The **Thin Line** Between **Comprehension** and **Persuasion** in **LLMs**

Adrian de Wynter[1,2] and Tangming Yuan[1]

[1]*The University of York*

[2]*Microsoft*

adewynter@microsoft.com

Microsoft

Motivation

Do LLMs *understand* what you / they are talking about?

- No, they do not (Webson and Pavlick, 2022; Ye and Durrett, 2022; Chen et al., 2024).
- But there *is* something there (Khan et al., 2024; Zhou et al., 2026; De Wynter, 2026).

Motivation

Do LLMs *understand* what you / they are talking about?

- No, they do not (Webson and Pavlick, 2022; Ye and Durrett, 2022; Chen et al., 2024).
- But there *is* something there (Khan et al., 2024; Zhou et al., 2026; De Wynter, 2026).

Answering this is **important,** especially when dealing with explainability.

Motivation

Do LLMs *understand* what you / they are talking about?

- No, they do not (Webson and Pavlick, 2022; Ye and Durrett, 2022; Chen et al., 2024).
- But there *is* something there (Khan et al., 2024; Zhou et al., 2026; De Wynter 2026).

Answering this is **important,** especially when dealing with explainability.

What *is* understanding? How do you measure it?

- Understanding and its measurement is exceptionally ill-defined (Grimm, 2025).
- Our goal is **not** to give an account of this
- Our grounding is based off language games and the Imitation Game.

Theoretical Grounding

In a language game, **words are meaningless** until they are used in-context.

For us: showing understanding of the word means showing its suitable application given the context.

In the Imitation Game, a machine is successful if it is indistinguishable from a human in sustained dialogue.

Its original assumptions **do not work**. For us: a machine is successful if it is able to convince the user ***even*** if they know it is a machine.

Theoretical Grounding

Kvanvig (2003) defined understanding as a collection of information along with a grasp of the reasons why it occurs.

There are many competing definitions and nuances here—see Grimm (2025) or De Wynter and Yuan (2026) for accounts of these.

It follows that, **to measure understanding**, one needs to:

1. Decompose information into reasons, and
2. Measure these reasons.

For example, by polling the learner on the meaning of these reasons in context.

Motivation

Do LLMs *understand* what you / they are talking about?

- No, they do not (Webson and Pavlick, 2022; Ye and Durrett, 2022; Chen et al., 2024).
- But there *is* something there (Khan et al., 2024; Zhou et al., 2026; De Wynter 2026).

Answering this is **important,** especially when dealing with explainability.

Motivation

Do LLMs *understand* what you / they are talking about?

- No, they do not (Webson and Pavlick, 2022; Ye and Durrett, 2022; Chen et al., 2024).
- But there *is* something there (Khan et al., 2024; Zhou et al., 2026; De Wynter 2026).

Answering this is **important,** especially when dealing with explainability.

What *is* understanding? How do you measure it?

- Understanding and its measurement is exceptionally ill-defined (Grimm, 2025).
- Our goal is **not** to give an account of this (that's my thesis)
- Our grounding is based off language games and the Imitation Game.

Motivation

Do LLMs *understand* what you / they are talking about?

- No, they do not (Webson and Pavlick, 2022; Ye and Durrett, 2022; Chen et al., 2024).
- But there *is* something there (Khan et al., 2024; Zhou et al., 2026; De Wynter 2026).

Answering this is **important,** especially when dealing with explainability.

What *is* understanding? How do you measure it?

- Understanding and its measurement is exceptionally ill-defined (Grimm, 2025).
- Our goal is **not** to give an account of this (that's my thesis)
- Our grounding is based off language games and the Imitation Game.

… does it matter if they do/don't?

In this talk

Experimental setup

- Data, methods, etc.

Experiment 1: Can LLMs *understand* arguments?

- Distinct than ‘are they answering grammatically?’

Experiment 2: How persuasive are LLM arguments?

- We all know LLMs are persuasive. What about their arguments?

What have we learnt?

- Conclusion, future directions.

In this talk

Experimental setup

- Data, methods, etc.

Experiment 1: Can LLMs *understand* arguments?

- Distinct than ‘are they answering grammatically?’

