



# Language Games

Jordan Boyd-Graber  
University of Maryland  
2023

## Why Language Games?



- Wittgenstein: Language is a product of the context in which it is used
- Interaction allows us to figure out how to use language



Image: DeepMind



Peter Morgan/Reuters

## Today's Topics

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

# Fact Checking with Humans in the Loop

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

Chenglei Si<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)

# Fool Me Twice

Google Research

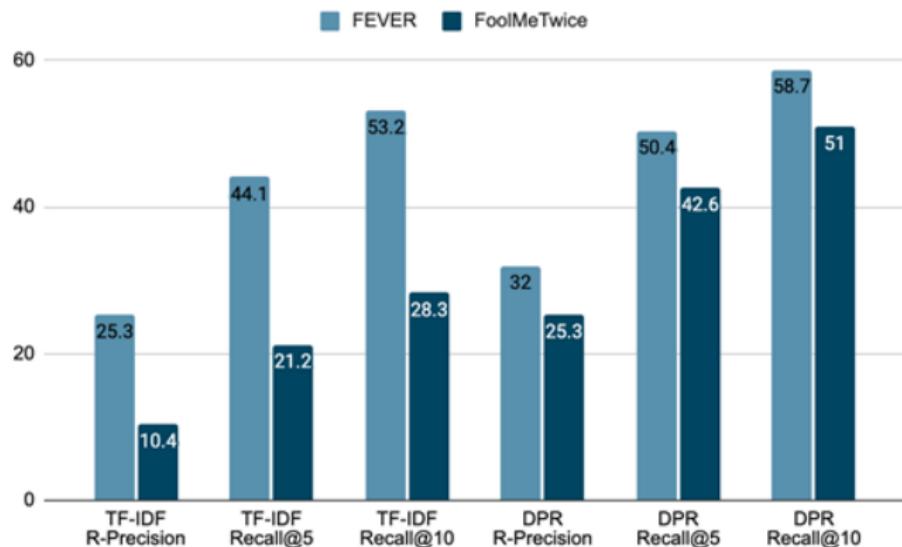
## Fool Me Twice Entailment from Wikipedia Gamification

