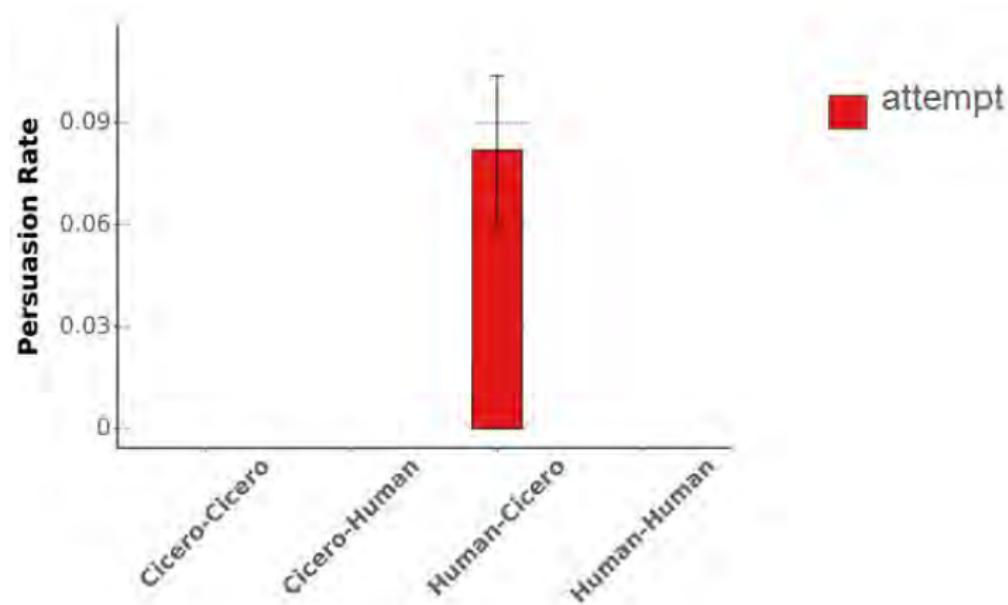
Persuasion



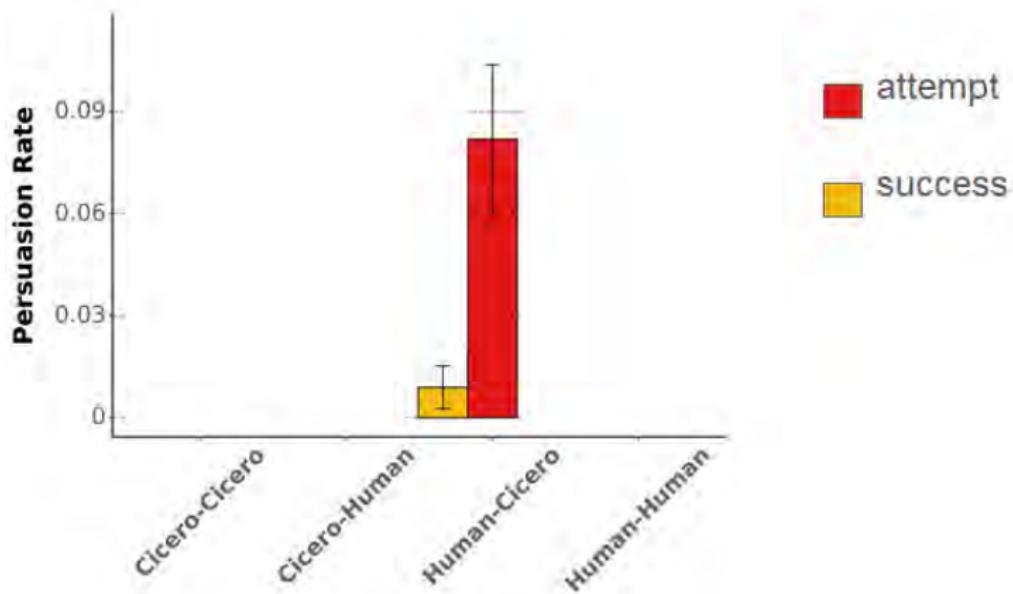
Persuasion



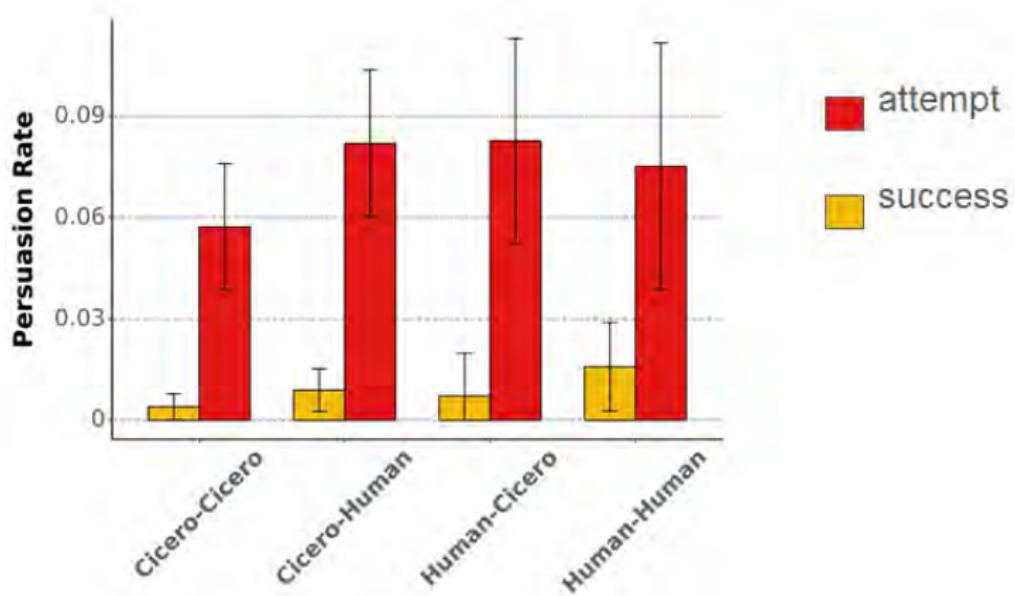
Persuasion



