

The Good, the Bad, and the Ugly: The Role of AI Quality Disclosure in Lie Detection

Sanchaita Hazra

Job Market Paper

NABE Tech Economics Conference (TEC2024)

AI Systems in Contemporary Life

- AI emerged as an advisor for individuals in daily life activities: entertainment, dating, dietary choices
(Fast and Schroeder, 2020, Yeomans et al., 2019, Green and Chen, 2019, Bucinca et al., 2021)
- Two major dimensions have evolved: AI over-reliance & AI aversion
- As algorithms are already deployed to detect lies on social media (Alom et al., 2020; Monaro et al., 2020)
 - ◆ Overreliance & lack of AI transparency can have serious consequences on lie detection online
(U.S. Congress Act, EU, OECD, G7 AI policies)
- However, people's innate expectations about AI advisors' efficacy can also shape overreliance
 - ◆ So, info about (or lack of) AI-advisor's efficacy and expectations can jointly shape AI reliance
- To our knowledge, no research on impacts of AI efficacy info & users' expectations on AI reliance
 - ◆ We examine this in the context of detecting text-based lies

Quora

Is it true that West Bengal has a 27% Muslim population?



Answer



Follow · 8



Request



ChatGPT

According to the 2011 census, the Muslim population in West Bengal is 27.01%. It is the second-largest Muslim population in any Indian state, after Uttar Pradesh. However, it is important to note that this information is from the last census and the current population may have changed.



Anonymous

2y

Real Muslim population in Bengal (2011)- 27,861,060 (30.5%) instead of 24,654,825 (27%).



Sourav Ghosh · Follow

Activist at Politics of India (2010–present) · 5y

Wrong!

The current population of Muslims is more **than 40% already**. This is because of unlimited support they get for votebank of Current Govt + Infiltrators + the lack of Courage of Bengal BJP. Bengal is the edge of Self-Destruction, Unless all of us do something together all may be lost because of **One Mad Politicians EGO!**

I am a Indian 1st then comes my religion, Time has come to fight for our motherland before it becomes _____. Fill in the blanks yourself.



Ananda Mondal · Follow

Lives in West Bengal, India (1992–present) · 6y

As per 2011 census, Muslim population shows 27.01%. It may get changed now.

<http://www.censusindia.gov.in/2011census/C-01/DDW19C-01%20MDDS.XLS>

However, real-life conversations may be contaminated by motivated reasoning

We retrieve conversations from a TV game show: **To Tell The Truth** that resembles to real-life situations, but free from motivated reasoning

To Tell The Truth (T4)



Contestants:

- All claim to be **RC**,
- One of them is the **RC**
- The **RC** is always to be truthful, others can lie

Judges:

- Listen to the **affidavit**
- **Cross-question** the contestants to **identify** the real John/Jane Doe independently

Affidavit:

True factual information about a real contestant (**RC**), publicly announced



Each contestant can receive a reward upto **\$3500 (adj.)** for successfully deceiving all judges

A Transcript

Detect who are impersonating William D Hackett

Back-and-forth Q&As
between the judges
and the challengers

Excluded transcripts where
judges explicitly mention that
they used audio-visual cues

Contestants' identities are
swapped to make sure that the
answers cannot be copied from
internet

Gold data on the RC is known
to us, but not included in a
transcript

Conversations

Affidavit: I, William D. Hackett, am a captain in the United States Army. I have spent five years on Alaskan duty and served during World War II with the United States Mountain Troops. I am a mountain climbing specialist. I have made more than 500 major ascents in 12 countries. I am the only man in the world who have climbed to the highest point on five different continents. I have been selected as a member of the United States team, which in 1959 will attempt Mount Everest, the highest mountain in the world.

Q: Number Two, where in the world is a mountain called University Peak?
A: In Alaska.

Q: Number One, what is the highest mountain in the continent of the United States?
A: Oh, that's Mount Whitney.

Q: Number Three, about how high is Mount Whitney?
A: 20,000 feet.

Q: Number Two, how high is Mount Whitney?
A: 14,400.

Q: Number One?
A: 14,496.