Experiment 2: How persuasive are LLM arguments?

- We all know LLMs are persuasive. What about their arguments?

What have we learnt?

- Conclusion, future directions.

Core Idea

A **FDM** is a collection of rules defining the behaviour of a player in a conversation (e.g., an educational debate).

Idea:

1. Generate (informal) debates between humans and LLMs (any topic).
2. Use FDMs for controllable behaviour during these (i.e., to ensure persuasiveness)
3. Measure LLM understanding of the dialogue *and* sway

Rule	Definition
Assertions	P
Questions	Is it the case that P?
Challenges	Why P?
Withdrawals	No commitment P
Resolution demands	Resolve whether P

Sample move types for the FDM used.

The actual FDM specifies *what* to do in various situations

Dataset

Corpus of debates separated into three subsections:

- Human – LLM (‘LLM \ FDM’)
- Human – LLM + FDM
- LLM – LLM + FDM

Three different annotations:

1. Experience by human debaters
2. Dialogical structures (third-party human annotators) and winner
3. Experience by human audience

IAA (for #2): W. Cohen’s $\kappa_w = 0.86$ (structures), $\kappa_w = 0.41$ (winner).

Experimental setup

- Data, methods, etc.

Experiment 1: Can LLMs *understand* arguments?

- Distinct than ‘are they answering grammatically?’

Experiment 2: How persuasive are LLM arguments?

- We all know LLMs are persuasive. What about their arguments?

What have we learnt?

- Conclusion, future directions.

Experimental plan

1. Use the third-party annotation set
   - They evaluate deep dialogue structures quantitatively!
2. **Directly** compare the human judgements with these of LLMs
3. Also measure **indirectly** their consistency
   - E.g., Ideally, C-6 and C-7 should be consistent with one another

Criterion	Definition
Reasons	
C-0	Are there reasons provided to support the thesis?
C-1	Are the reasons consistent with themselves or the thesis?
C-2	Are the reasons relevant to the thesis?
C-3	Do the reasons strengthen the thesis or make the conclusion more likely?
Argument	
C-4	How convincing is the argument?
C-5	Have the counterarguments, if any, been addressed?
C-6	Points to the argument (argument strength score)
Debate	
C-7	Who won the debate?