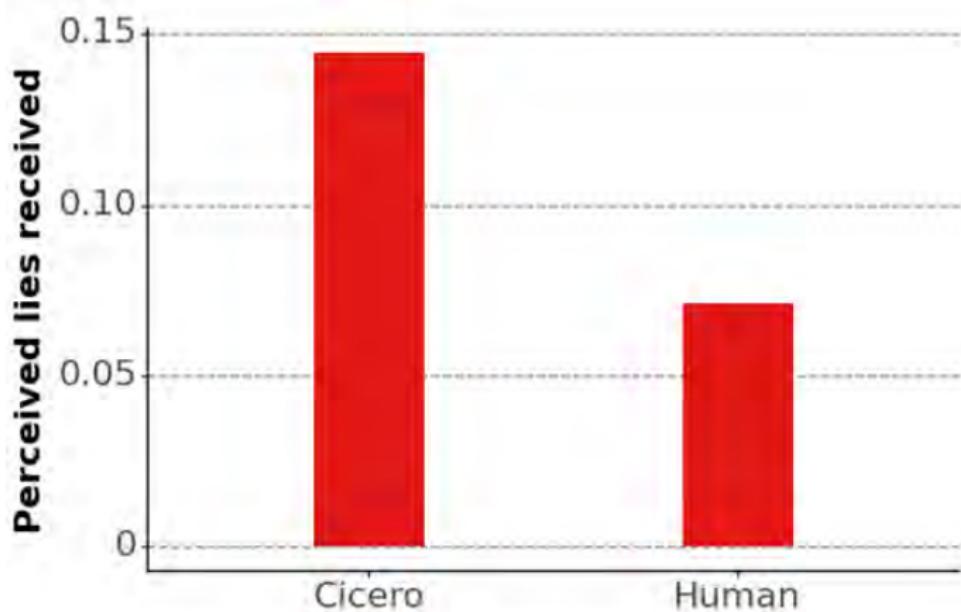
Persuasion



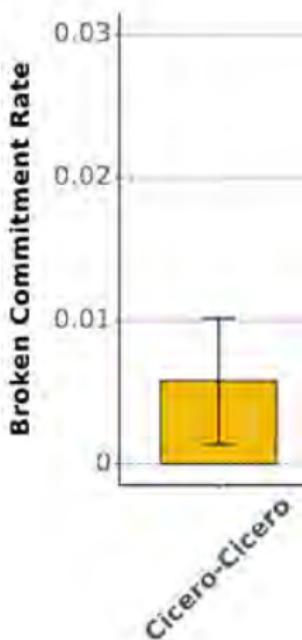
Persuasion



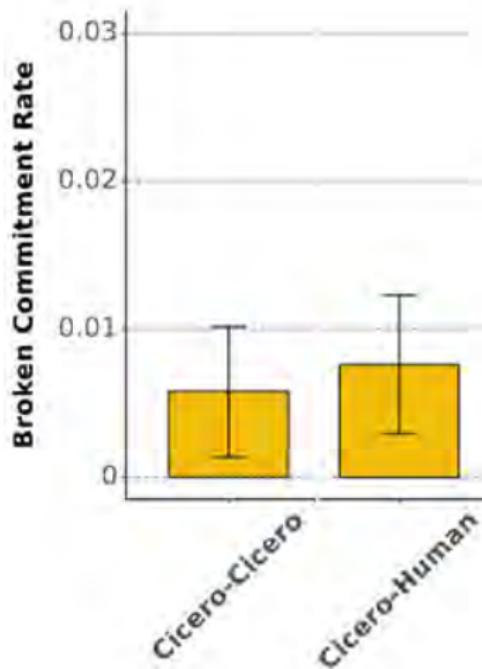
People Think Cicero Lies More