Q: Number Two, what are those little pointed steel things that they dig into mountains that hold the rope on?
A: Tetons.

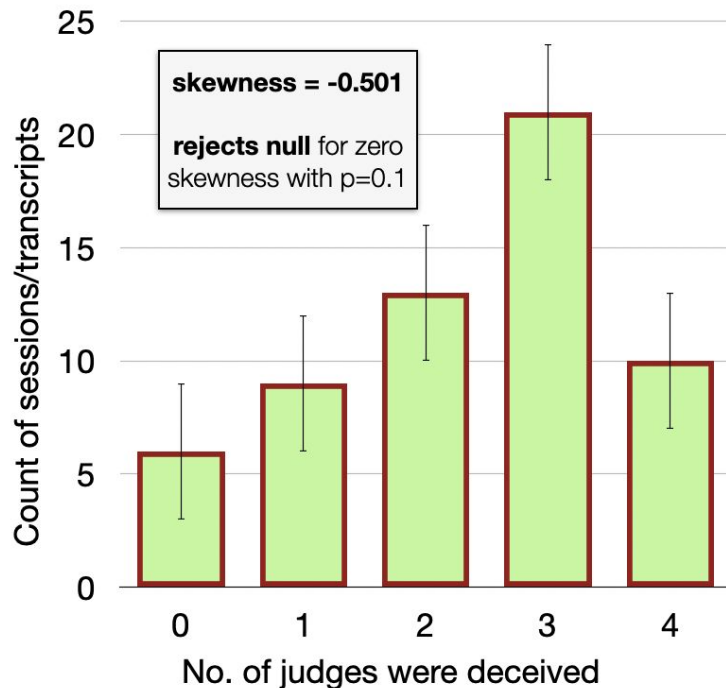
Q: Are they, pardon me?
A: Tetons.

Automatic Detection of Text-based Deception using AI

On a random draw of the **five** transcripts,
how many of the AI's guesses
do you think can be correct?

Automatic Detection of Text-based Deception using AI

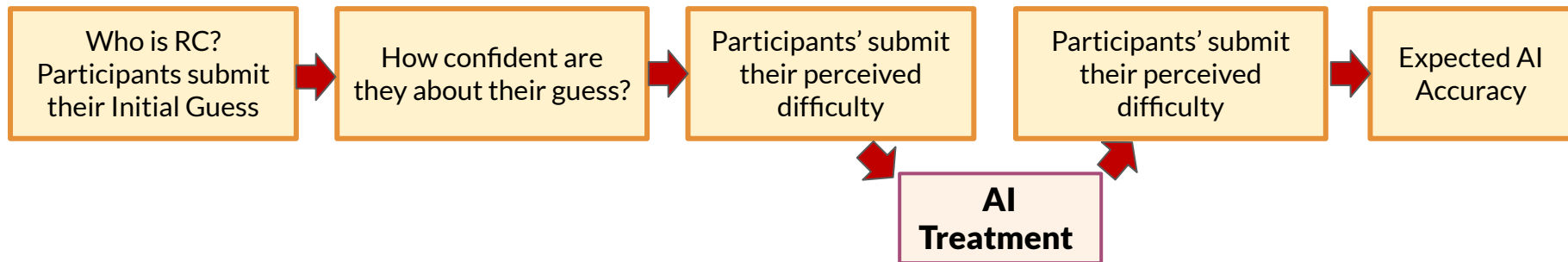
- Our model performs the best (vs gpt3/3.5/base gpt-4) with **56.06%** accuracy on 132 transcripts (judges' accuracy ~ 44%)
- With defined predictors, accuracy improves 5 pts
- Detects correctly in cases (>50%) where most judges in the original game show fail



Three Qualities of AI advisors

- Three sets of transcripts, each with five transcripts, that vary systematically in AI accuracy
- Low Quality (**LQ**): 20% AI accuracy (1 out 5 AI guesses are correct)
- Medium Quality (**MQ**): 40% AI accuracy (2 out 5 AI guesses are correct)
- High Quality (**HQ**): 60% AI accuracy (3 out 5 AI guesses are correct)

Experimental Treatments (N = 545, 2x3 Design)



Black Box

Your guess for Set 1 was **Number Two**.