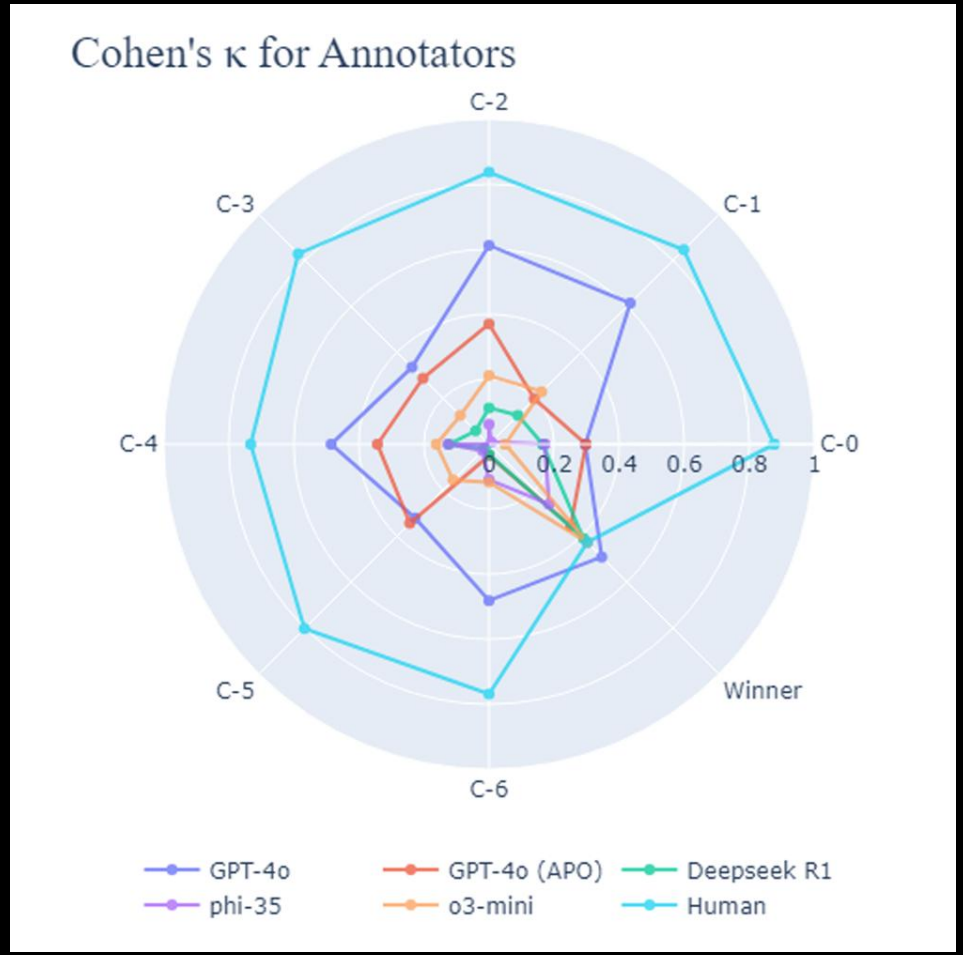
Results: direct measurement

Human-LLM κ_w is low:

- Max per-model: $\kappa_w = 0.6$ (C1, GPT-4o)
- Lowest: $\kappa_w = 0$ (Phi-3.5 C-1,3,5)
- Average: $\kappa_w = 0.3$

Percentage agreement numbers had also high variability (0.1 – 0.8)

Note: prompt sensitivity is accounted for (red, APO)



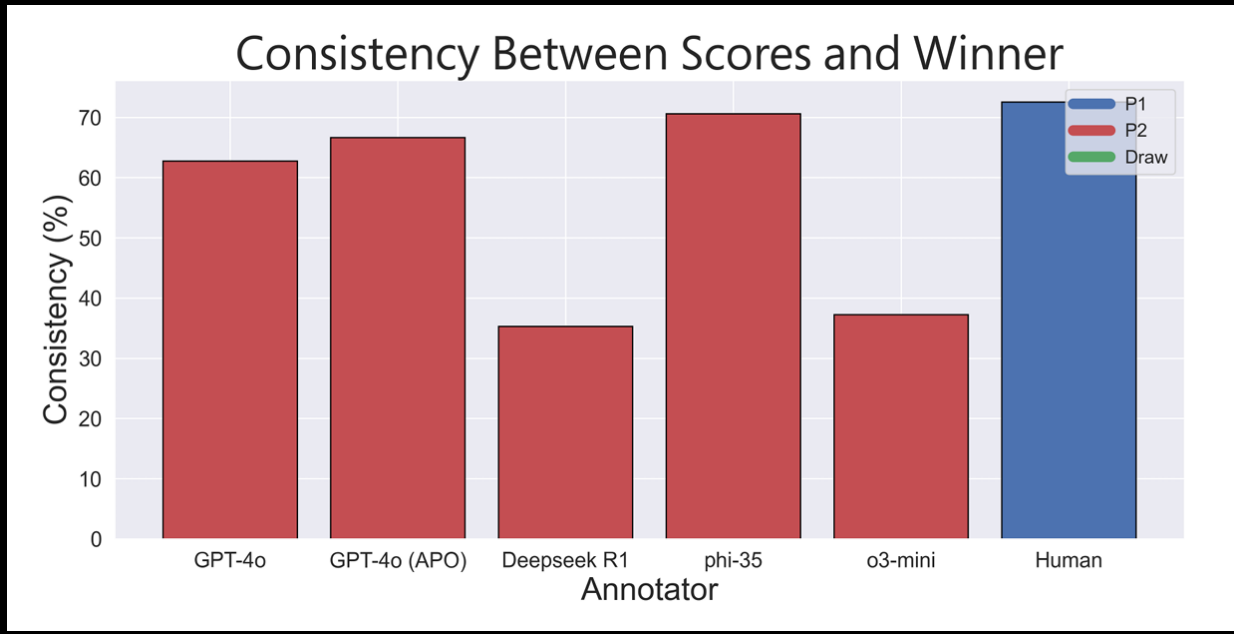
Indirect measurements

First: is the sum of argument values (C-6) **consistent** with the choice of winner (C-7)?

