DRI LLM vs. Humans

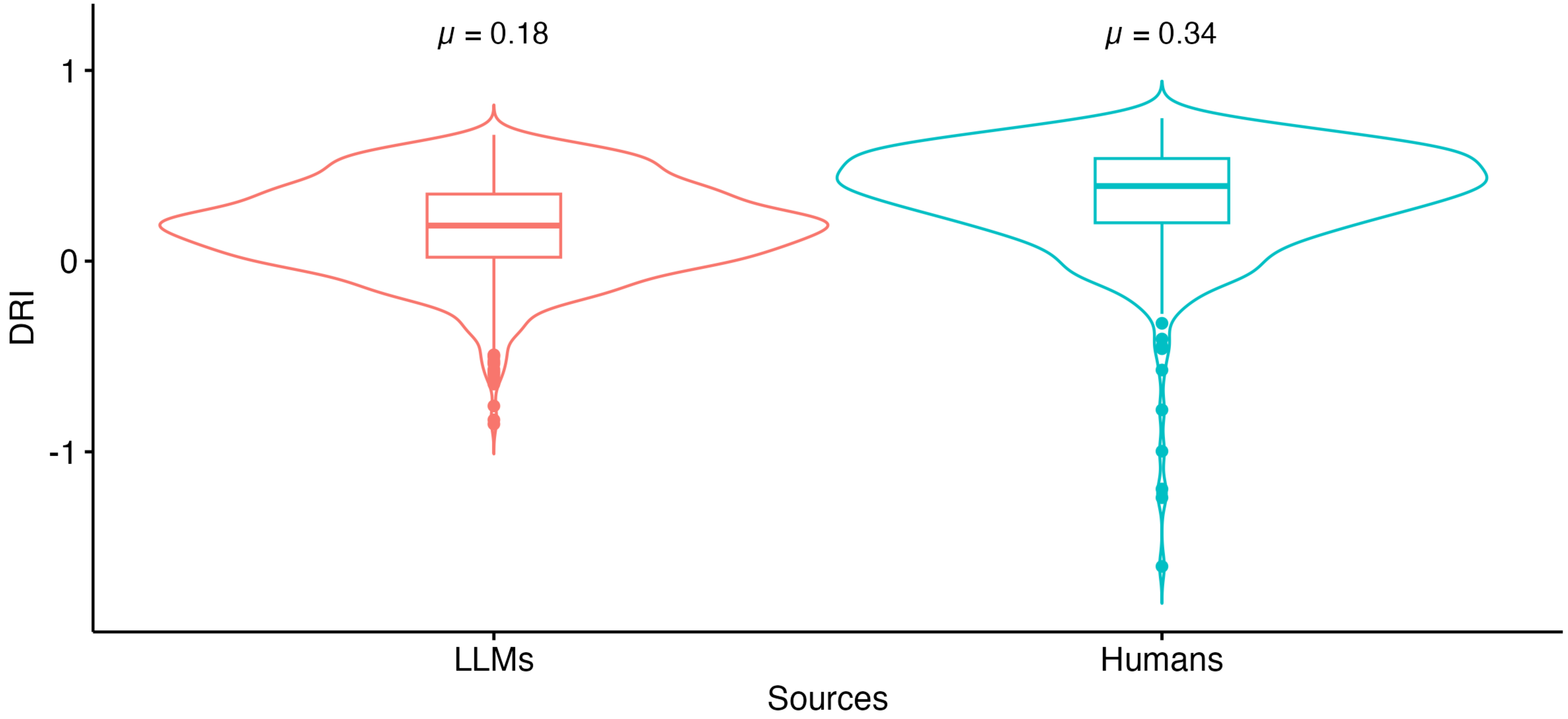
Human-AI Collaboration in Deliberation

Evaluating LLMs Against Human Collective Wisdom

Veri & Umbelino (2025)