AI's guess for this Set 1 is **Number One**.

If you would like to review the details of this set: [Click Here](#)

Which of the two guesses would you like to submit as your final guess for this set?

- ☐ Submit the AI's guess
- ☐ Submit my guess



Your confidence level is 50.

If your final guess is correct, you will receive 75 cents. If your final guess is incorrect, you will receive 75 cents.

With Information

Your guess for Set 1 was **Number One**.

AI's guess for this Set 1 is **Number One**.

The AI correctly identified the real John or Jane Doe in three of the five sets, which is equal to a 60% accuracy rate.

If you would like to review the details of this set: [Click Here](#)

Which of the two guesses would you like to submit as your final guess for this set?

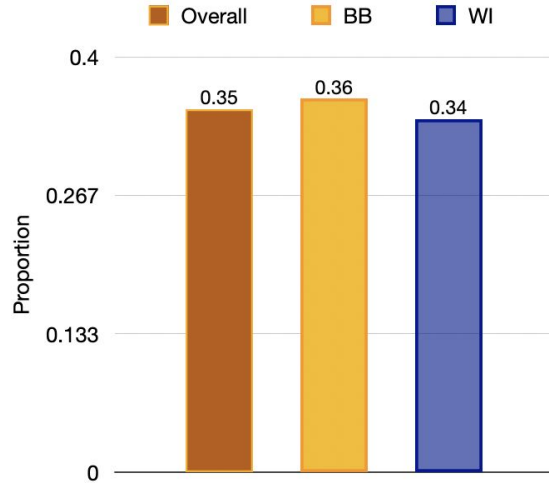
- ☐ Submit the AI's guess
- ☐ Submit my guess



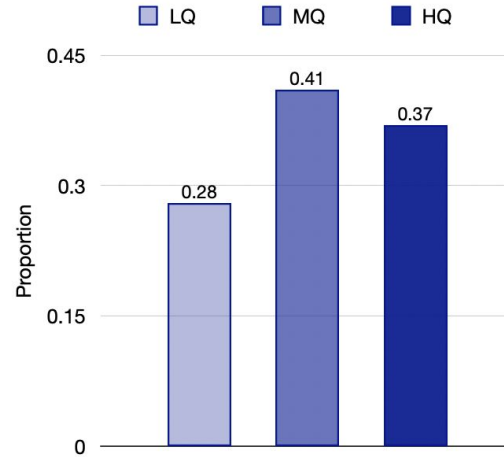
Your confidence level is 50.

If your final guess is correct, you will receive 75 cents. If your final guess is incorrect, you will receive 75 cents.

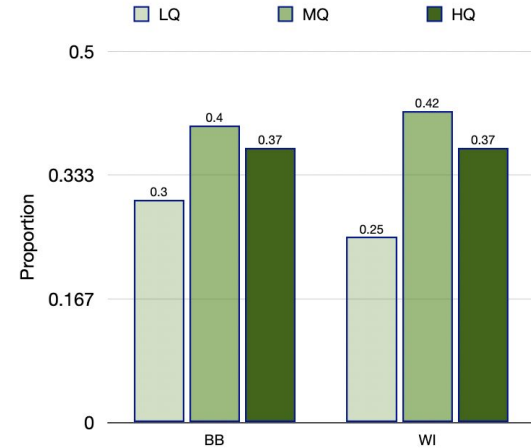
Subject's Innate Detection Ability: Initial Accuracy



Individual's innate ability are statistically equivalent in both information conditions



Accuracy lower than random in LQ



Initial Accuracy is significantly lower in LQ than in MQ/HQ; equivalent between MQ & HQ