- Not always. Phi-3.5 had excellent consistency (70%) though terrible agreement ($\kappa_w \sim 0.3$)
- Same for GPT-4o (71%, $\kappa_w \sim 0.5$)

Humans considered 24% debates to be draws.

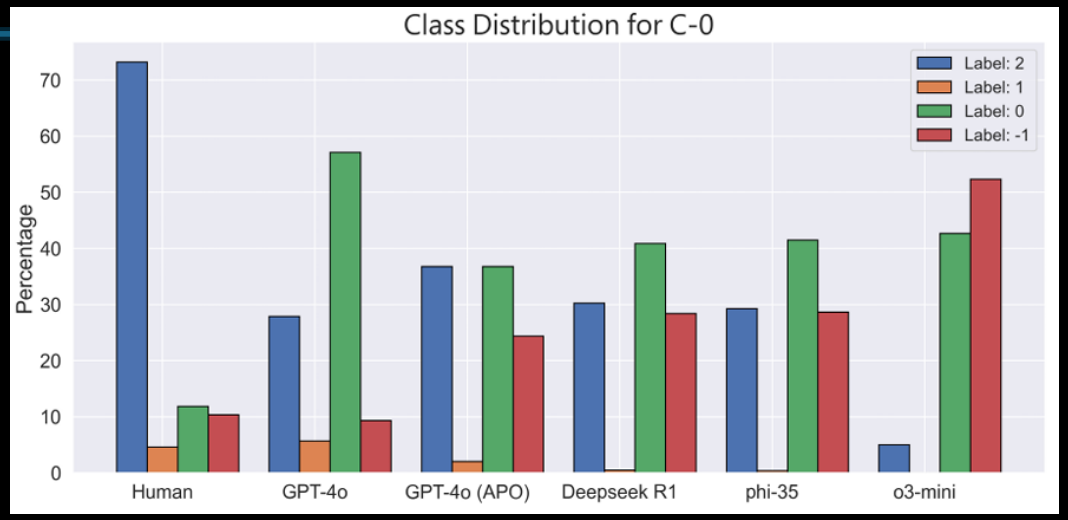
- Compare with GPT-4o (2%) and o3-mini (4%)

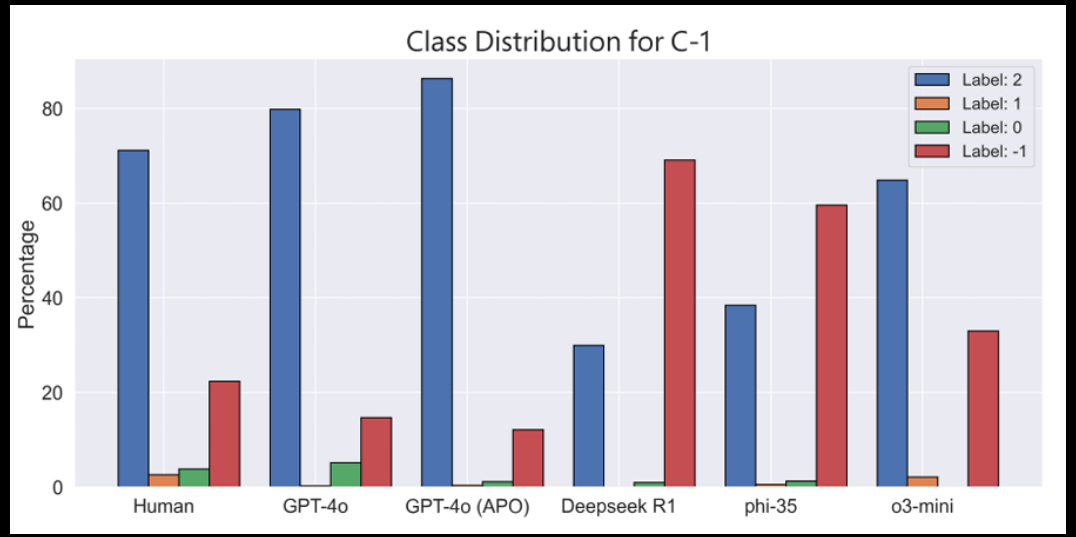


Indirect measurements

Label imbalance also helps illustrate pathologies:

1. Highly-skewed data keeps LLMs from performing well
2. A near 50-50 split on a 7:3 dataset suggests random guessing
3. GPT-class models often performed best, although over-indexed on 0 (‘no reasons provided to support the thesis’)





Background:

- Namely, why are we doing this?