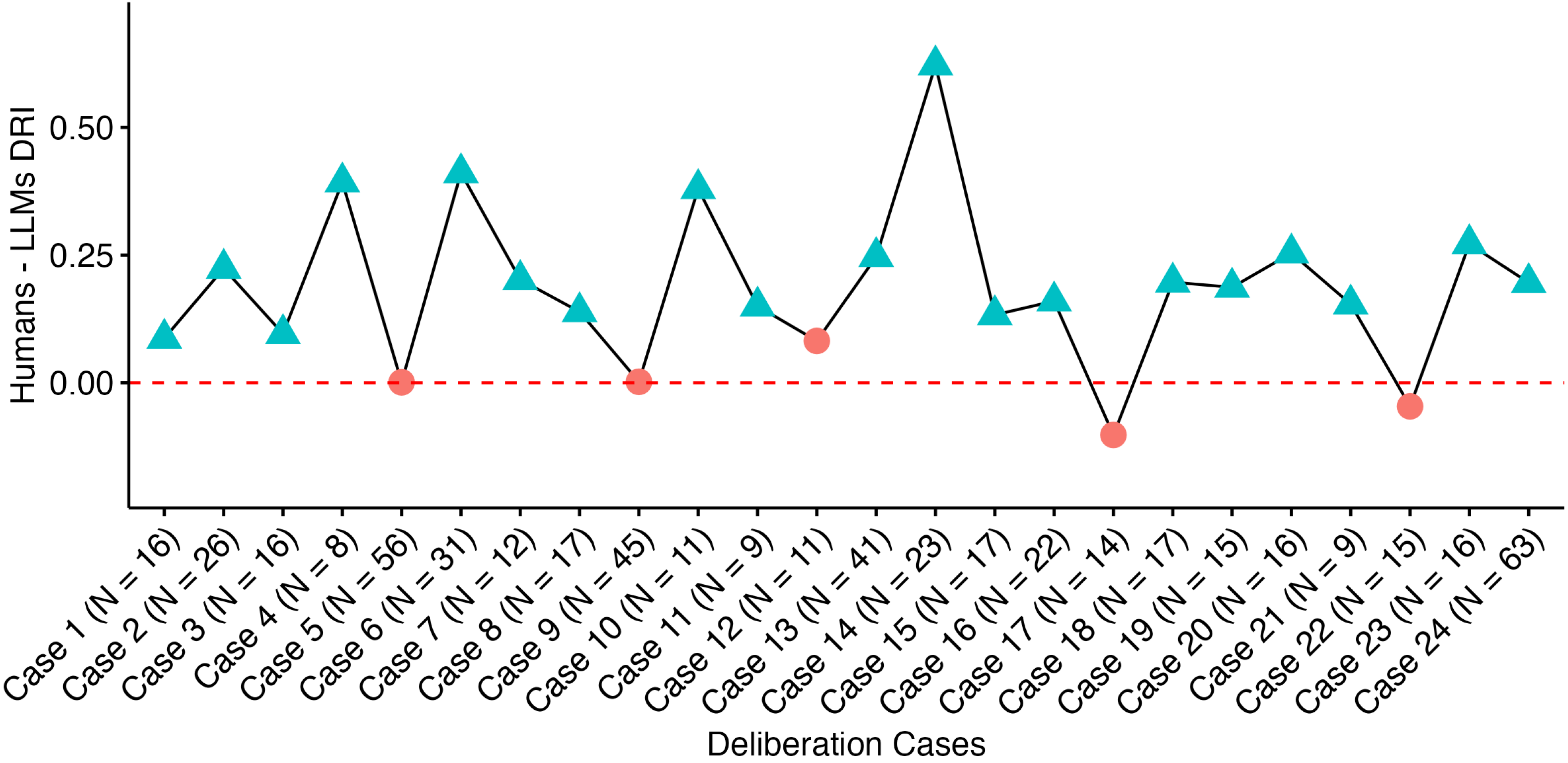
1. LLMs are promising for emulating human reasoning