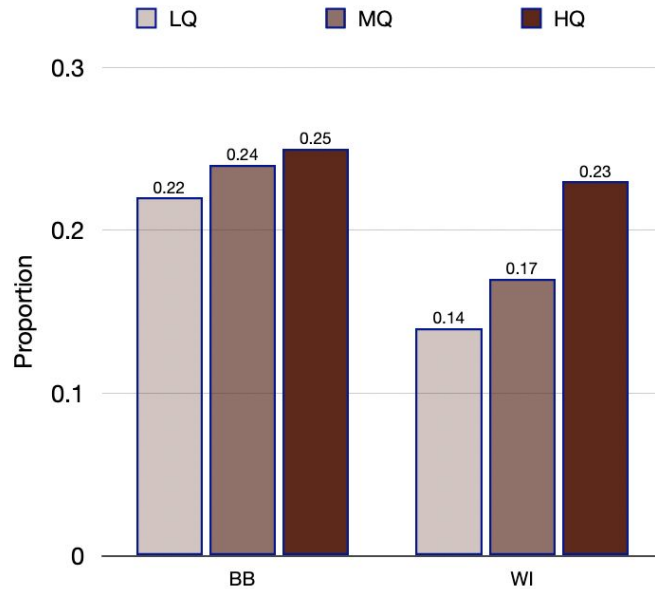
Subjects' initial accuracy lowest for the set of transcripts where AI advisor quality lowest

Subject's Innate Detection Ability: Initial Accuracy

- Subjects' innate ability to detect truth does not differ between the information conditions, overall as well as for each AI advisor
- Any variation in subjects' AI reliance across information conditions for any given AI advisor quality can be attributed to the AI efficacy information

	Initial Accuracy
$BB - WI$	0.7114 (0.4769)
$BB_{LQ} - WI_{LQ}$	1.5542 (0.1205)
$BB_{MQ} - WI_{MQ}$	-0.609 (0.5427)
$BB_{HQ} - WI_{HQ}$	0.3845 (0.7007)
t-statistic for H_0 : LHS accuracy - RHS accuracy = 0 & H_1 : LHS accuracy - RHS accuracy $\neq$ 0 reported in the table with p-value in parentheses.	

Switch: Individual's Reliance on AI

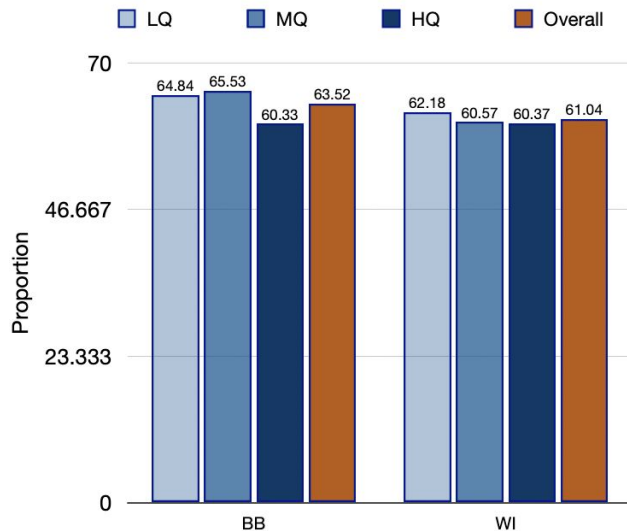


Switch rate significantly lower in FI than in BB (overall; LQ & MQ)
However, switch rate equivalent between FI and BB for HQ

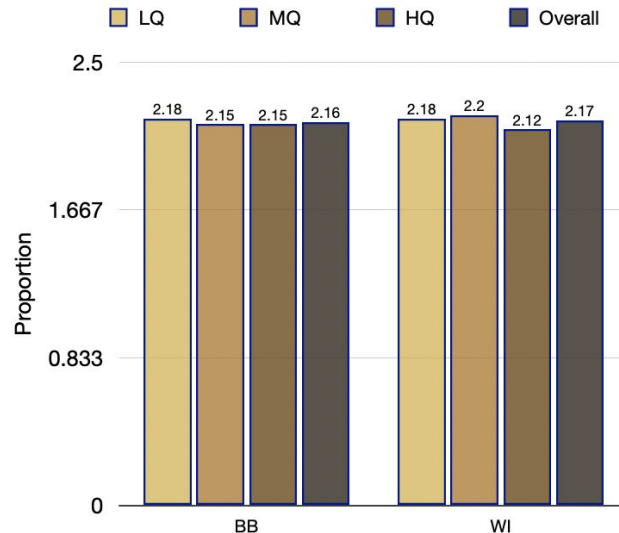
	Switch Rate
$BB - WI$	3.6884*** (0.0002)
$BB_{LQ} - WI_{LQ}$	3.1486*** (0.0017)
$BB_{MQ} - WI_{MQ}$	2.6517*** (0.0082)
$BB_{HQ} - WI_{HQ}$	0.7120 (0.4766)
t-statistic for $H_0 : \text{LHS} - \text{RHS} = 0$ & $H_1 : \text{LHS} - \text{RHS} \neq 0$ reported in the table with p-value in parentheses.	