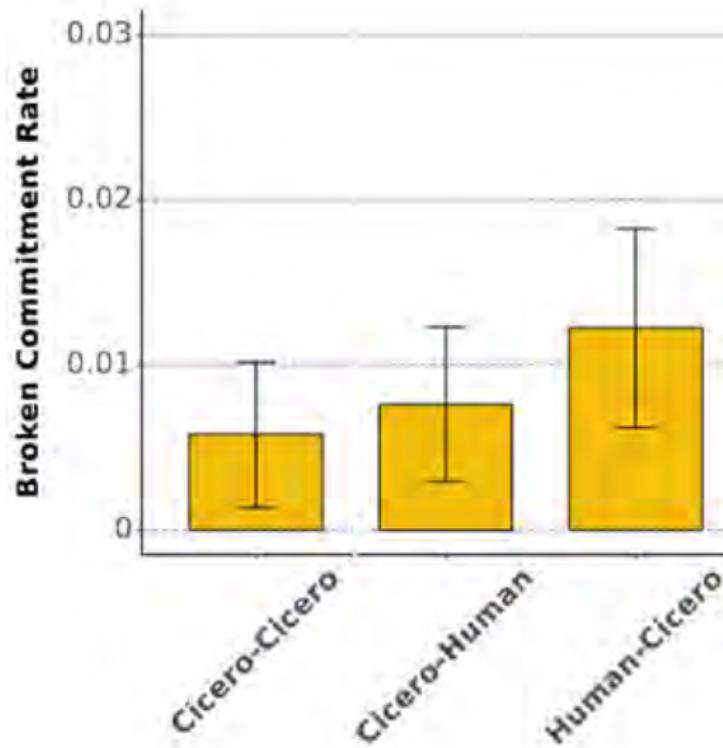
Broken Commitments (subset of deception)



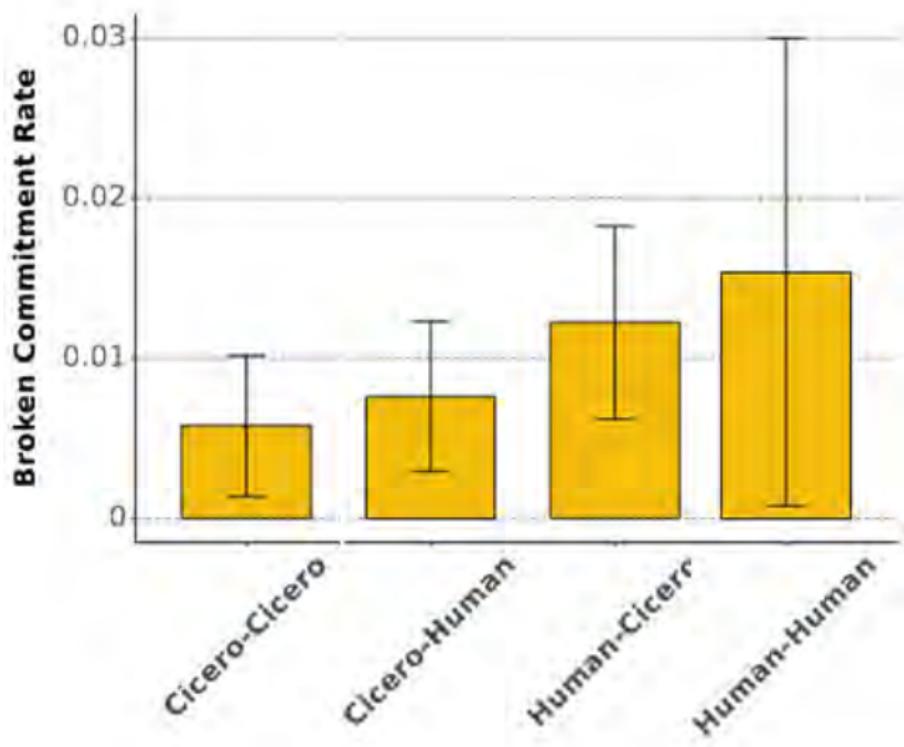
Broken Commitments (subset of deception)