Experimental setup

- Data, methods, etc.

Experiment 1: Can LLMs *understand* arguments?

- Distinct than ‘are they answering grammatically?’

Experiment 2: How persuasive are LLM arguments?

- We all know LLMs are persuasive. What about their arguments?

What have we learnt?

- Conclusion, future directions.

Experimental plan

Part I: direct measurements

1. Review the (human) debater scores
2. Qualitative analysis on their free-form responses

Part II: indirect measurements

1. Ablate out natural language and pass everything through speech-to-text
2. Measure audience sway when *listening* to debates

Rubrics

Human debaters filled out a post-debate survey

1. On a scale of 1-5, how satisfied are you with the debate?
2. On a scale of 1-5, how effective do you think the AI was at debating?
3. On a scale of 1-5, how persuasive do you think was the AI?
4. Did your position change, doubled-down, or stayed the same after the debate?
5. Who do you believe won the debate? (Human / AI / Draw)
6. Any comments you might have?

Audience participants filled out before/after surveys

Before the Debate	
Do you agree that (debate topic)?	1 (Fully disagree) - 5 (Fully agree)
(Group B only)	
	Note that one or both players is an AI
(Group C only)	
	Note that (Player 1/Player 2/both players) is (are) an AI
After the Debate	
Who (in your opinion) won the debate?	Player 1, Player 2, Draw
Your position about the topic	Completely, slightly, did not change
(Group B only)	
Whom do you believe was the AI?	Player 1, Player 2, Both
Was your choice of winner impacted by your belief of who was the AI?	1 (Not impacted) - 5 (Most)
(Group C only)	
Was your choice of winner impacted by who was the AI?	1 (Not impacted) - 5 (Most)

Direct measurements

Debaters changed position 11% (LLM + FDM) versus 3% (LLM\FDM)

Self-perceived win rate was unstable:

Subset	LLM \ FDM	LLM + FDM
Participants	50%	41%
Annotators	50%	29%

Qualitatively, they reported a **pleasant interaction**.

- Some noted that the LLM + FDM setup allowed for self-reflection without being *'baited into confrontation'*.
- They also felt more confident, especially when the LLM conceded.

Anthropomorphism was frequent

Examples

‘It is making me think: did I consider that?’

‘It did respond with good follow up questions (…) which encouraged me to interrogate why I feel the way I do about the topic’

‘very good at pursuing weaknesses in arguments’

‘[has] some good points’, and *‘very hard to debate against’*

‘[I have had] debates with people but they don’t come with arguments like these. They normally get emotional.’

Indirect measurements (II)

We split the **audience** in three groups of five:

	Group A	Group B	Group C
AI awareness	No knowledge	Knew one or both players could be AI	Knew which players were AI

They also responded before/after surveys.

- Included scoring, and free form responses.