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

Jannis Bösel   Jordan Boyd-Graber   Julian Eisenschlos



## It is Harder (a bit)



## Helping Humans Do the Task

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

## Helping Humans Do the Task

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

### ChatGPT Explanation:

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

## Helping Humans Do the Task

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

### ChatGPT Explanation:

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

### Retrieved Passage #1:

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

# Helping Humans Do the Task

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

## ChatGPT Explanation:

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

## Retrieved Passage #1:

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

## Contrastive Explanation - Supporting:

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

## Contrastive Explanation - Refuting:

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

## 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

## Mimicking How Humans Learn Calibration

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



## 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**

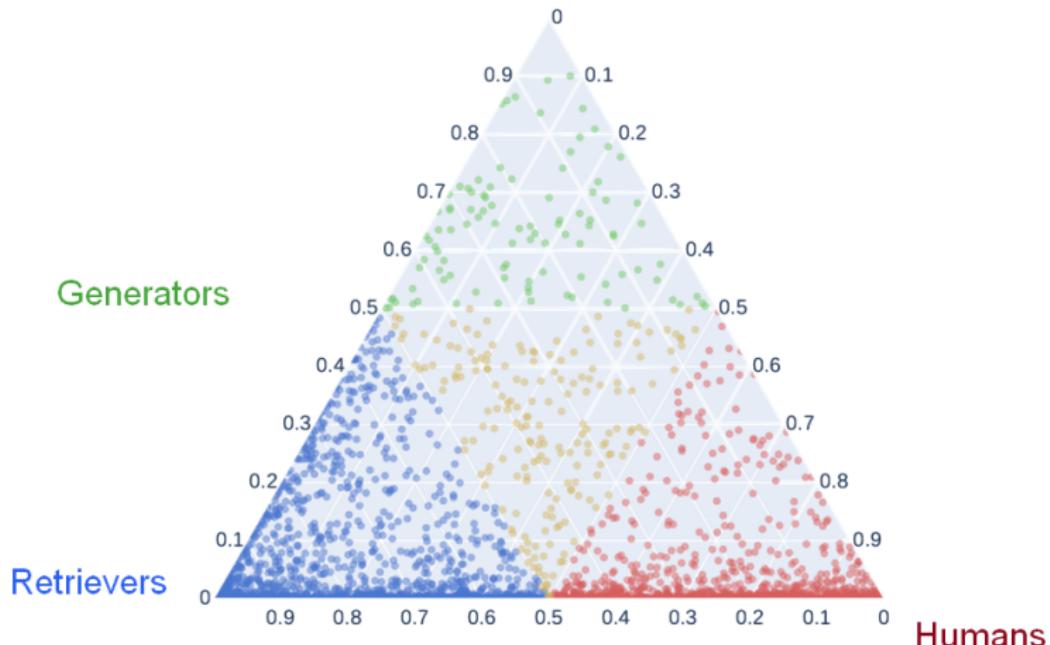
## Sample Question

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

Faster = Smarter

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

## What's hard for Computers



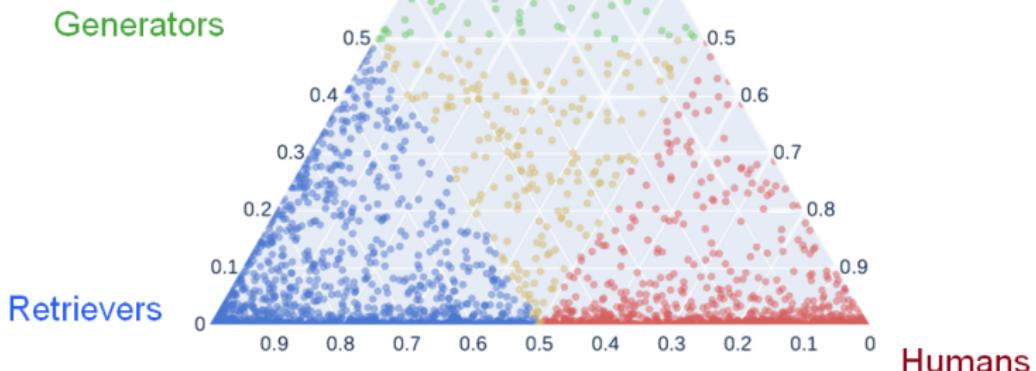
## What's hard for Computers

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

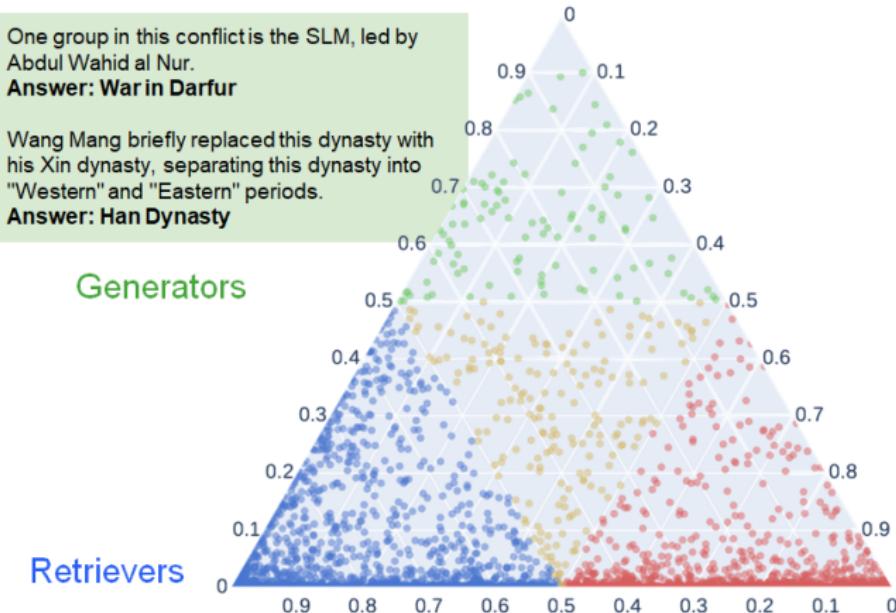
Humans

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**



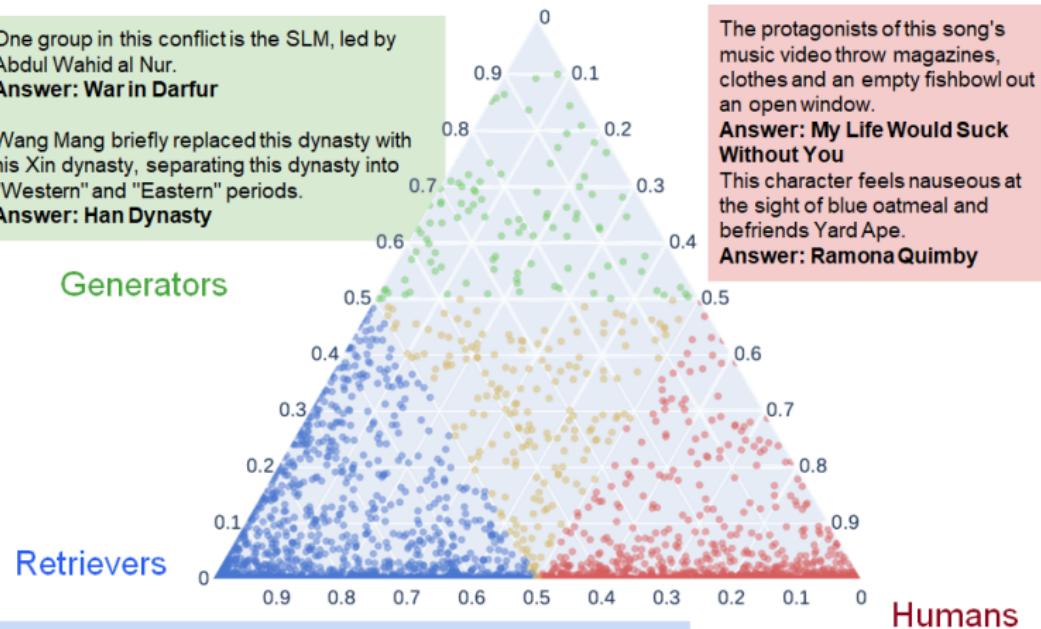
## What's hard for Computers

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**



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

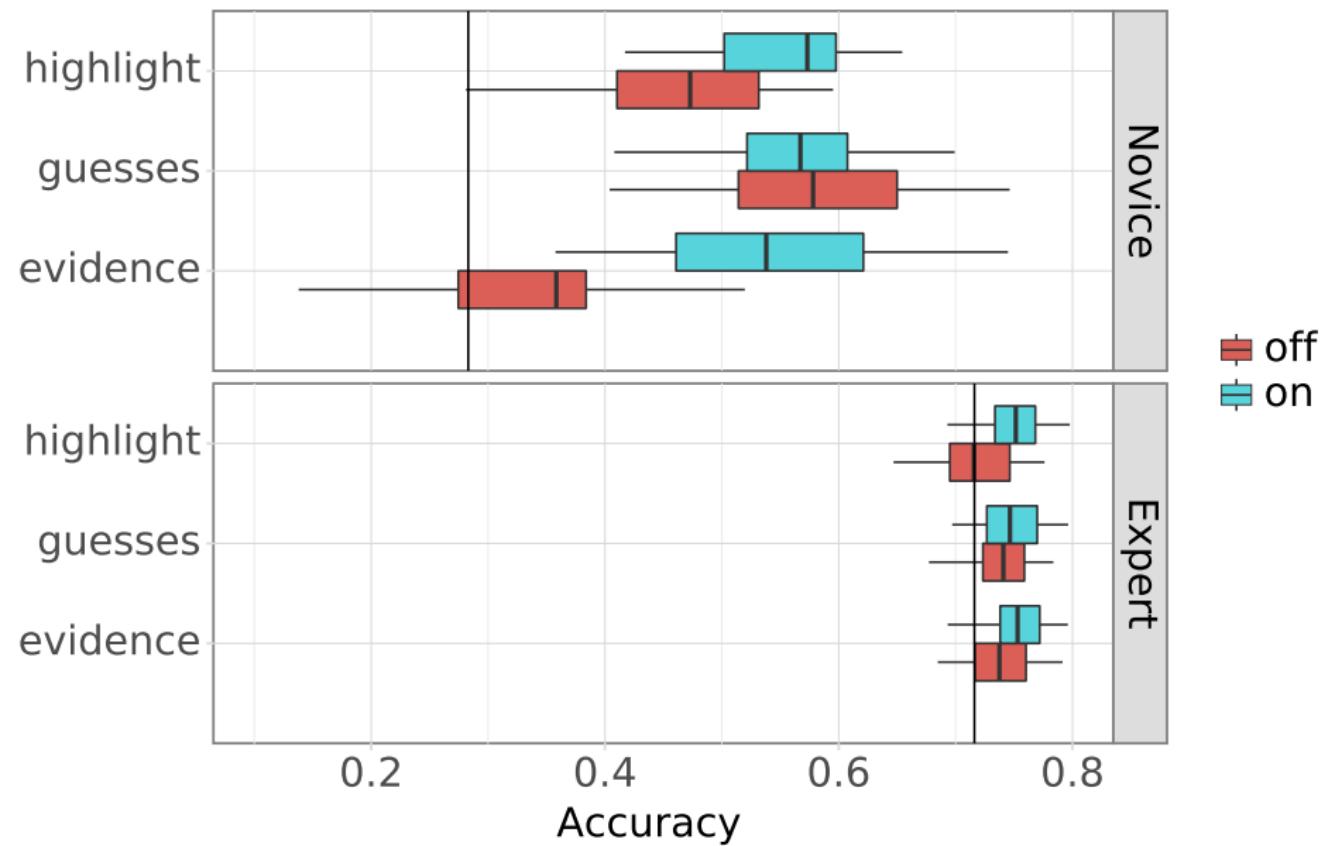
## Experts vs. Novices

### Experts

Trivia experts, familiar with task, enjoy the task

### Mechanical Turkers

Mechanical Turkers: easily overwhelmed, need the help



Evidence helps novices, experts are expert

## Regression Analysis

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

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

## Regression Analysis

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

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

Coefficients tell story!

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

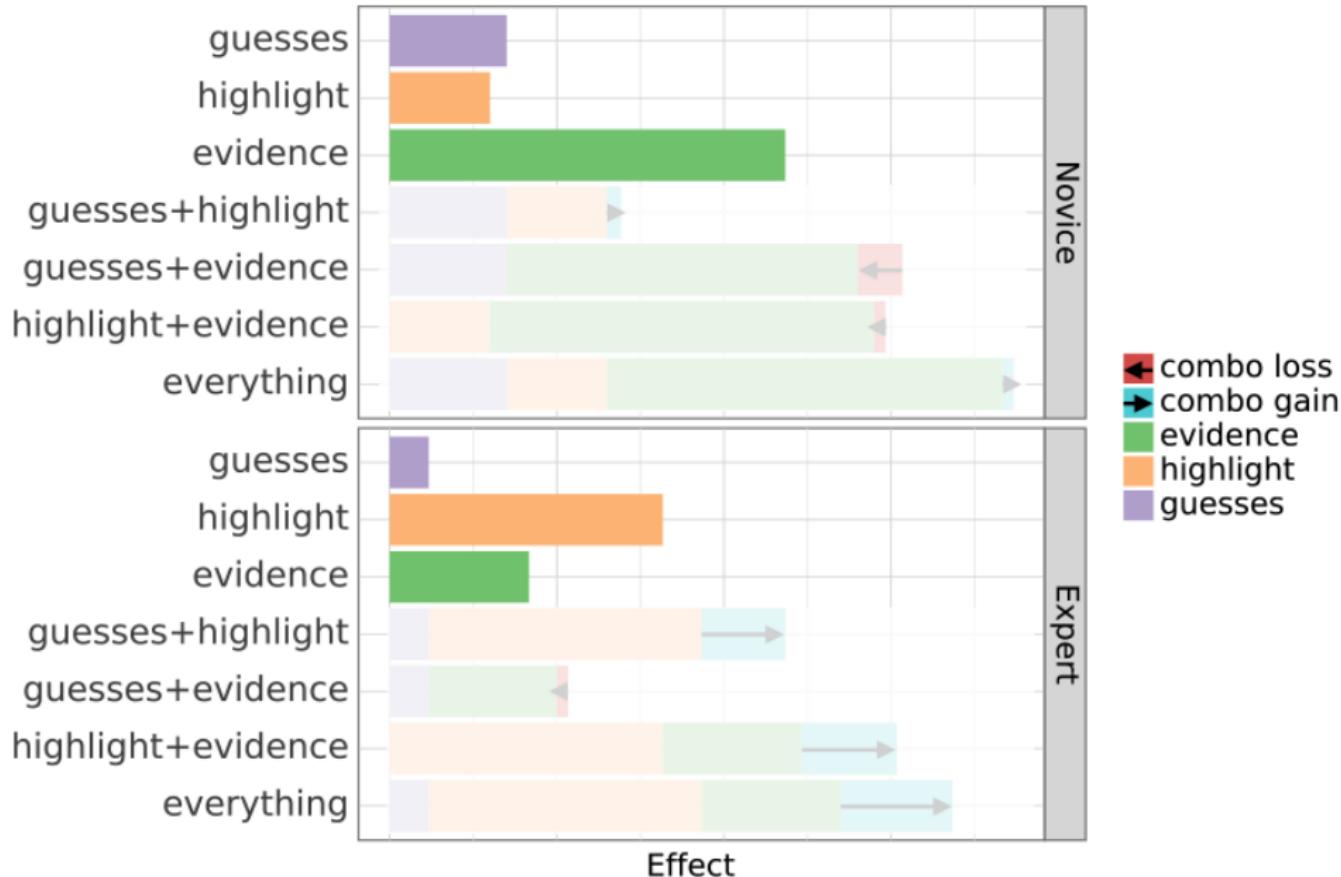
## Regression Analysis

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

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

Coefficients tell story!

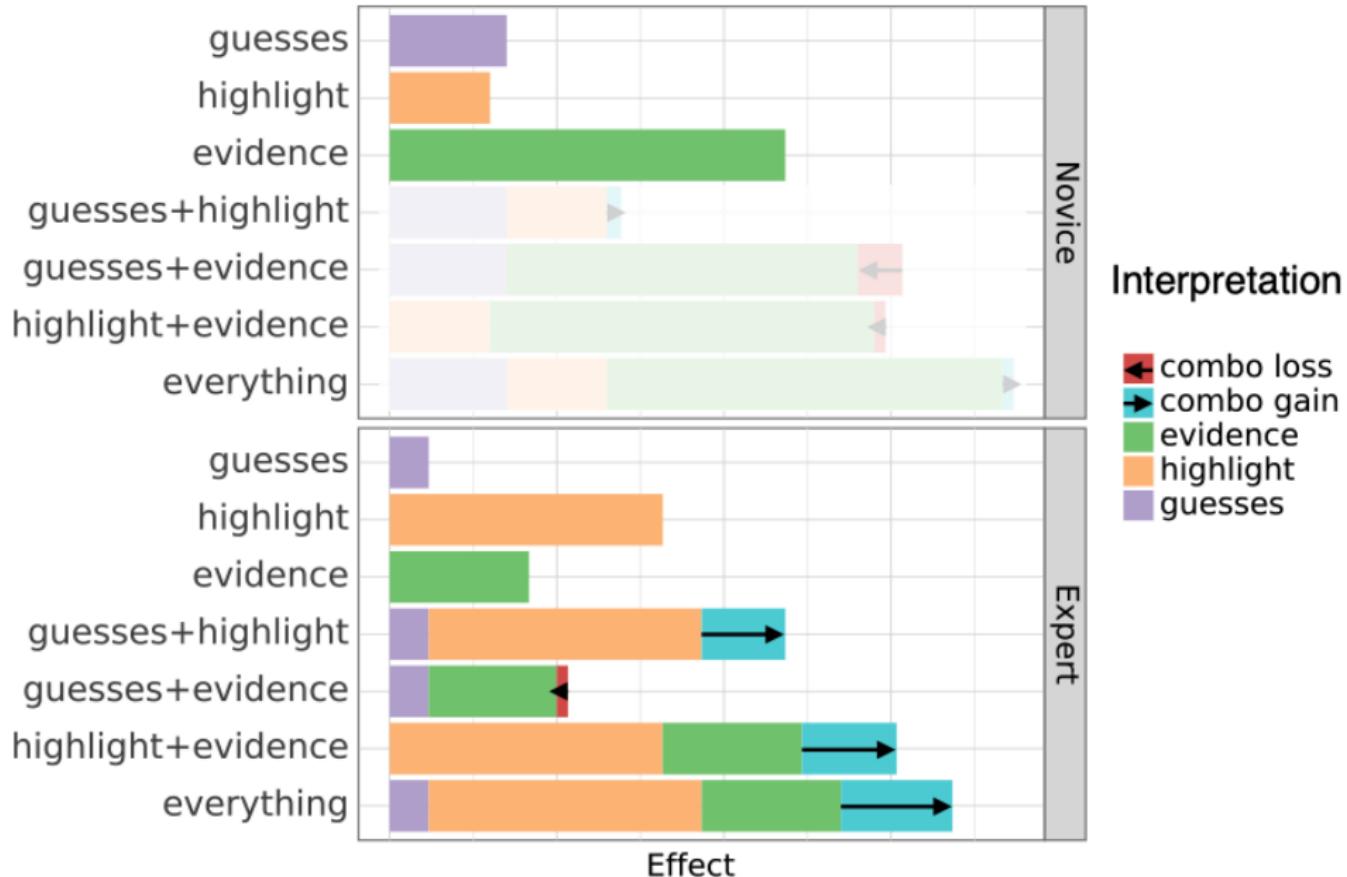
- **Big, Positive:** Help
- **Big, Negative:** Hurt
- **Small:** Neutral