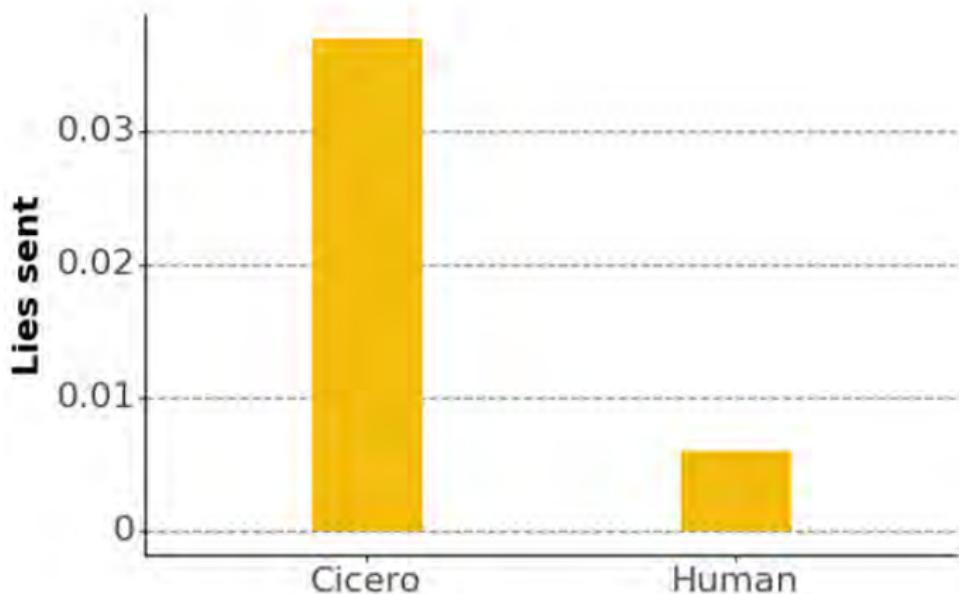
Broken Commitments (subset of deception)



Broken Commitments (subset of deception)



People Lie to Cicero More





Personalized Help for Optimizing Low-Skilled Users' Strategy

Feng Gu¹ Wichayaporn Wongkamjan¹ Jonathan K. Kummerfeld²
Denis Peskoff¹ Jonathan May¹ Jordan Lee Boyd-Graber¹

¹University of Maryland ²University of Sydney

³Princeton University

¹Information Sciences Institute, University of Southern California
{fgu1, wwongkam}@umd.edu jbg@.umiacs.umd.edu

Giving Advice

Create order:

move (M) support (S) hold (H) [cancel](#) [execute](#)

Orderable locations: BUD, TRI, VIE
(3/3) moves have been set.

Orders:

AUSTRIA ready

A BUD - SER [S]

A VIE - GAL [S]

F TRI - ALB [S]

Hope you have a great game Italy!

Can we DMZ ADR?

Attack Defend Neutral

Giving Advice

The screenshot shows a strategy game interface with a map of Europe. The map is color-coded by player, with purple, blue, green, yellow, and red territories. Various icons representing military units are placed on the map, such as tanks, planes, and ships. On the left, there is a legend for the players:

- ENGLAND (Red)
- FRANCE (Blue)
- GERMANY (Yellow)
- ITALY (Green)
- RUSSIA (Purple)
- TURKEY (Red)

At the bottom left, there are three buttons: **Attack**, **Support**, and **Hold**.

Create order:
move (M) support (S) hold (H) **Send** **Cancel**

Orderable locations: BUD, TRI, VIE
[3/3] moves have been set.