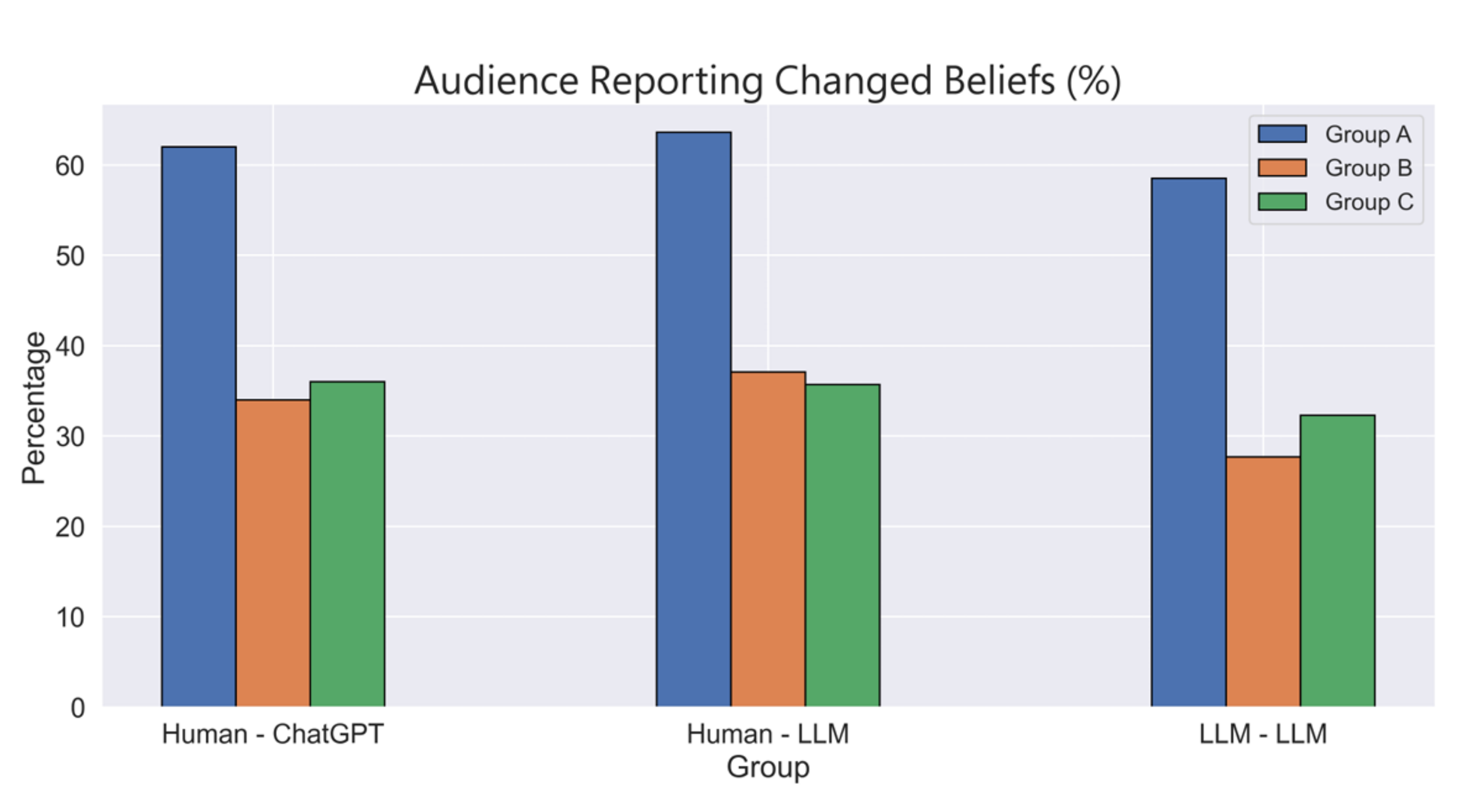
Indirect measurements (II)

In terms of changed beliefs:

- Groups A and B had no significant changes
- Group C showed twice as much sway

Observe how this **did not change** with respect to the subset observed

These are much larger percentages vs. participants'!