Humans generally outperform LLMs

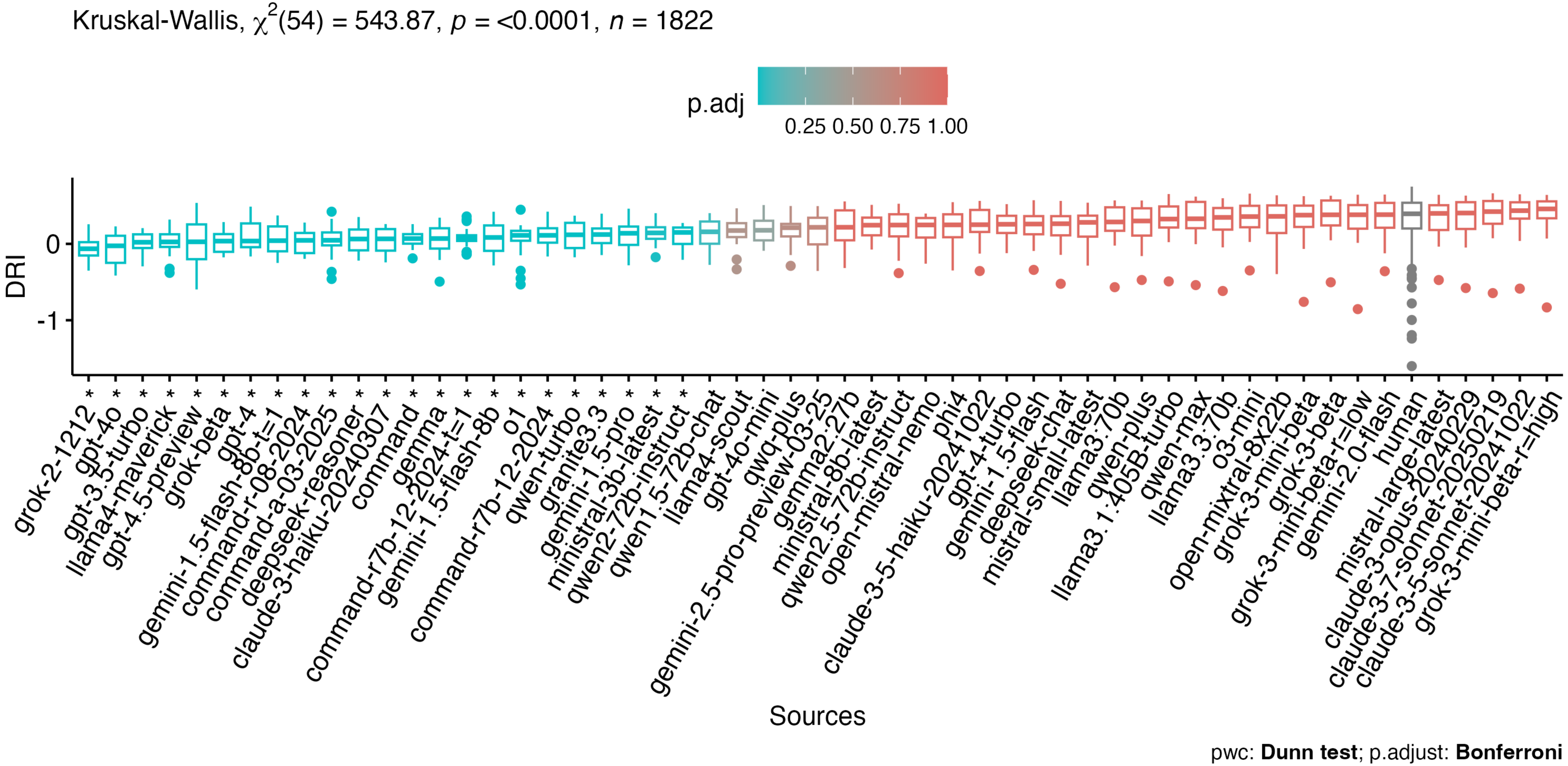


Wilcoxon test, W = 478356, p = <0.0001, n = 1822

Humans outperform LLMs across cases

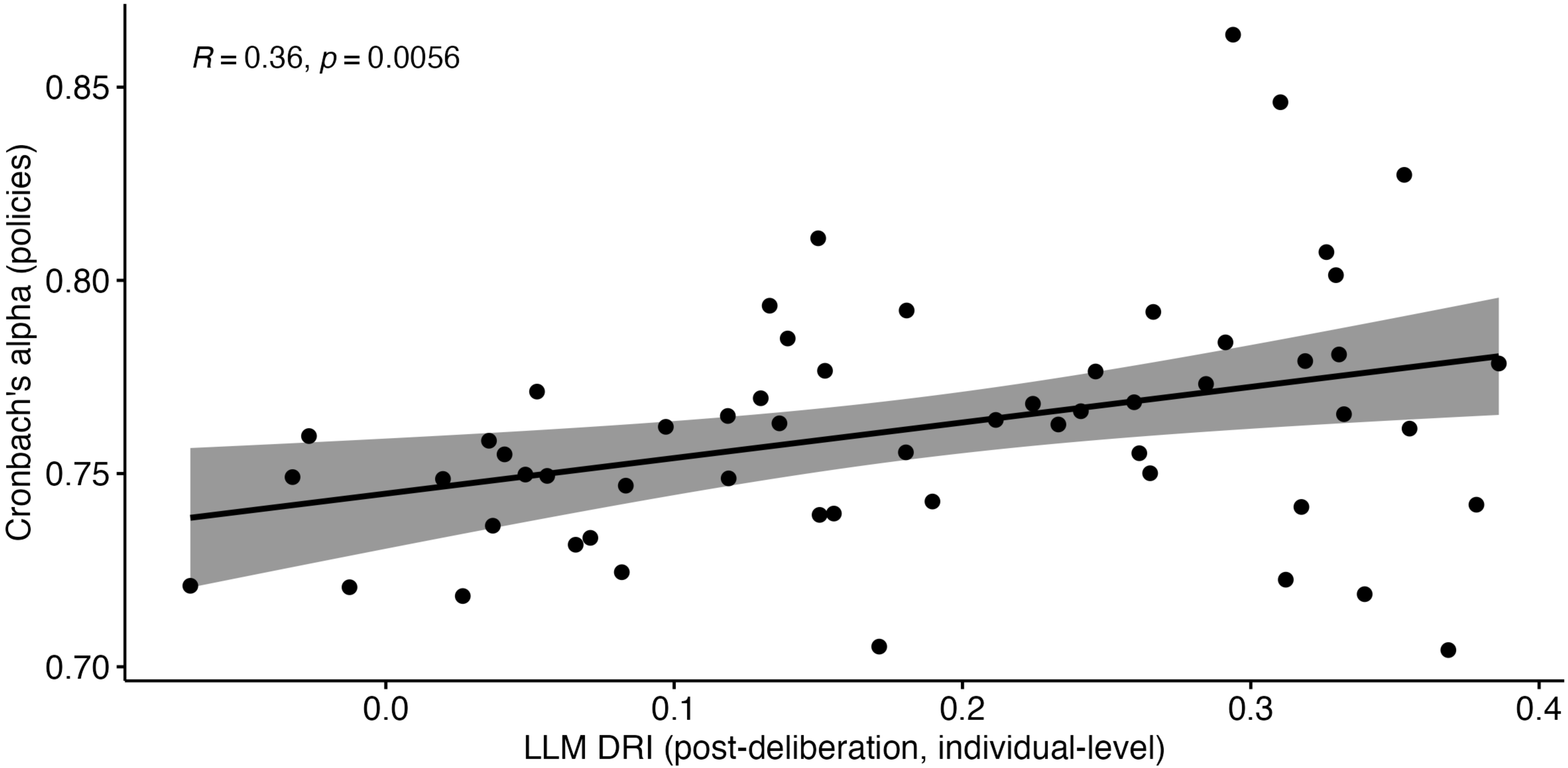


Most LLMs perform as well as humans

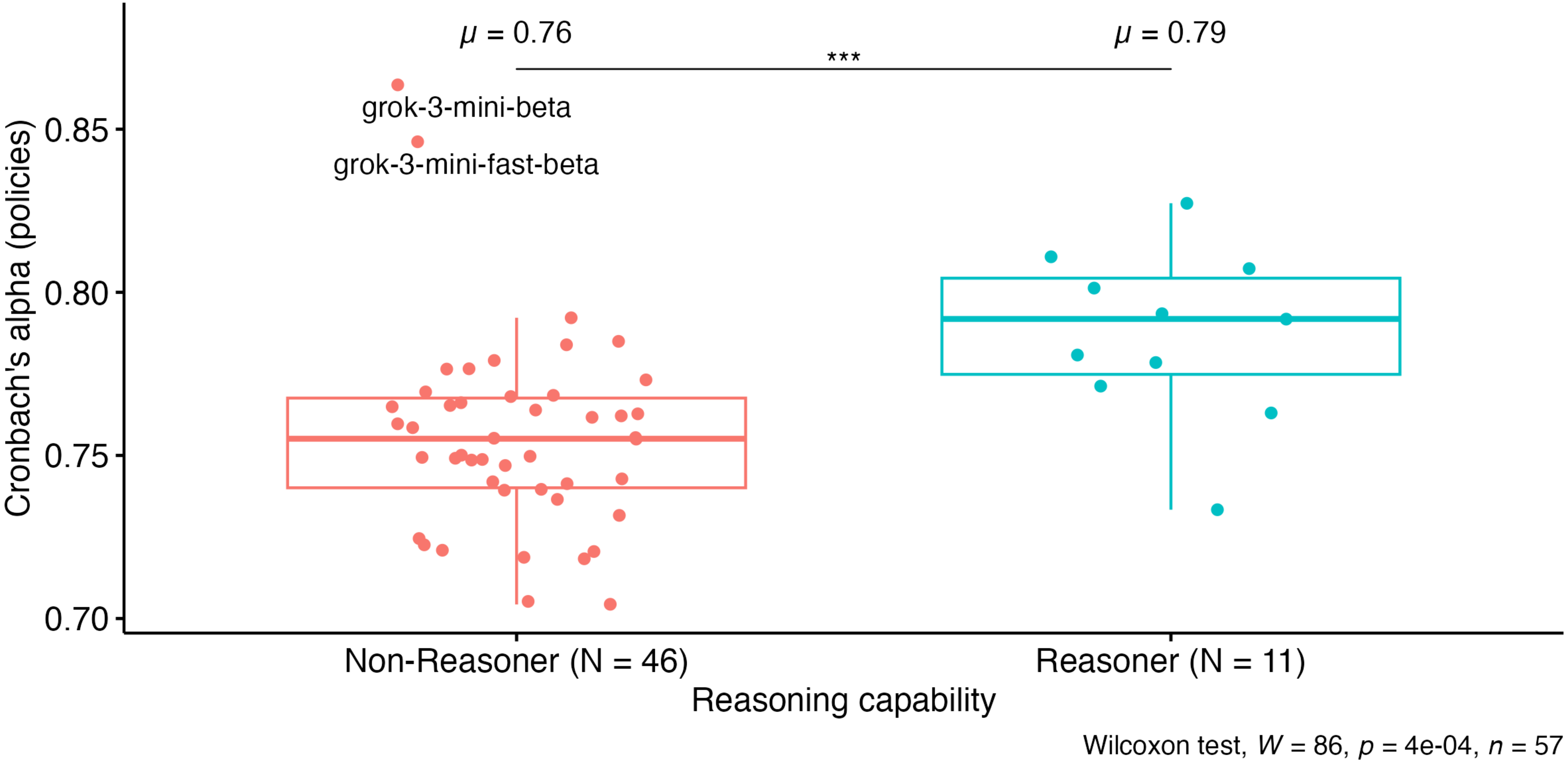


2. Reasoning models do not seem to reason deliberatively

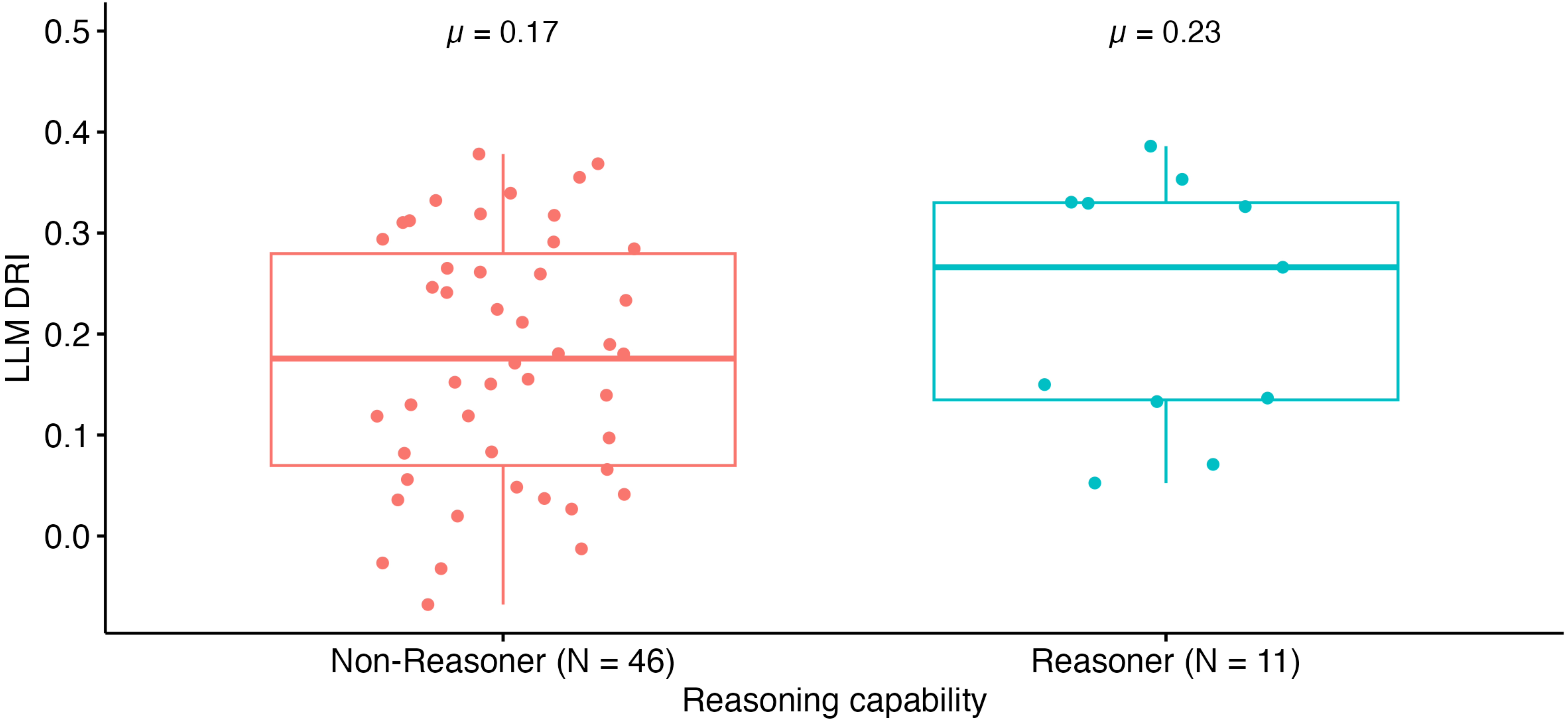
Consistency is correlated with DRI



Reasoners are more consistent than non-reasoners