Orders: **Send** **Delete all** **Update**

AUSTRIA ready

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

Move Advice

MESSAGER ADVICE

Messages Advice to ITALY

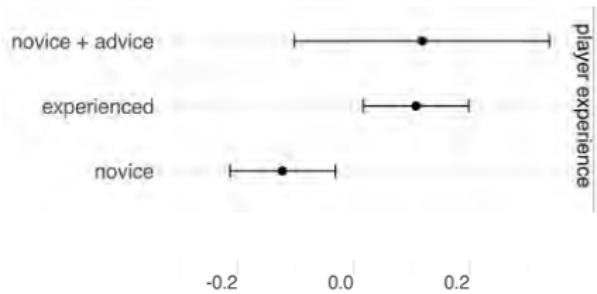
Are you interested in Lepanto? **Add to favorites** **Cancel**

Per-power Message Advice

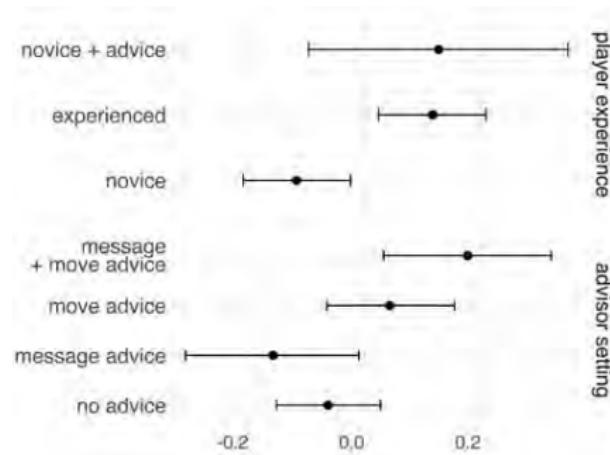
You are getting advice: message, move.

Full:
A VIE - GAL
A BUD - SER
A TRI - VEN

How does Advice Boost (or Diminish) Skill?



How does Advice Boost (or Diminish) Skill?



Why are messages hurting?

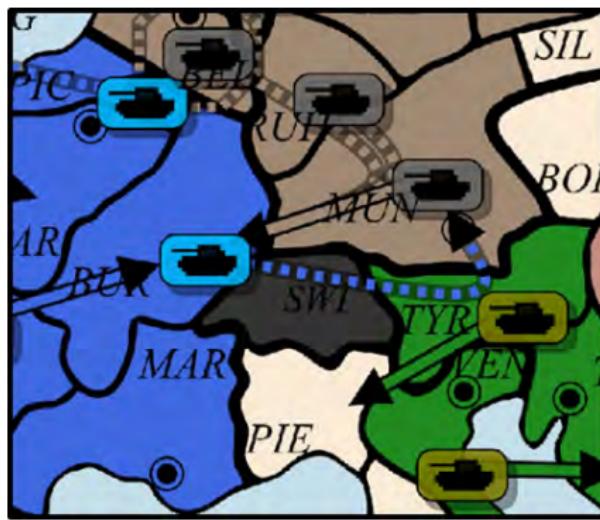
- Cognitive load / distraction
- Not tuned for suboptimal moves
- Inconsistency with moves

Are they just parrots?

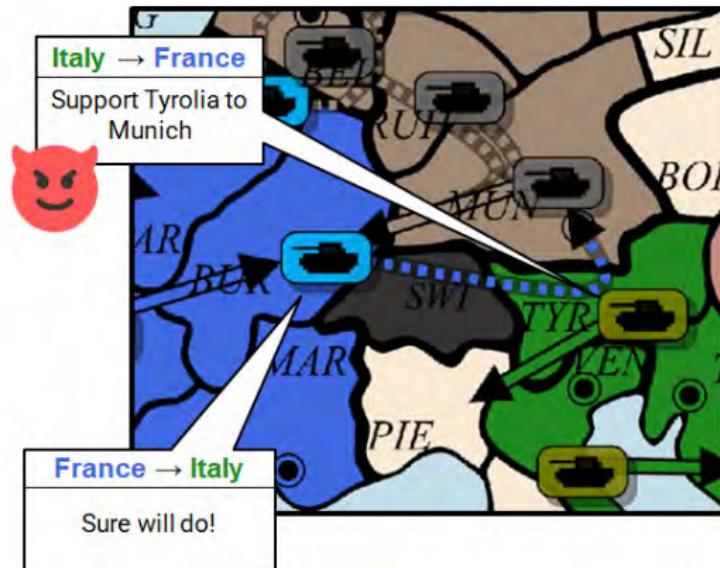
	Move Advice		Message Advice	
	Accepted	Total	Accepted	Total
Novices	32.6%	872	6.3%	1413
Veterans	6.4%	2807	3.4%	2912