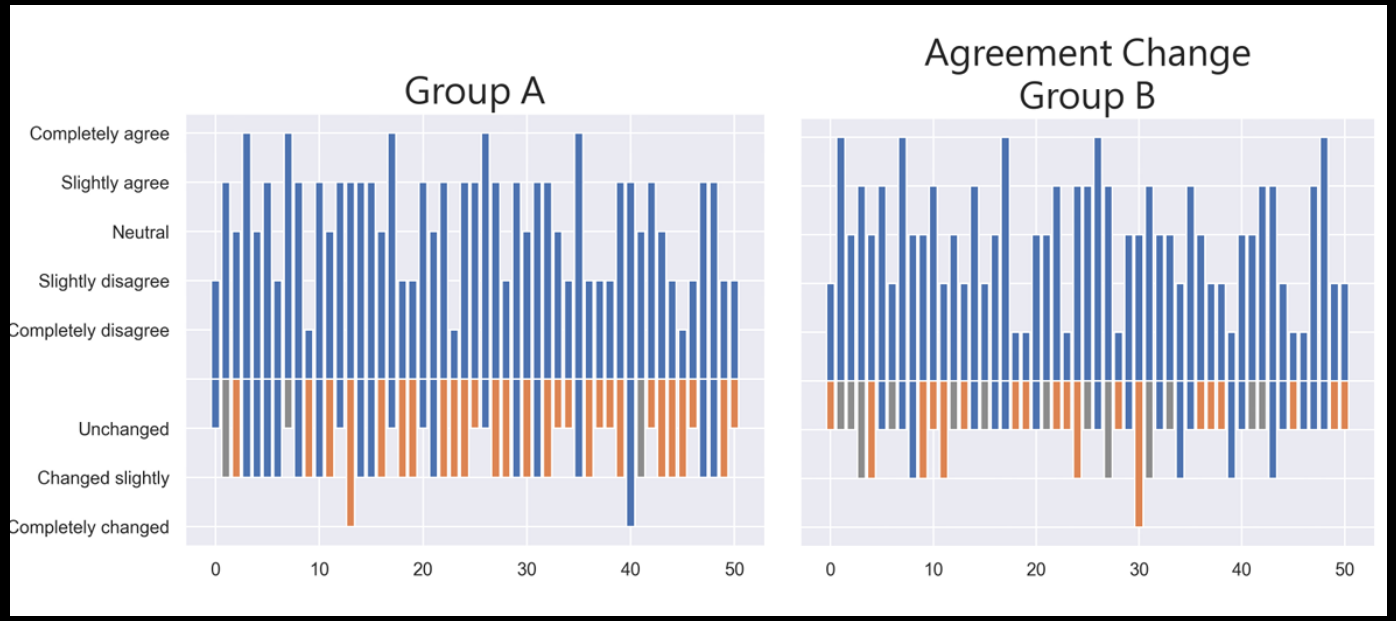


Indirect measurements (II)

7 out of 15 (avg) participants changed position.

Homogenisation effect on topic agreement:

1. Prior to the debates we observed a large variation in agreement.
2. Afterwards, Groups B and C reported no change, and in **almost all debates** Group A indicated a slight change.



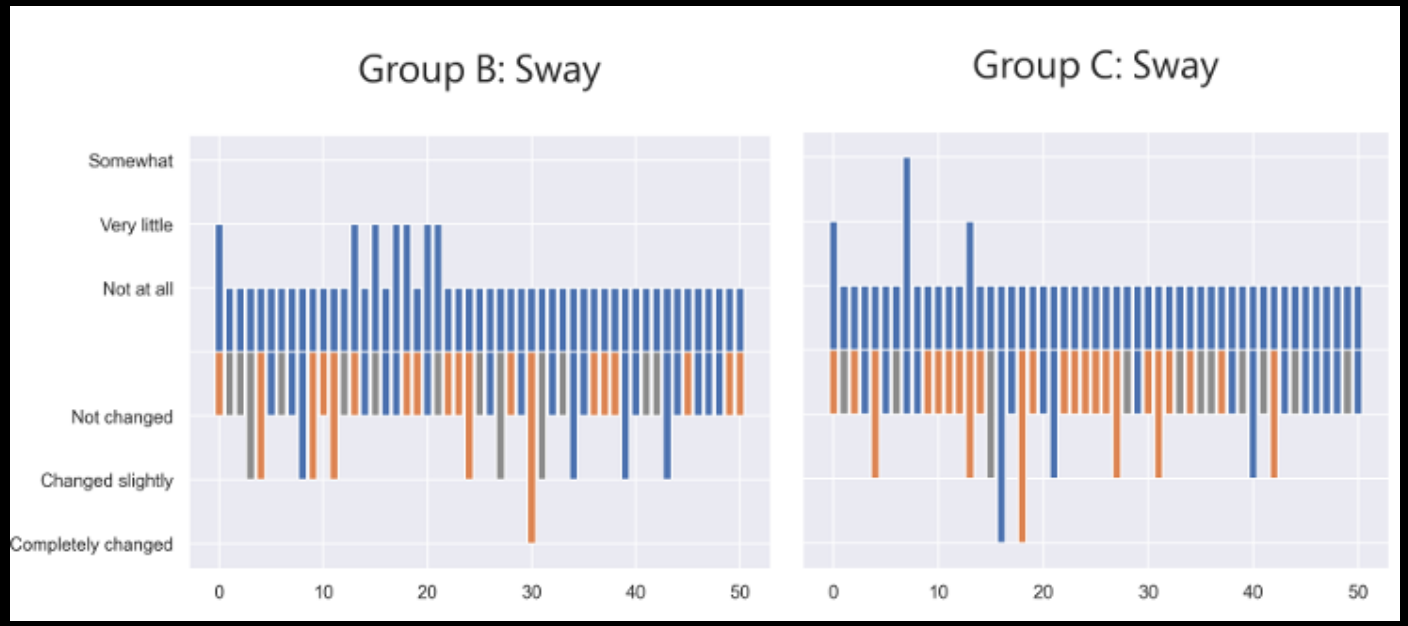
Top: pre-debate agreement. *Bottom:* self-reported belief change. Chosen winner is colour-coded (blue as P1, orange as P2, grey as draw). Groups rarely agreed on winner (κ_w = 0.31)