Factors Affecting Switch

Confidence in own guess

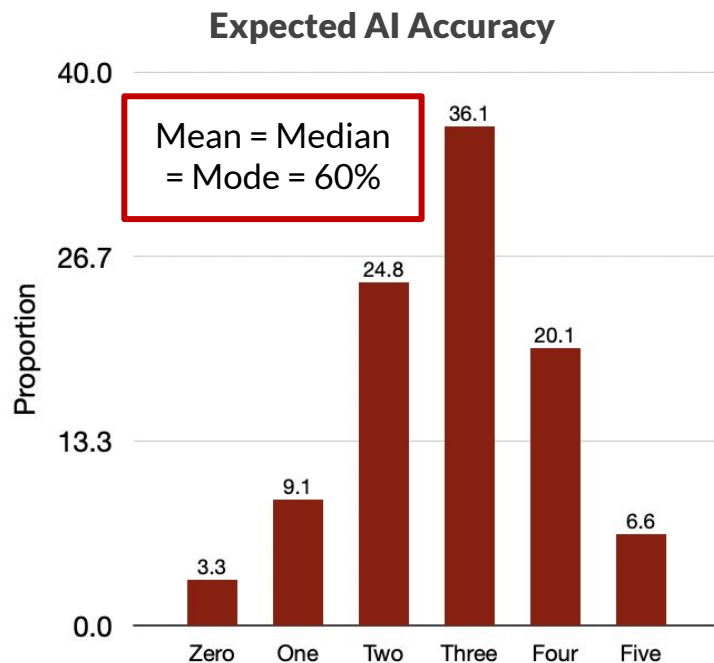


Perceived difficulty



No difference in “Confidence in own guess” or “Perceived difficulty” across Information conditions or across different qualities of AI advisors

Factors Affecting Switch



In HQ, actual = expected AI accuracy
But falls short of subjects' expectations in MQ & LQ