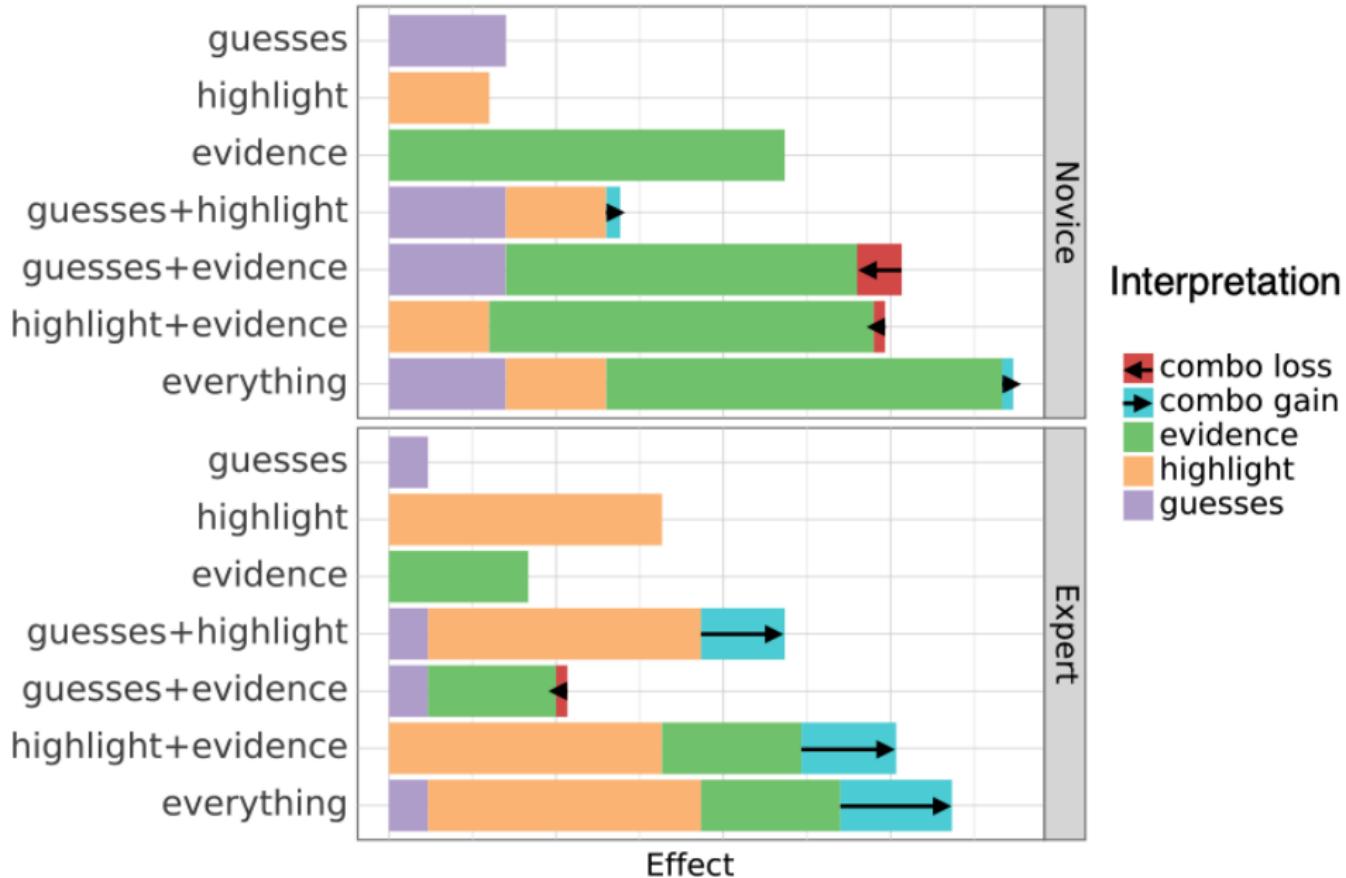
Everything helps: Evidence for novices, Highlight for experts