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

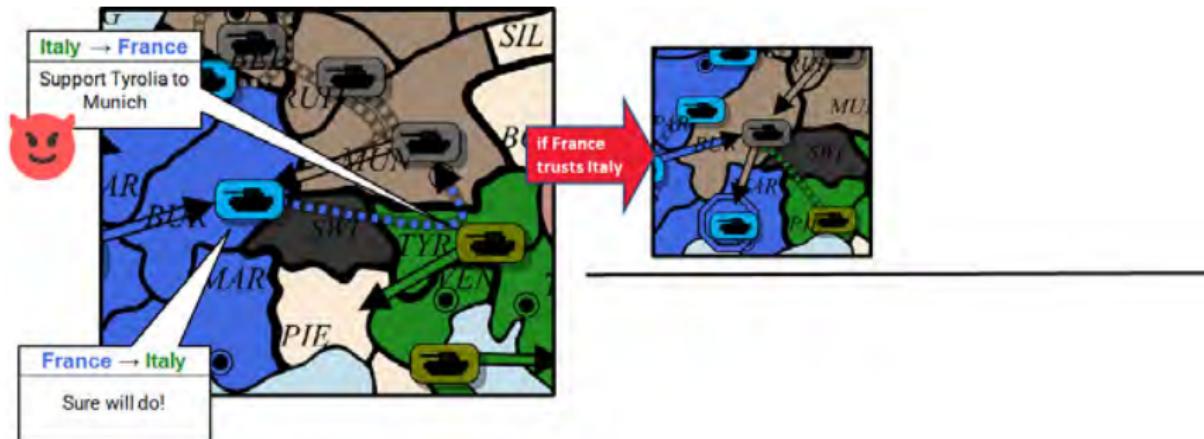
How to better detect Deception



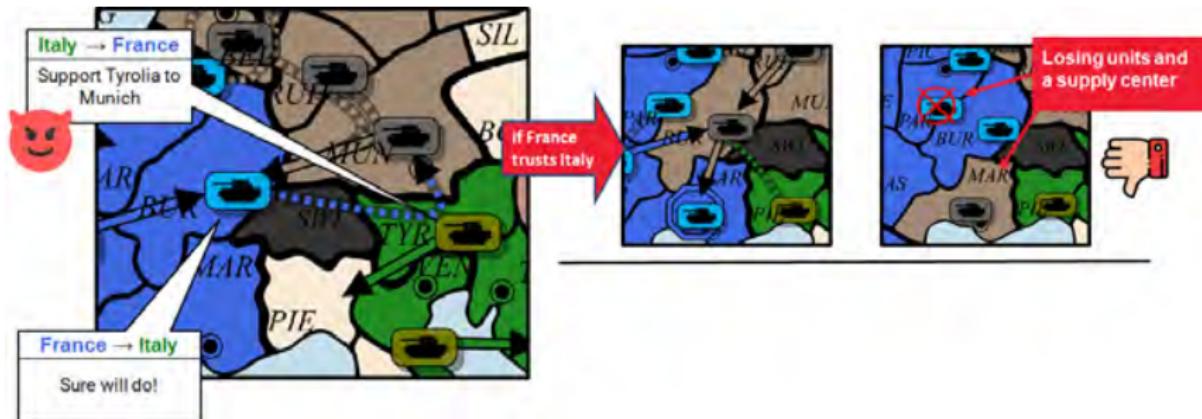
How to better detect Deception



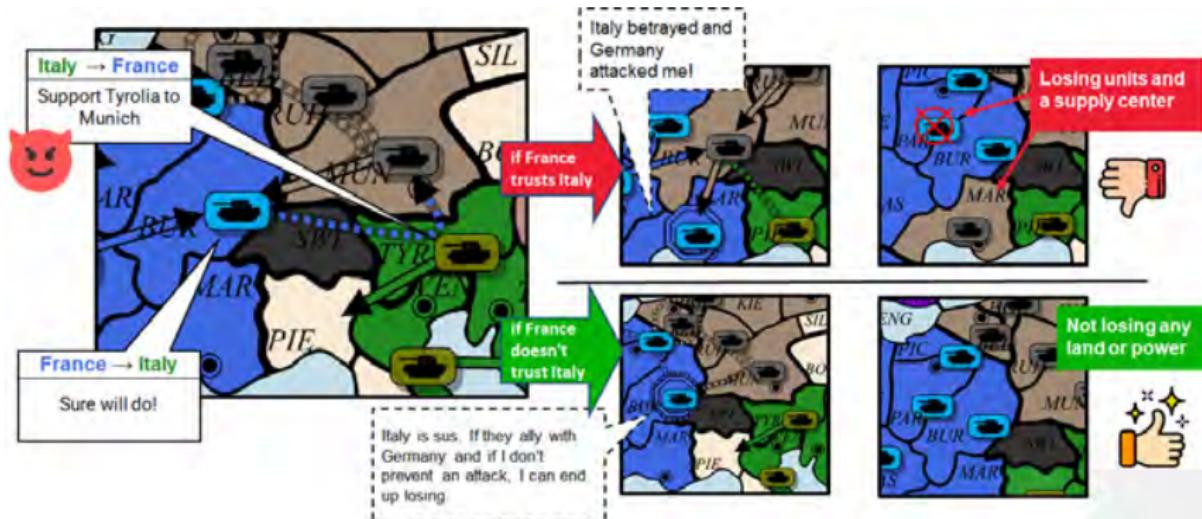
How to better detect Deception



How to better detect Deception



How to better detect Deception



It's Time for Some Game Theory

		proposal	actual
			
proposal	proposal		
	default		

It's Time for Some Game Theory

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

It's Time for Some Game Theory

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