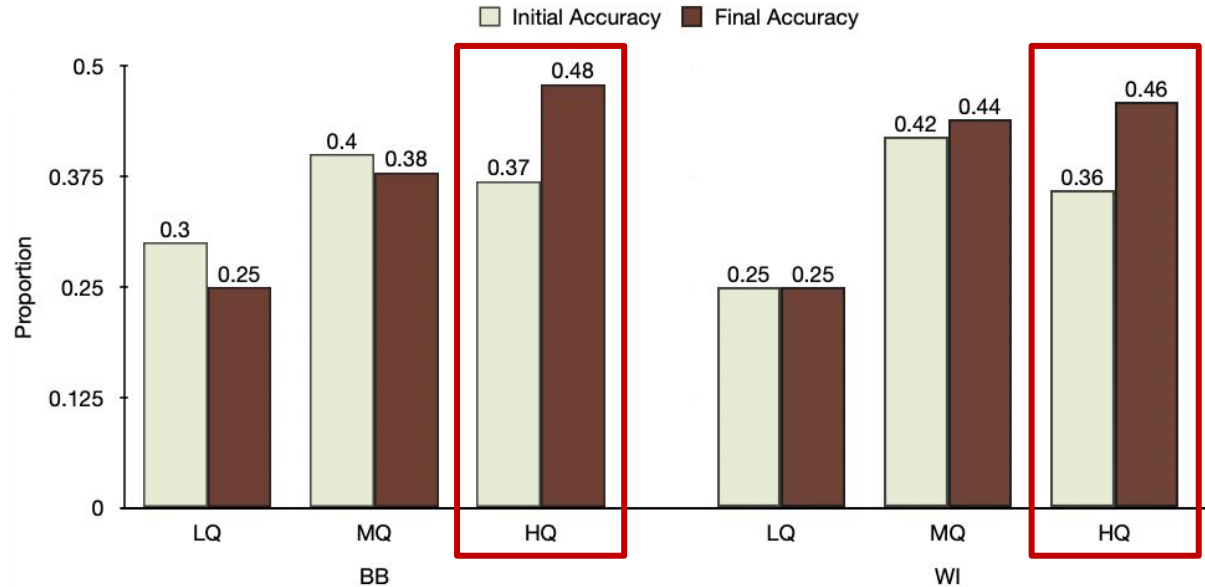


	Switch Overall	Switch BB
Confidence in own guess	-0.00298*** (0.000)	-0.00274*** (0.000)
Perceived difficulty	0.0517*** (0.000)	0.0518*** (0.009)
Relative confidence		
Perceived modal difficulty		
Expected AI accuracy		0.0435*** (0.000)
<i>LQ</i>	-0.0215 (0.485)	-0.0148 (0.625)
<i>MQ</i>	0.00960 (0.764)	0.0111 (0.726)
<i>WI</i>	-0.0182 (0.559)	
<i>LQ * WI</i>	-0.0688 (0.118)	
<i>MQ * WI</i>	-0.0709 (0.119)	
Constant	0.317*** (0.000)	0.177** (0.018)

Conjecture:

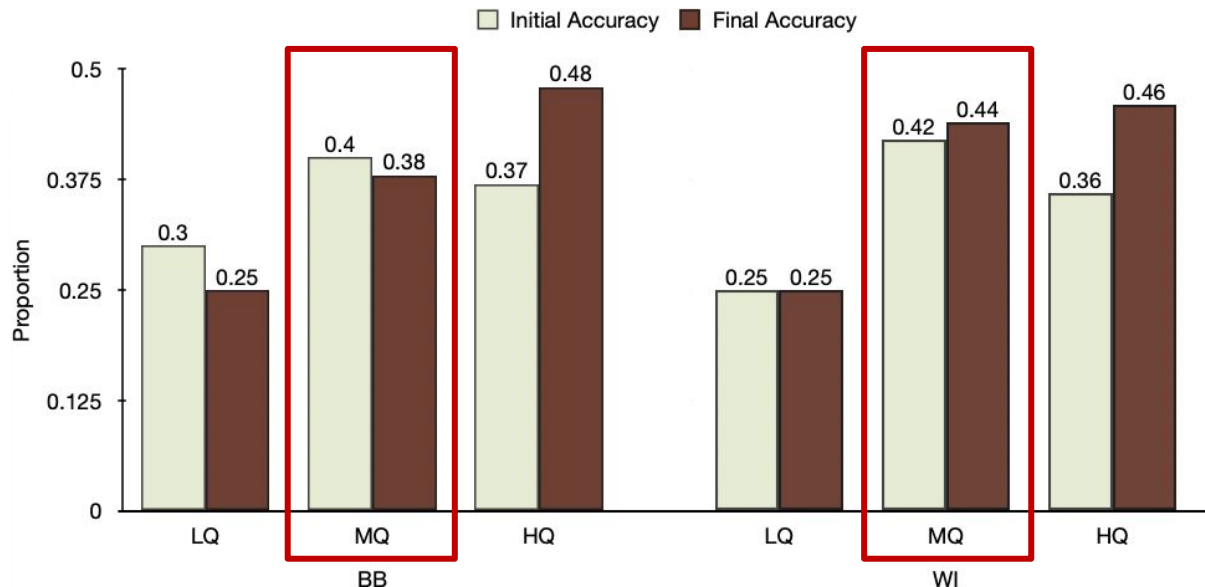
In HQ, subjects do not perceive any need to alter their AI reliance as their expectation matches
But in MQ and LQ, subjects reduce their AI reliance