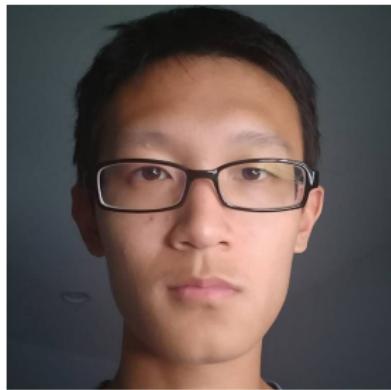
Synergistic effects



Highlight and evidence help experts most



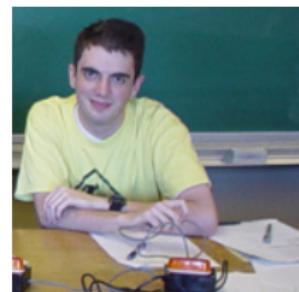
For novices, less synergy



Learning to Explain Selectively

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

## Measuring Interpretability

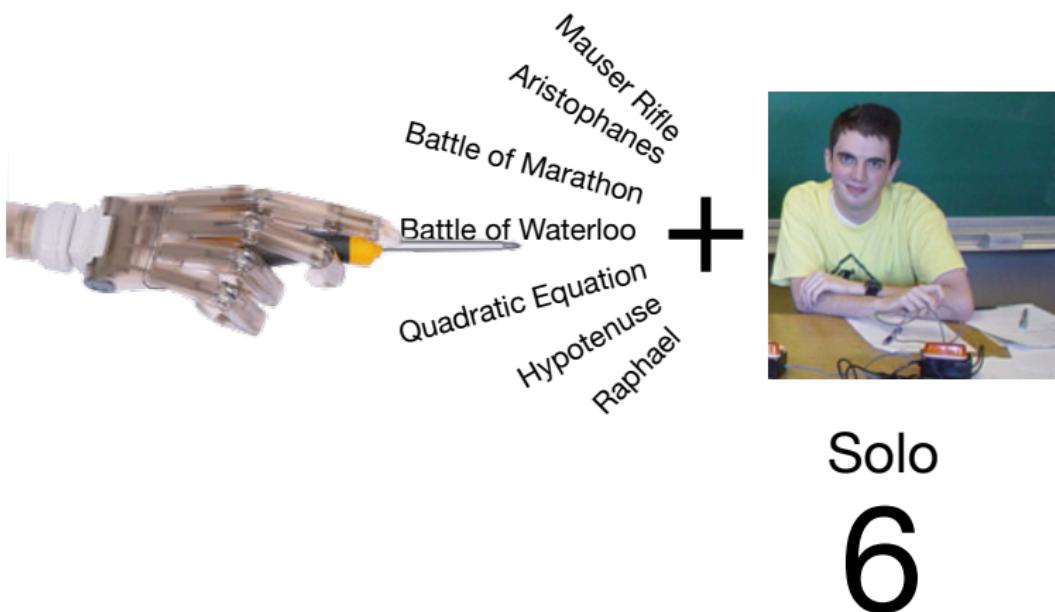


## Measuring Interpretability

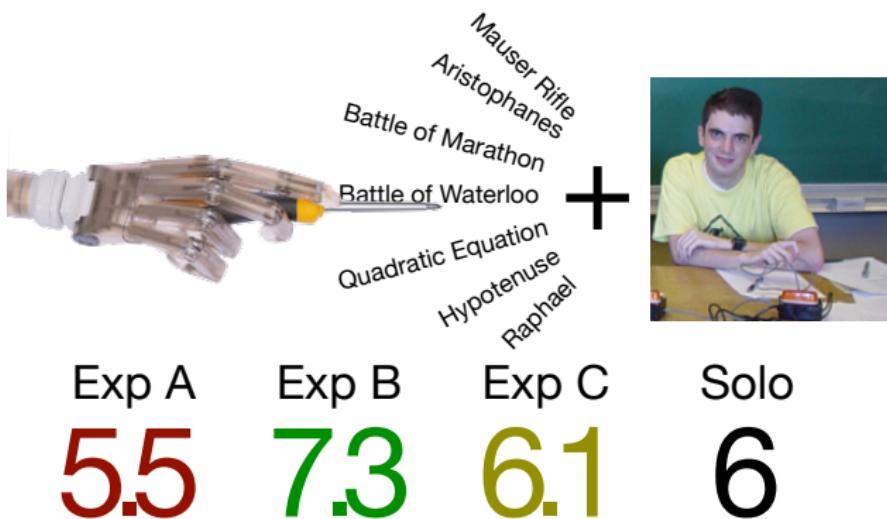


Solo  
**6**

## Measuring Interpretability

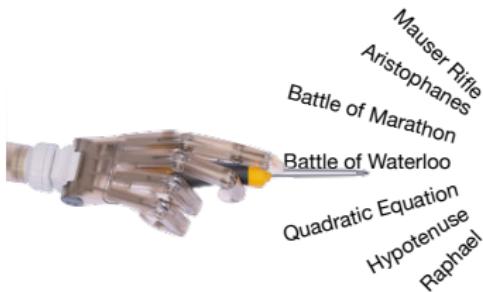


## Measuring Interpretability



## Improvement through Bandit Algorithms

### Visualization

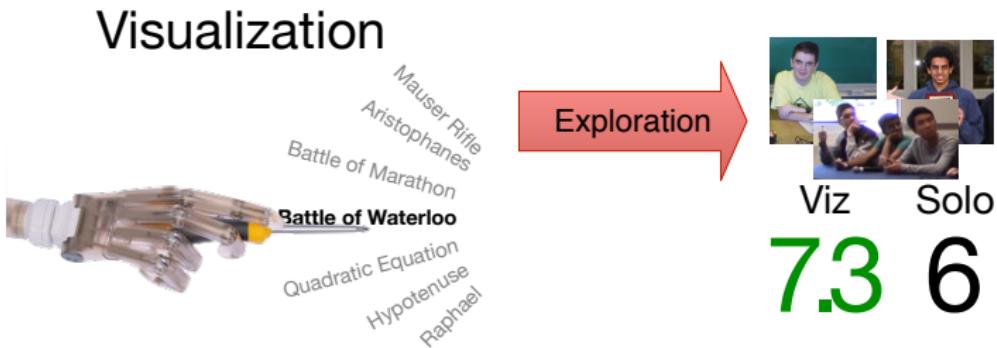


Viz Solo  
**7.3** **6**

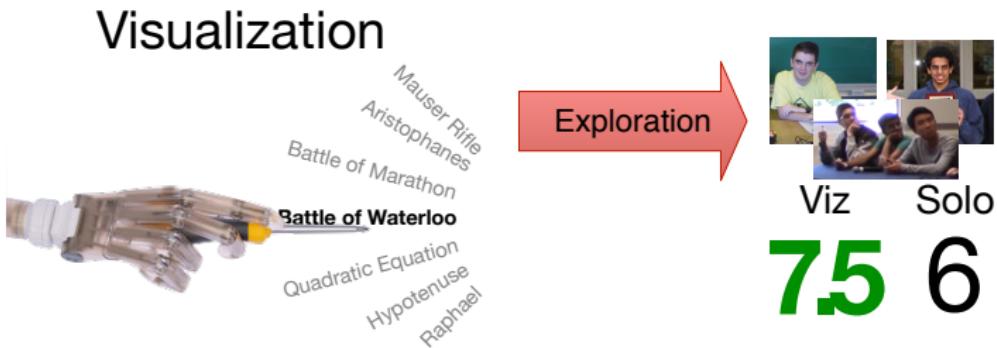
## Improvement through Bandit Algorithms



## Improvement through Bandit Algorithms



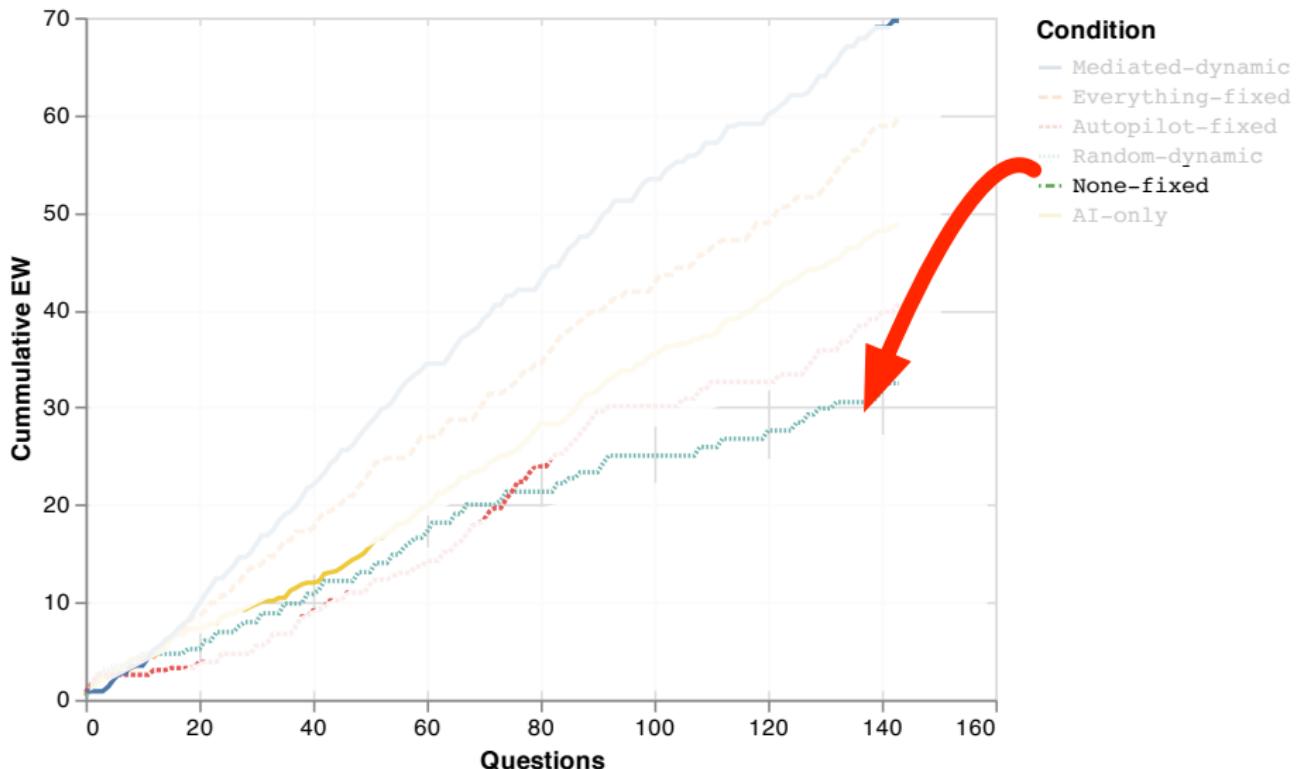
## Improvement through Bandit Algorithms



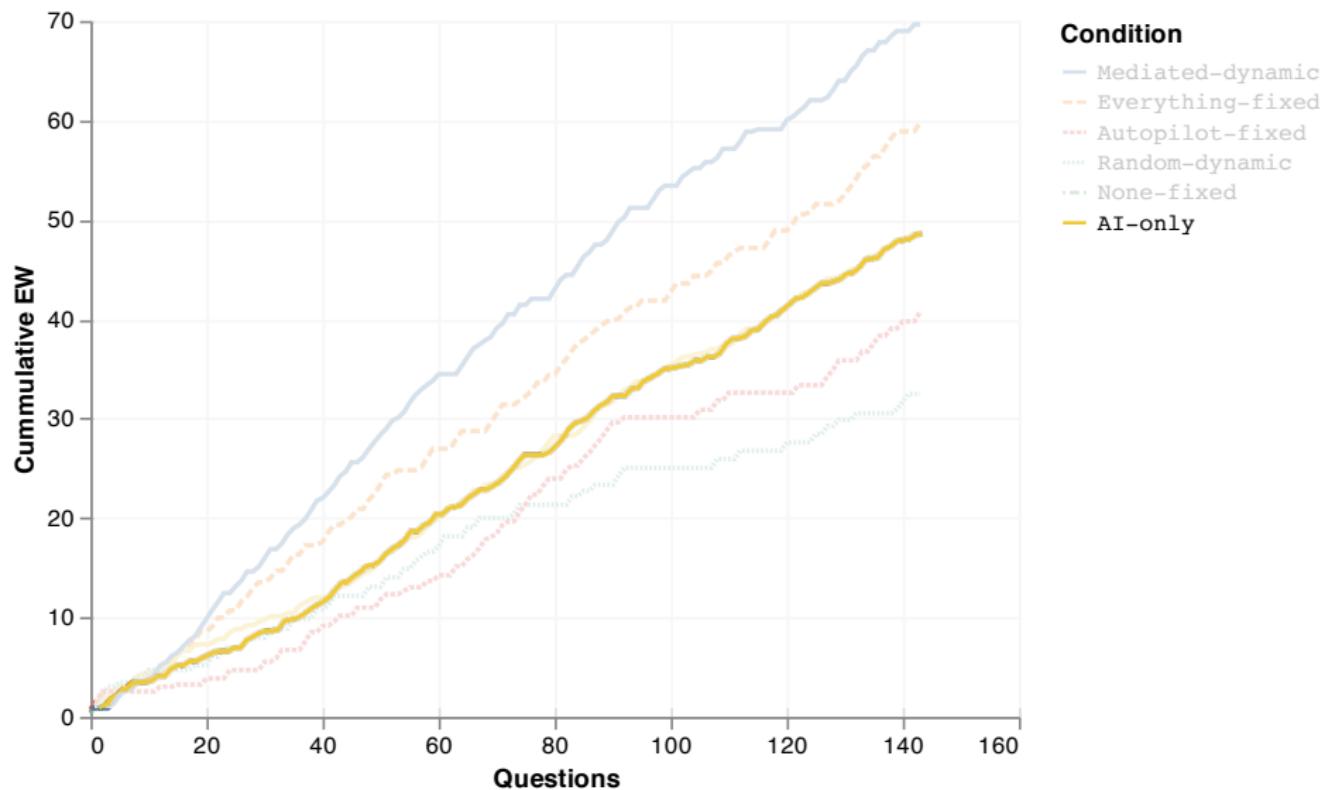
## Improvement through Bandit Algorithms



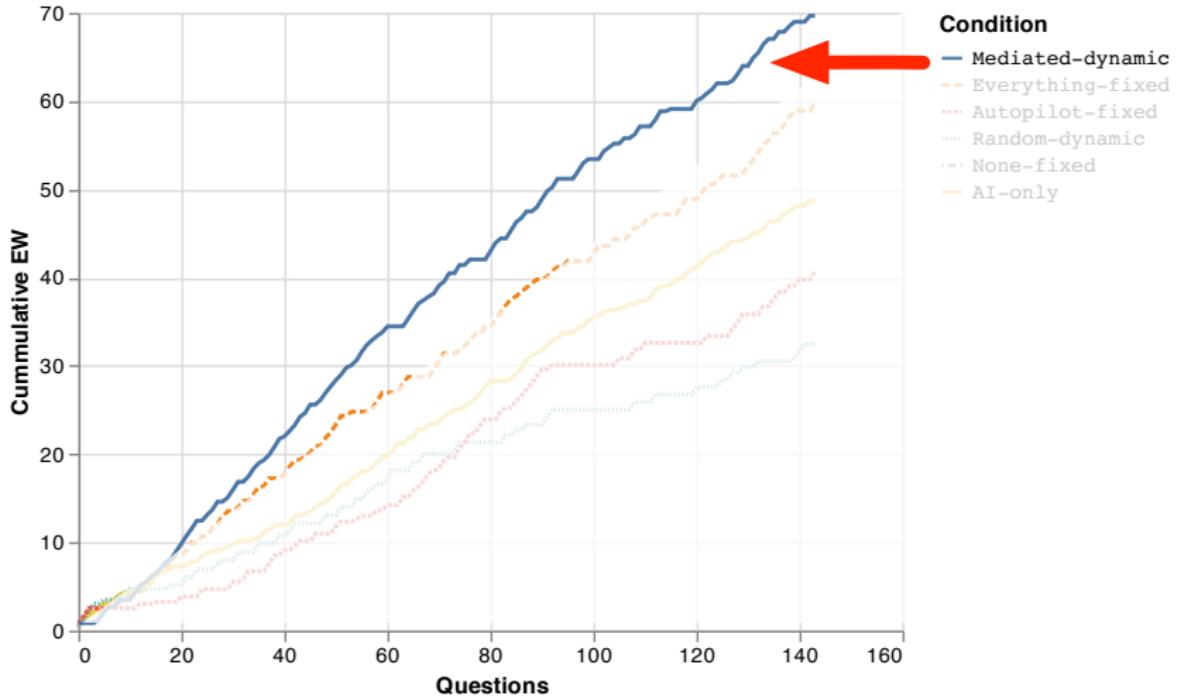
Bandit actions [?]: turn each of the explanations (Guess, Highlight, Evidence) on or off.