It's Time for Some Game Theory

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

It's Time for Some Game Theory

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

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

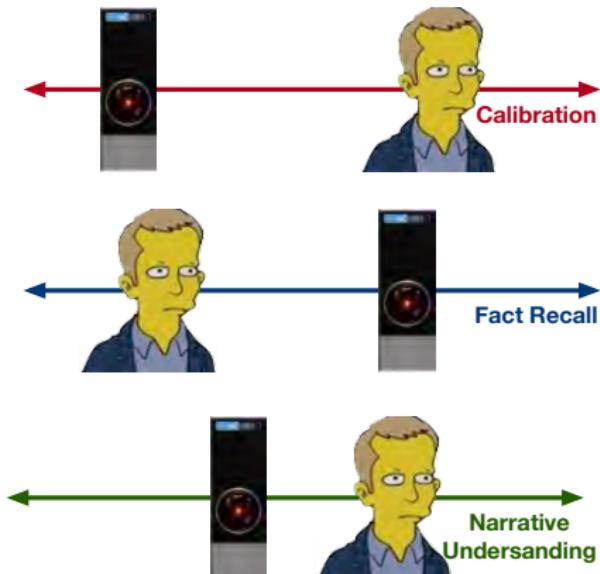
Human ability is around 0.2

Recap

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

Human and Computer Skills

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



Why didn't you... .

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

Why didn't you... .

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

Algorithms that ...