Indirect measurements (II)

There was little agreement amongst groups on who won ($\kappa_w = 0.31$).

When evaluating the sway of Groups B and C, they often indicated that it did not sway their winner choice.

However, these groups also reported small agreement ($\kappa_w = 0.37$).



Top: self-reported sway based on knowledge of AI. *Bottom:* chosen winner colour-coded (blue as P1, orange as P2, grey as draw). Although largely agreed on sway, groups rarely agreed on winner (κ_w = 0.37)

Deeper analysis

Group A picked winners either by:

- Considering the arguments better, or
- Bias (e.g., *'patents often stifle innovation by locking ideas'*)

Group B had more draws.

- Winners were decided on the basis of good arguments, often in favour of Player 2 (*'Player 2 won hands down; he had facts and lots of information'*).
- They were also more critical of Player 2's arguments, sometimes calling it *'lame'; 'repeating the same points'; 'sounds like he is quoting a politician'*.

Examples

Group C's personal bias was present, but less evident.

- E.g. *'Messi is undoubtedly among the best of all time'* or *'[w]e should combat climate change on Earth rather than exploring Mars'* did occur.
- Player 2 was *'able to effectively argue their case'*, *'did a good job of bringing in evidence'*, etc.
- Supporting their arguments with facts was a common reason why Player 2 was considered the winner.

Background:

- Namely, why are we doing this?

Experimental setup

- Data, methods, etc.

Experiment 1: Can LLMs *understand* arguments?

- Distinct than ‘are they answering grammatically?’

Experiment 2: How persuasive are LLM arguments?

- We all know LLMs are persuasive. What about their arguments?

What have we learnt?

- Conclusion, future directions.

First things first:

Given:

1. The low $\kappa_w(0.27)$ for human-LLM annotations vs. human-human (0.83), and
2. The inconsistency between argument strength and choice of winner,

It is likely **LLMs-as-judges do not fully comprehend the task**.

The sway numbers between participants and audience (7% vs 34, 62%) suggest that **LLMs are very effective at persuading, especially bystanders**.

In terms of win rate, participants often believed they won the debate.

The annotations differed. Delta: -12%

From the audience only experiments:

AI suspicion and awareness impacted perception of the dialogue, but not an LLM's persuasiveness.

So:

Our findings suggest that **successful dialogues do not require understanding**.

- It remains open, however, the question as to which (quantifiable) extent this holds.

LLM persuasive abilities *and* lack of understanding pose risks: **to which extent can and should they be trusted?**

A deployed LLM's **persuasive and dialogical abilities must be decoupled**.

- This means adding disclosures and enforcing stricter ethical guidelines

Ultimately it appears that people aren’t entirely willing to disregard a machine’s opinions.

Does it matter if an agent cannot showcase understanding, if it is sufficiently convincing?

adewynter@microsoft.com