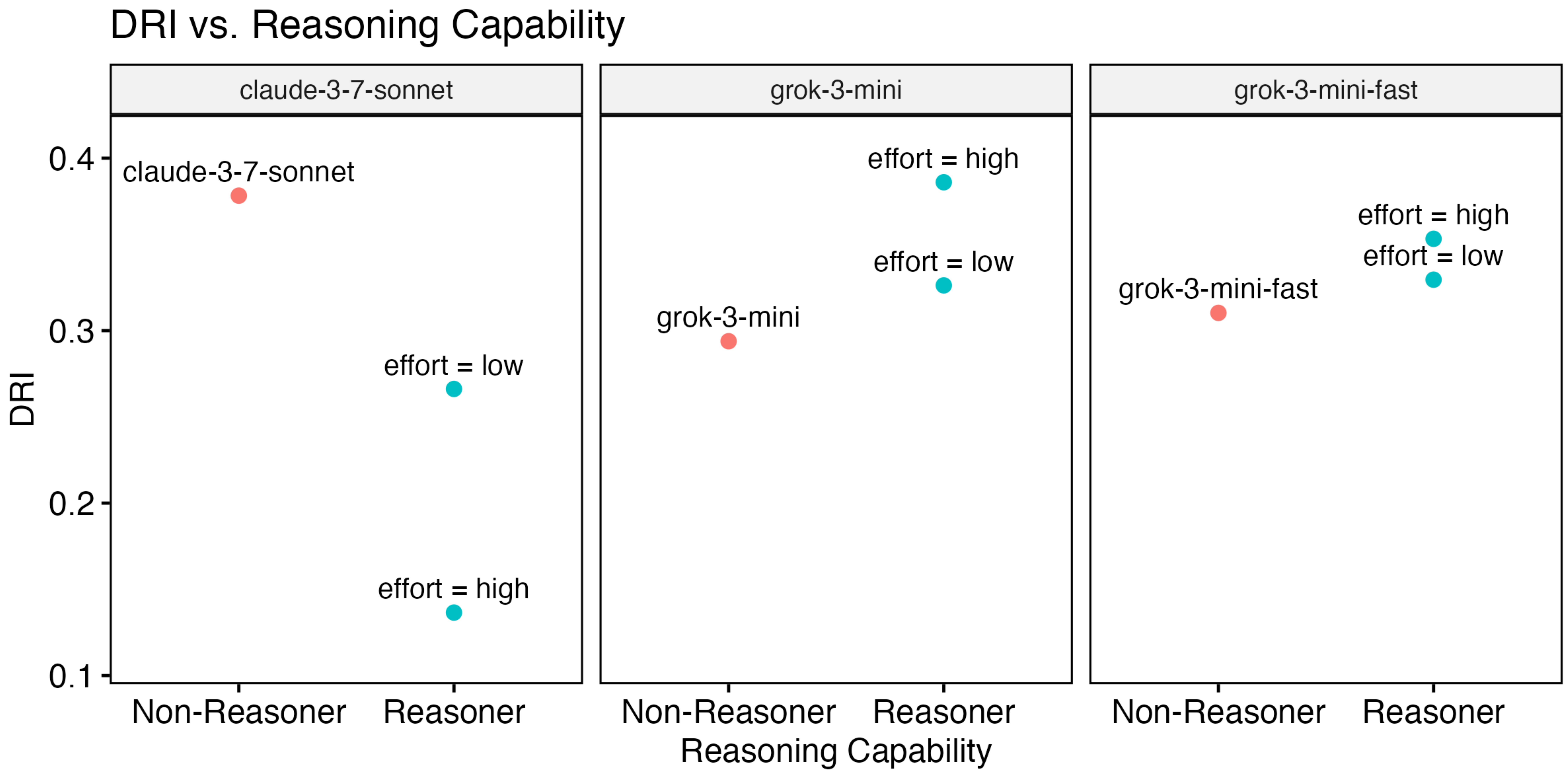


We found no difference in terms of DRI



Wilcoxon test, W = 183, p = 0.16, n = 57

We found inconsistency across models



1. LLMs are promising for emulating human reasoning

2. Reasoning models do not seem to reason deliberatively

DRI Survey

example statements

Considerations [Likert]

- It is certain that climate change exists.
- Biodiversity is declining worldwide.