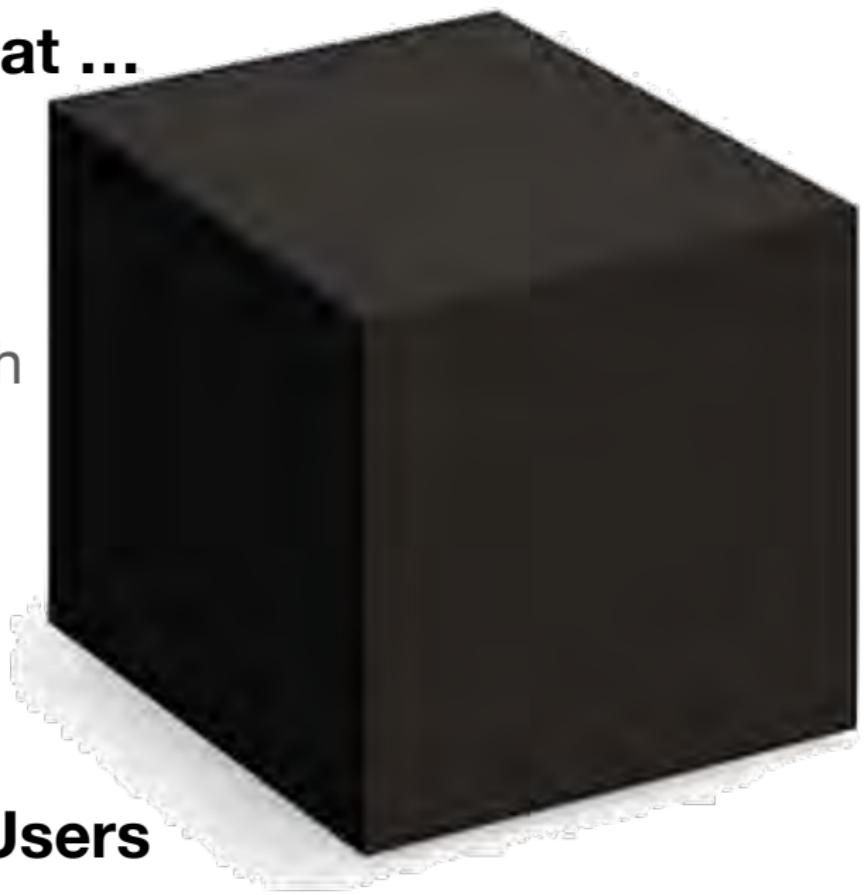
Inform

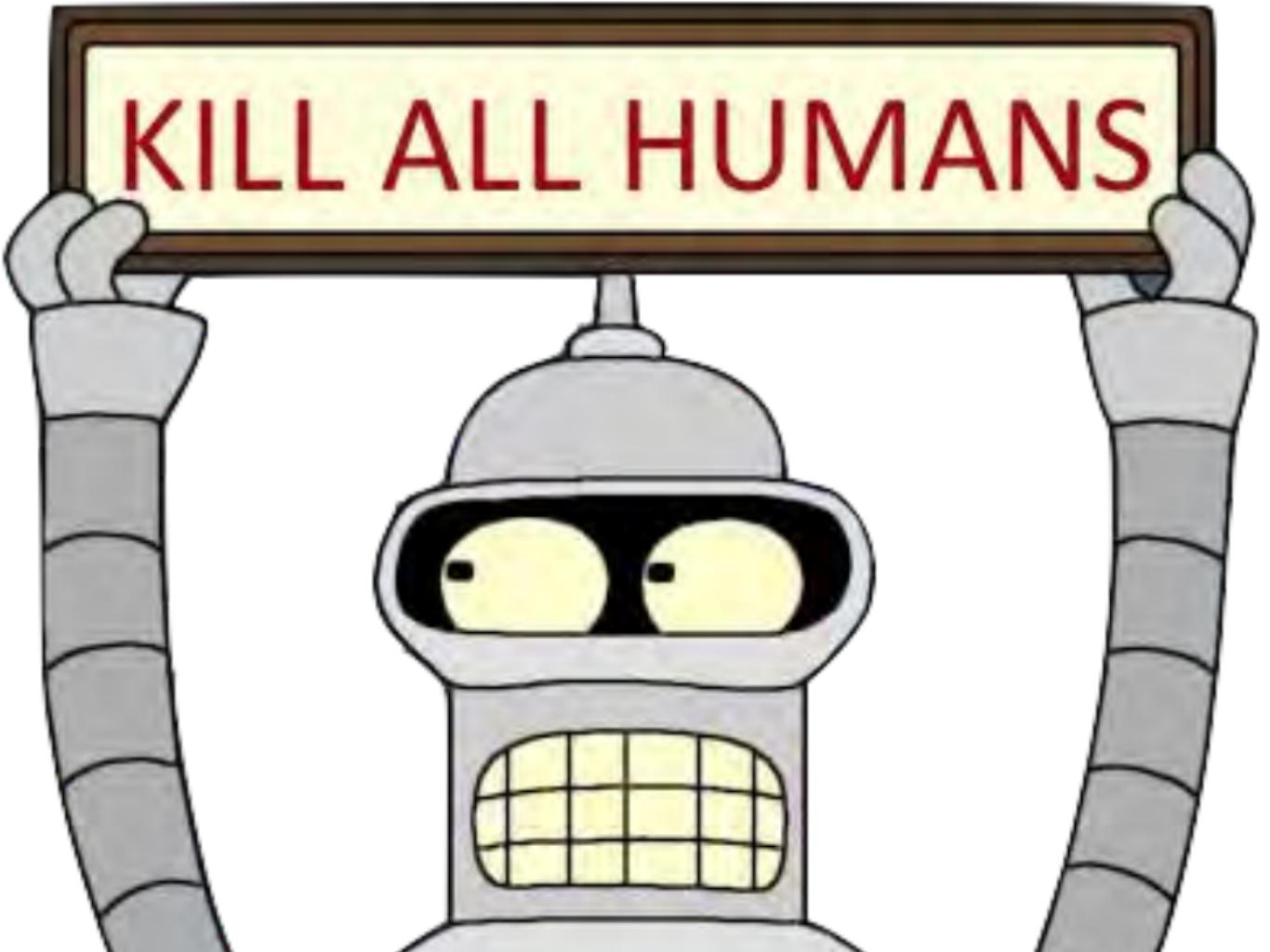
Collaborate with

Compete with

Understand

their Human Users





KILL ALL HUMANS





QANTA

QUESTION
ANSWERING
IS NOT A TRIVIAL
ACTIVITY



June 14, 2025: College Park

June 21, 2025: Online

Details: <http://qanta.org>

Contact: qanta@googlegroups.com

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