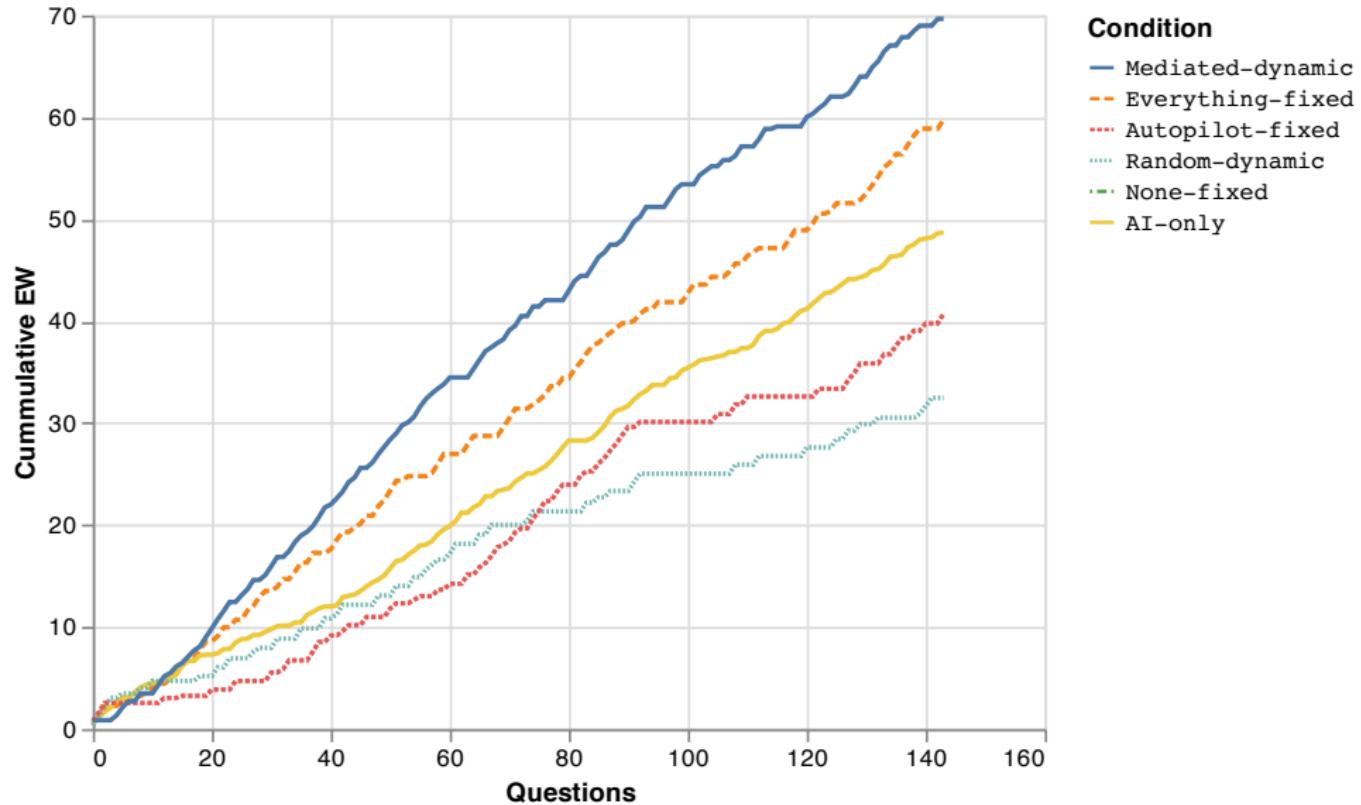
Human alone without an AI teammate



AI alone without a human teammate

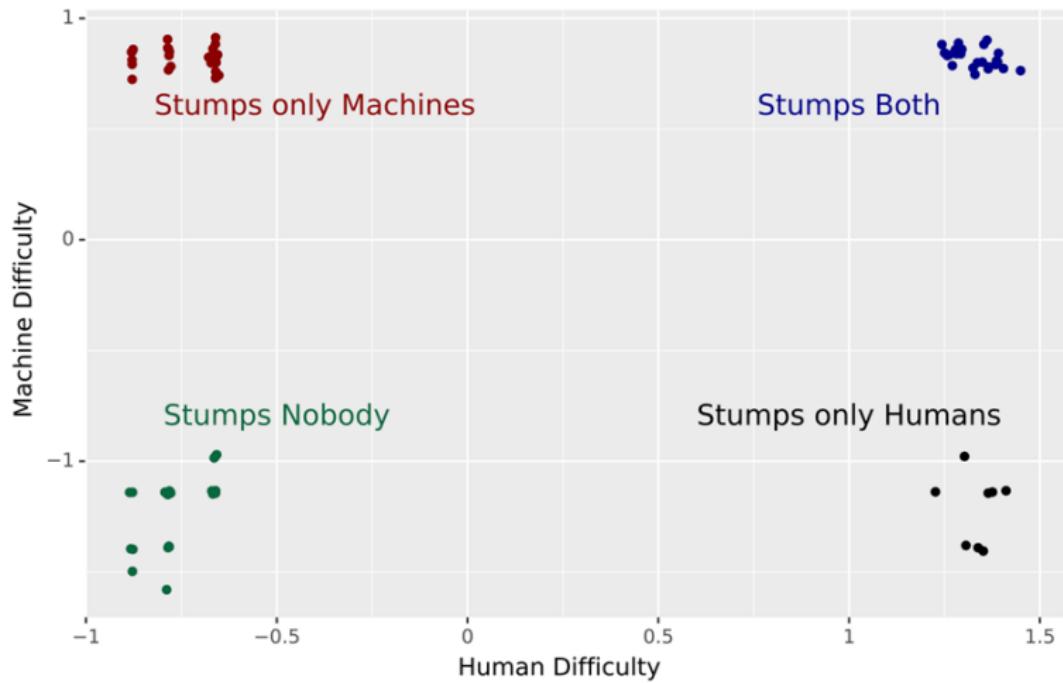


Dynamic assistance to human

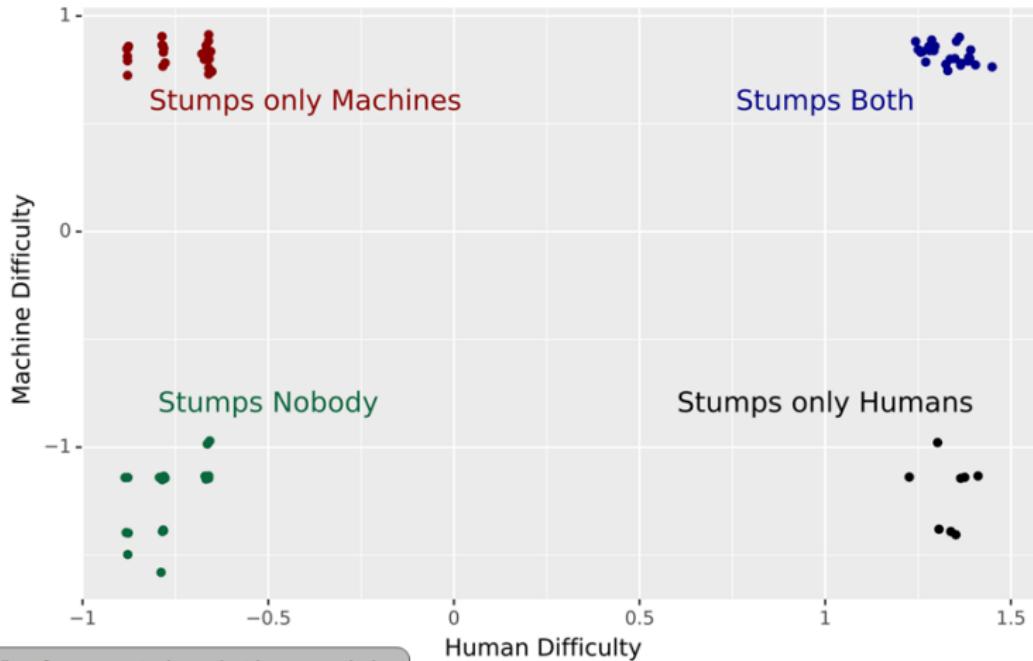


Better than showing everything!