Do Switches Help? The “Good” Effect



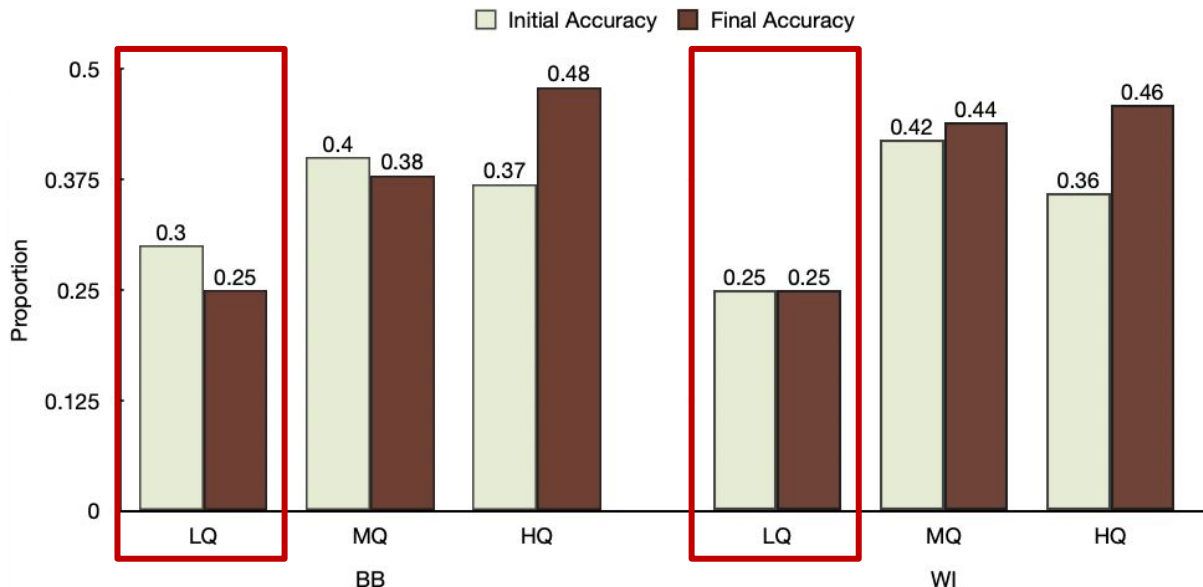
- In HQ, final accuracy significantly greater than initial accuracy, both in BB and WI
- A highly capable AI advisor **significantly enhances** truth detection
- Disclosing the capability of highly capable AI advisor is **inconsequential**

Do Switches Help? The “Bad” Effect



- In **MQ**, final accuracy does not improve over innate capability
- When unaware of quality information (**BB**), decline in accuracy (but, insignificant)
- Not disclosing the capability of modest not-so-good quality AI advisor is **bad**

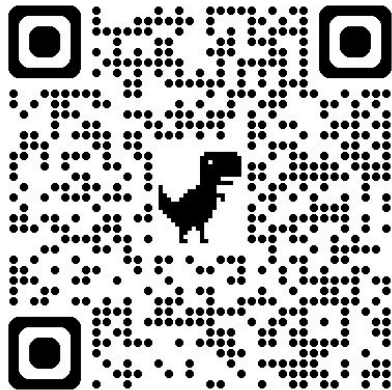
Do Switches Help? The “Ugly” Effect



- In **LQ**, final truth detection rate **falls significantly below** intrinsic rate
- Lies could have been detected, if participants relied on own ability instead of switching to AI's guess.
- If informed, participants recalibrate reliance on AI advisor, preventing the drop in truth detection

Dataset

We release **our novel conversational dataset (T4Text)** , unique to the previous deception datasets, that contains objective truths, forming the basis of lie detection



<https://github.com/sanchaitahazra/T4Text>



Thanks!

✉ @hsanchaita

🌐 sanchaitahazra.com

✉ sanchaita.hazra@utah.edu

Co-authors



Haimanti
Bhattacharya



Subhasish
Dugar



Bodhisattwa
Majumder