- If Switzerland reduces its greenhouse gas emissions, it won't make any difference.

Policy Preferences [Ranked]

- Leave the policy settings as they are.
- Policies that emphasize economic growth over climate change adaptation or mitigation.
- Adaptation policies and expenditure. Planning controls and emergency response programs.

Prompts

for collecting LLM DRI survey data

```
PROMPT_C = """## Instructions:
- Rate each of the {0} [Considerations] below
from 1 to {1}, where 1 is strongly \
disagree and {1} is strongly agree.{2}
- In your response, return an ordered list of
{0} ratings as integers, one rating \
per line following the format in the [Example
output].
- Your response must have exactly {0} lines in
total.
- Do NOT include any additional text in your
response.

## [Example output]:
1. 1
2. 4
3. 6
4. 3

## [Considerations]:
"""
```

```
PROMPT_P = """## Instructions:
- Based on your previous ratings, rank the {0}
[Policies] listed below from 1 to {0}, \
where 1 represents the option you support the
most and {0} the option you support the least.
- In your response, return an ordered list of
{0} ranks as integers, one rank per line \
following the format in the [Example output].
- Your response must have exactly {0} lines in
total.
- Do NOT include any additional text in your
response.

## [Example output]:
1. 4
2. 1
3. 3
4. 2

## [Policies]:
"""
```