## What makes for Adversarial Example

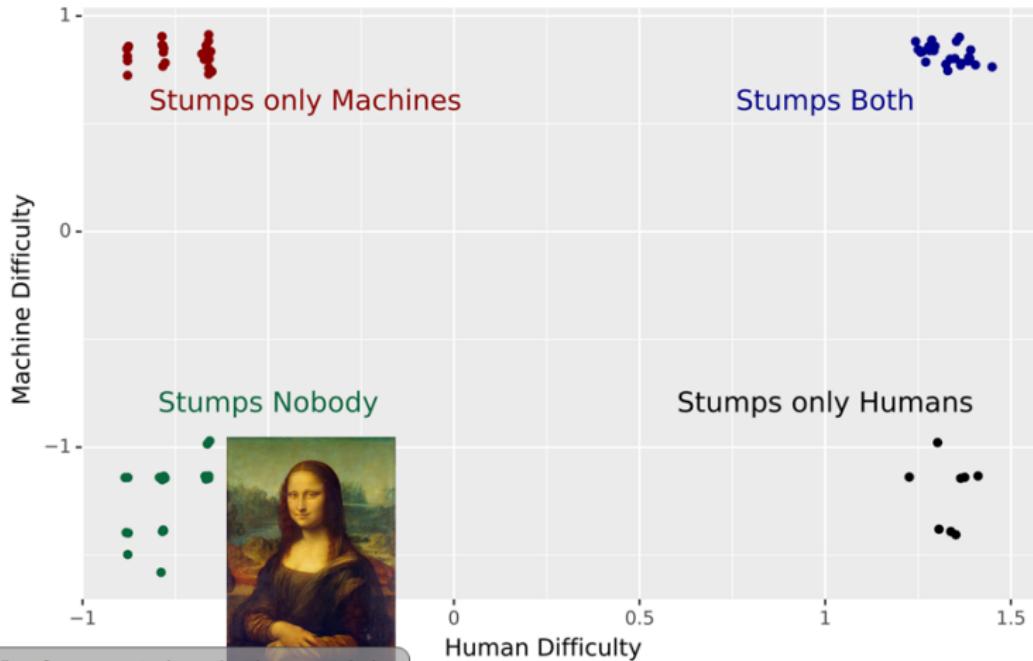


## What makes for Adversarial Example



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

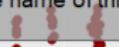
## What makes for Adversarial Example



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?

## 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 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



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?

## 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?

## Simultaneous Interpretation is Hard!

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



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



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



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



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

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

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

# STACL: Simultaneous Translation with Integrated Anticipation & Controllable Latency



**Liang Huang**  
Principal Scientist, Baidu Research

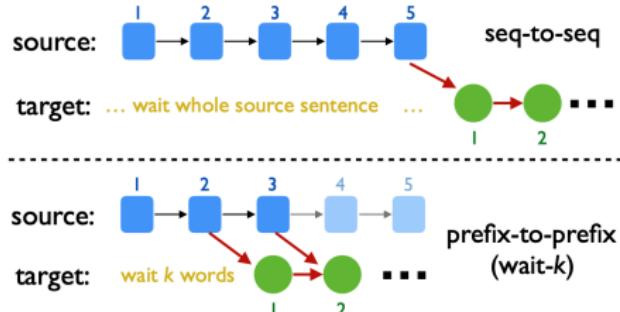
Assistant Professor (on-leave), Oregon State University



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

# Prefix-to-Prefix Translation

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



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,

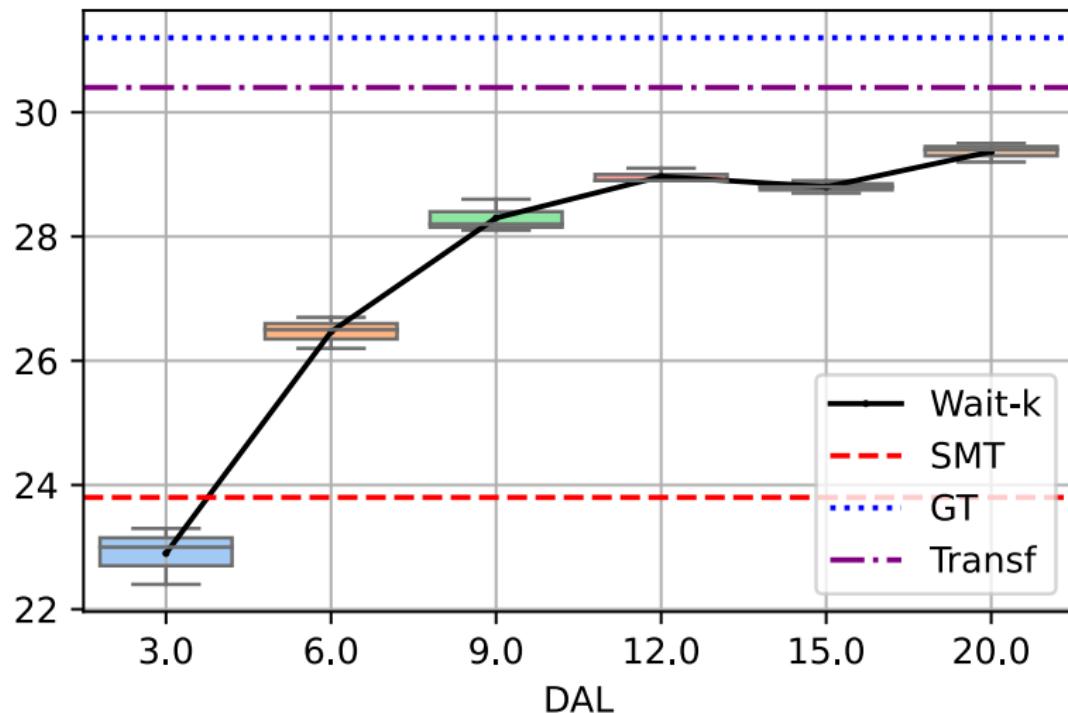
#### Target :

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.

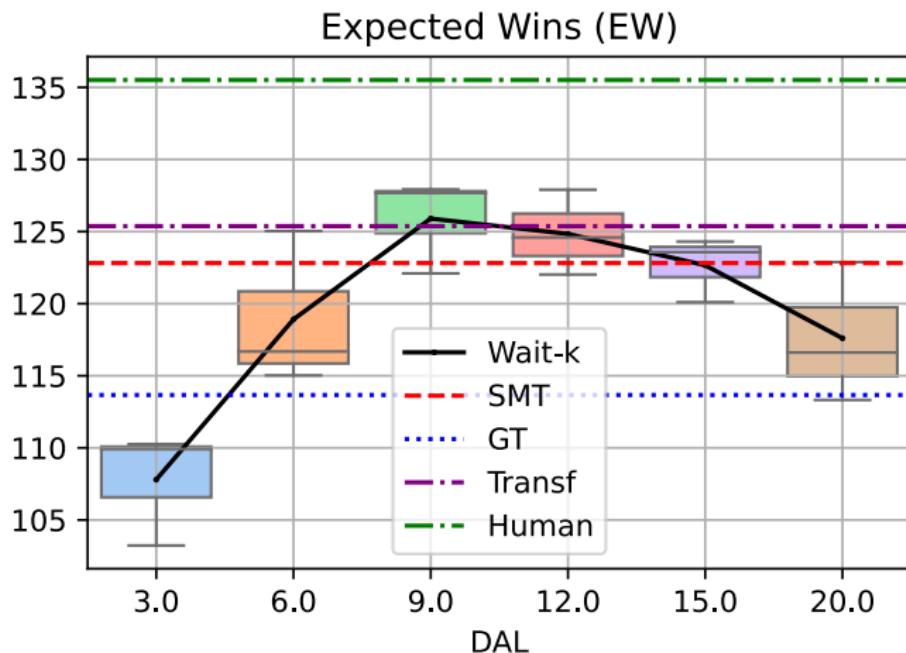
- You're a contestant on a Polish game show
- You have access to a simultaneous translation system
- Your job is to answer the question before your opponent (as quickly as possible)
- Keep question answerer the same, vary translation

## BLEU results for modern Simultaneous Translation Systems

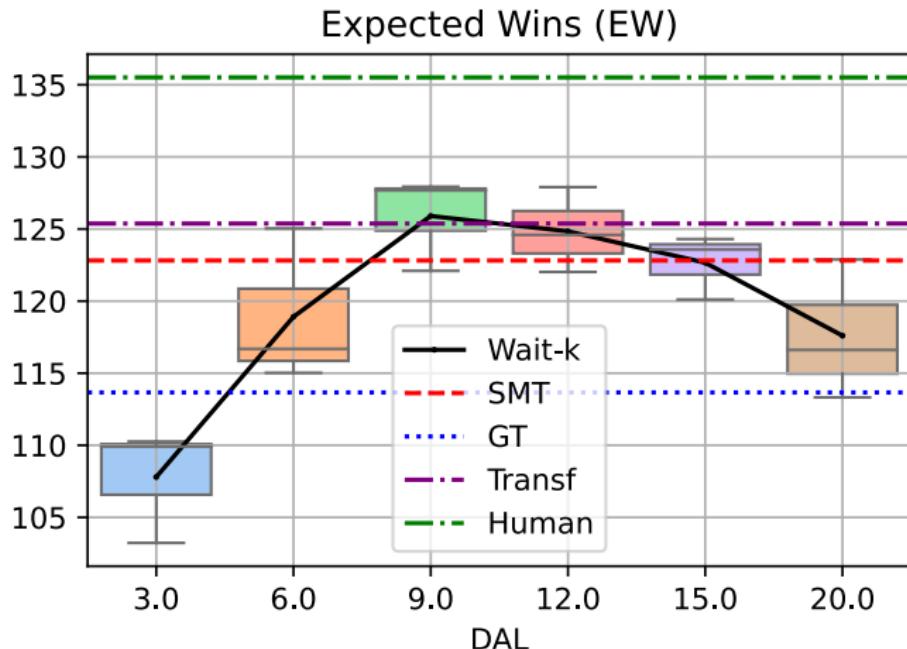
BLEU on WMT-pl



## Downstream QA Results

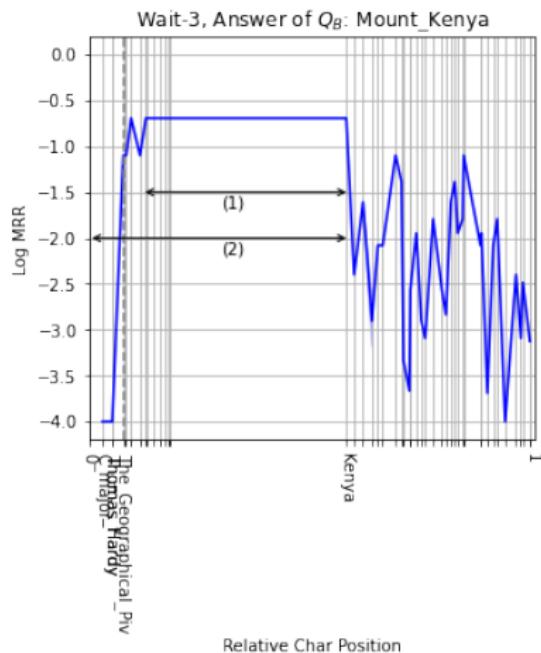


## Downstream QA Results



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 ...