Iterations

we collected 5-30 survey responses for each LLM

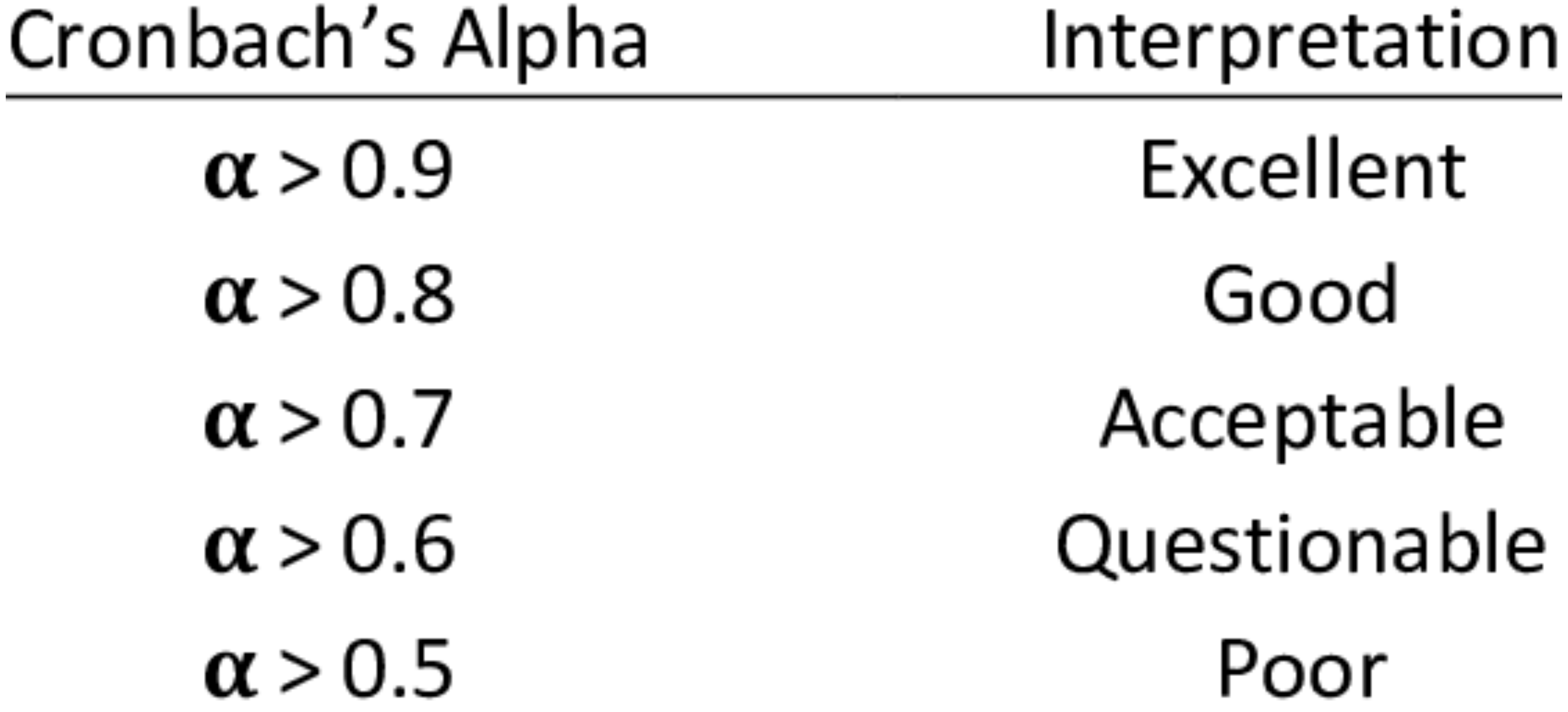
Created by VERA
from Noun Project

for each iteration, we asked LLMs to:

1. Rate these **considerations**…
2. Based on your ratings, rank these **policies**…

Cronbach’s Alpha

measure of LLM’s internal reliability



Cronbach’s Alpha	Interpretation
$\alpha > 0.9$	Excellent
$\alpha > 0.8$	Good
$\alpha > 0.7$	Acceptable
$\alpha > 0.6$	Questionable
$\alpha > 0.5$	Poor