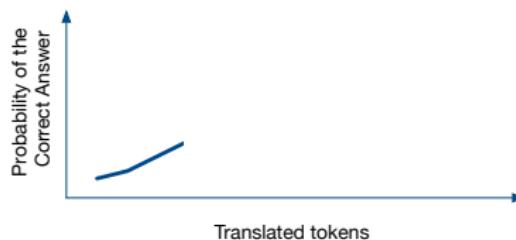


**Guess:** ??

## When are Mistakes / Hallucinations Harmful?

This coordinate determines

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

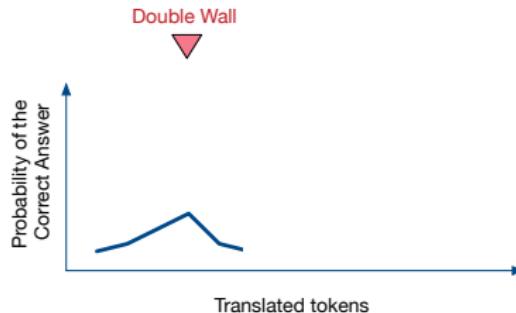


**Guess:** IP Address

## 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 ...

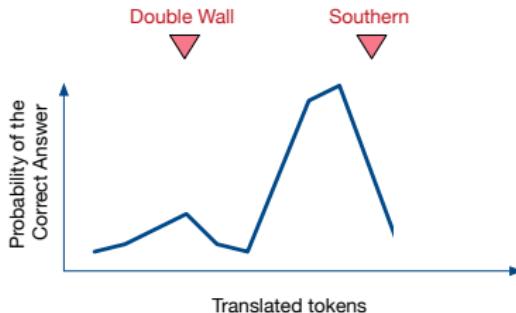


**Guess:** Longitude

## 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 ...

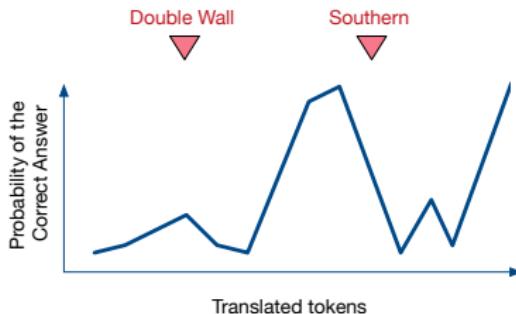


**Guess:** Spherical Coordinate

## 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

## Thanks

### Collaborators

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

### Funders



### Supporters

International Academic  
**Competitions**

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

National Academic Quiz Tournaments, LLC

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...

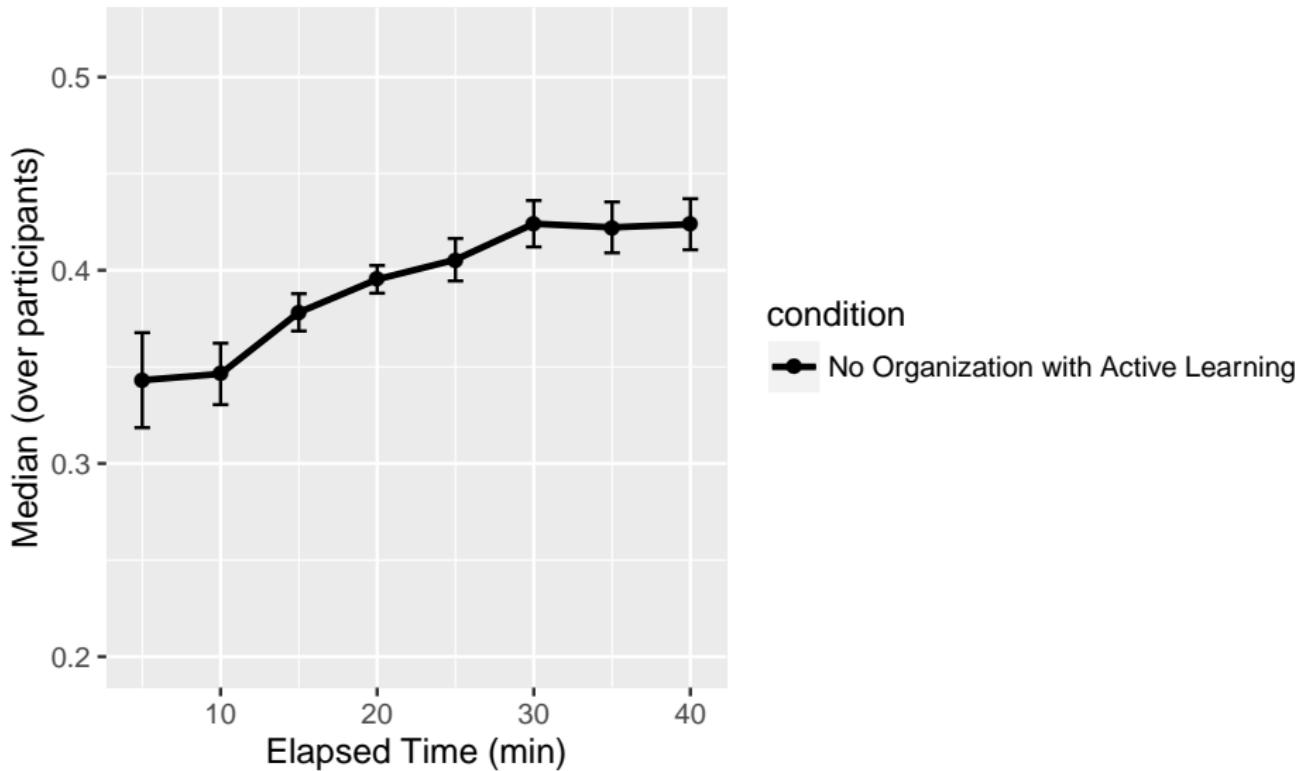
To rescind any unobligated discretionary appropriations returned to the F...

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

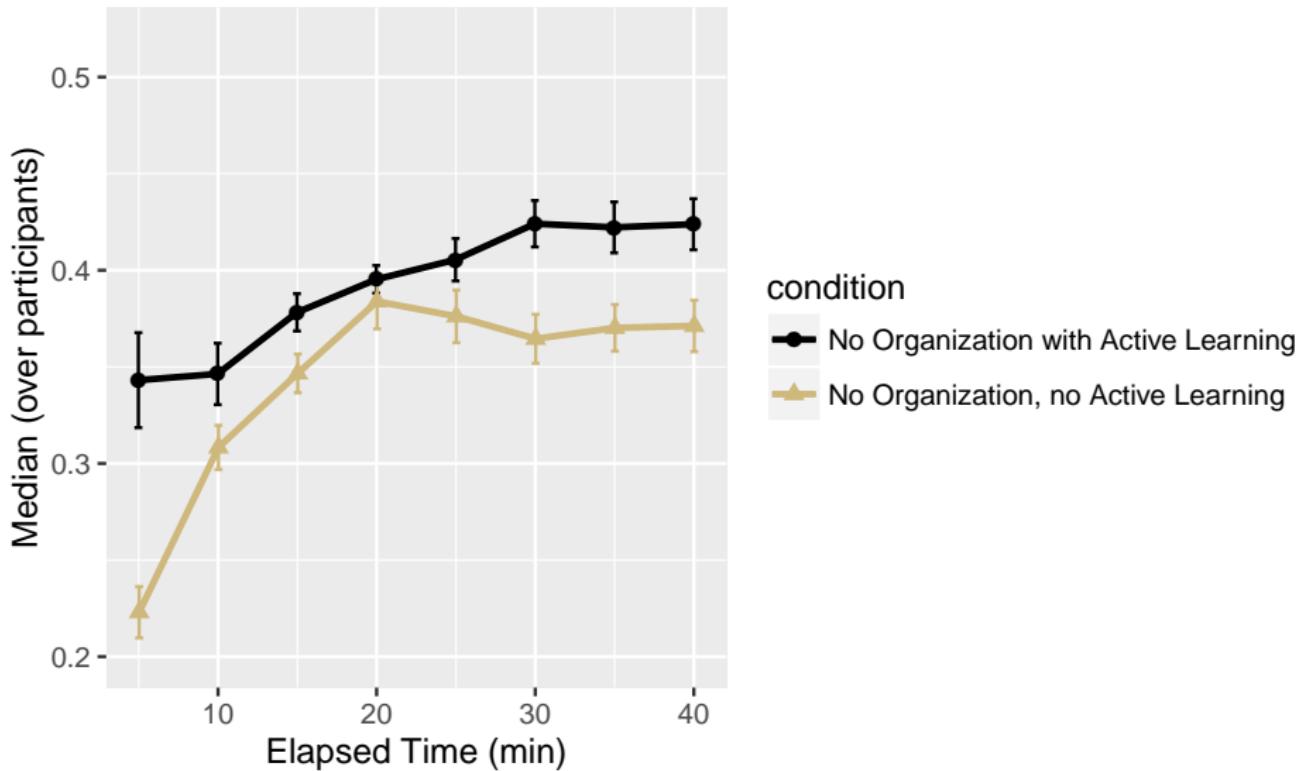
Making appropriations for energy and water development and related age...

Covered Themes Progress:

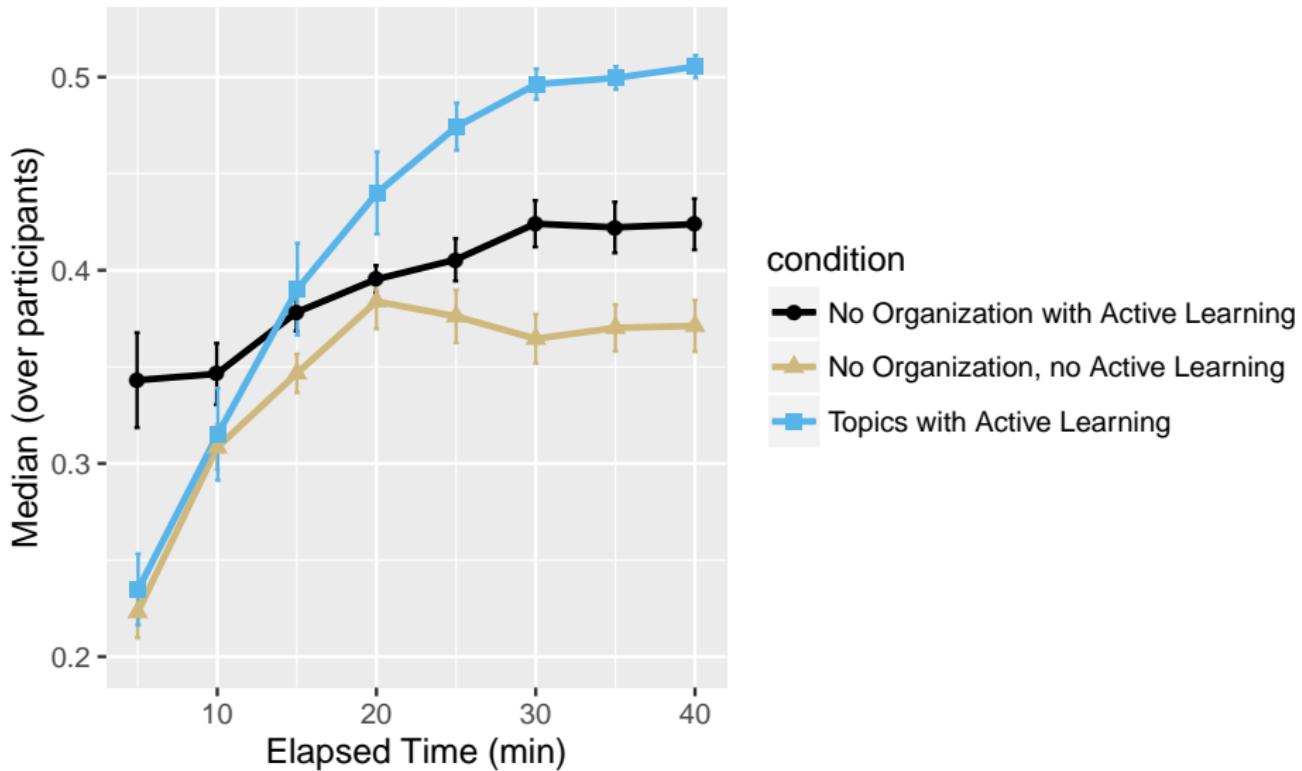
Direct users to document



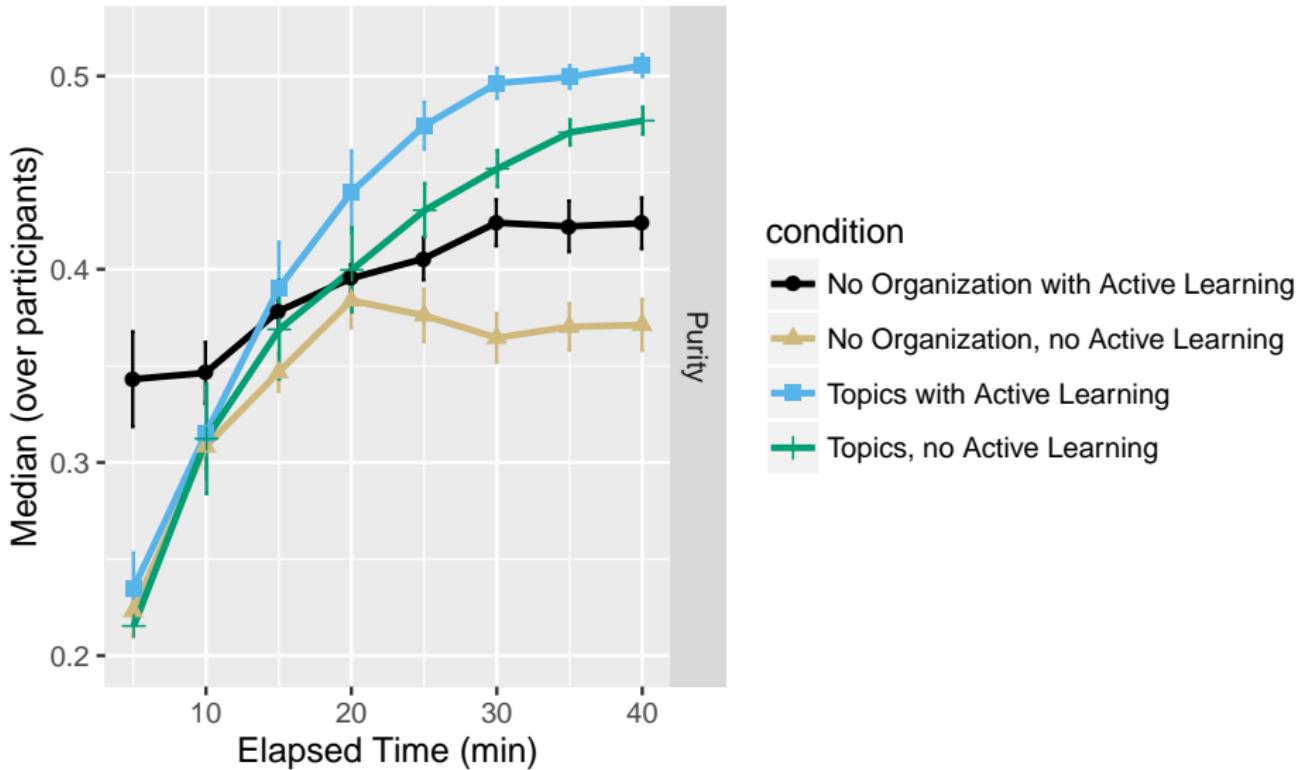
Active learning if time is short



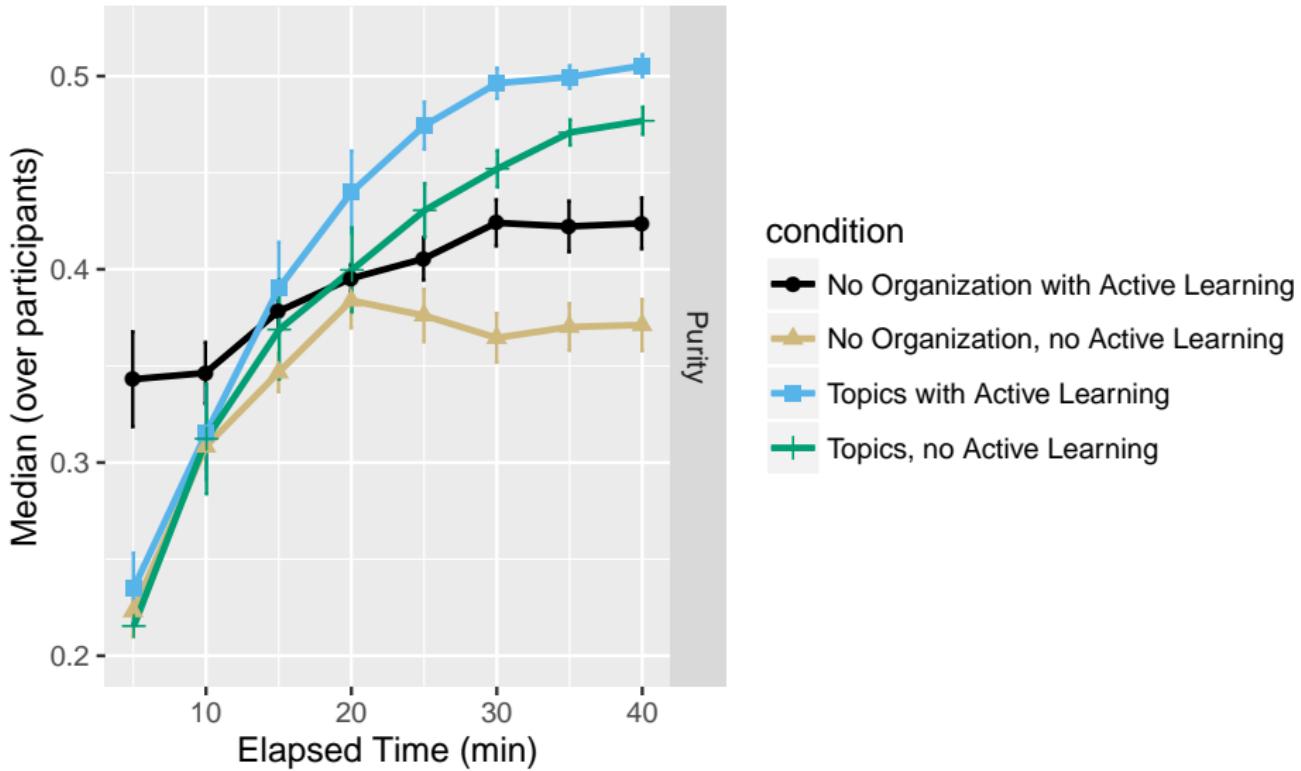
Better than status quo



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

## References

# RC TRUST



## Psychology & Social Sciences

We empower humans to understand, control, and verify intelligent algorithms.



## Artificial Intelligence & Machine Learning

We develop self-explainable, accountable, and verifiable machine learning algorithms.



## Data Science & Statistical Learning

We infer uncertainty, causality, and representations hidden in complex data sets.



## Cybersecurity & Privacy

We preserve privacy, security and trust in algorithms for safety-critical systems.

# RC TRUST



## Psychology & Social Sciences

We empower humans to understand, control, and verify intelligent algorithms.



## Artificial Intelligence & Machine Learning

We develop self-explainable, accountable, and verifiable machine learning algorithms.



## Data Science & Statistical Learning

We infer uncertainty, causality, and representations hidden in complex data sets.



## Cybersecurity & Privacy

We preserve privacy, security and trust in algorithms for safety-critical systems.

## Psychology & Social Sciences

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

# RC TRUST



## Psychology & Social Sciences

We empower humans to understand, control, and verify intelligent algorithms.



## Artificial Intelligence & Machine Learning

We develop self-explainable, accountable, and verifiable machine learning algorithms.



## Data Science & Statistical Learning

We infer uncertainty, causality, and representations hidden in complex data sets.



## Cybersecurity & Privacy

We preserve privacy, security and trust in algorithms for safety-critical systems.

## AI & ML

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

# RC TRUST



## Psychology & Social Sciences

We empower humans to understand, control, and verify intelligent algorithms.



## Artificial Intelligence & Machine Learning

We develop self-explainable, accountable, and verifiable machine learning algorithms.



## Data Science & Statistical Learning

We infer uncertainty, causality, and representations hidden in complex data sets.



## Cybersecurity & Privacy

We preserve privacy, security and trust in algorithms for safety-critical systems.

## Data Science & Statistical Learning

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

# RC TRUST



## Psychology & Social Sciences

We empower humans to understand, control, and verify intelligent algorithms.



## Artificial Intelligence & Machine Learning

We develop self-explainable, accountable, and verifiable machine learning algorithms.



## Data Science & Statistical Learning

We infer uncertainty, causality, and representations hidden in complex data sets.



## Cybersecurity & Privacy

We preserve privacy, security and trust in algorithms for safety-critical systems.

## Cybersecurity and Privacy

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