Do people have more trust in humans or AI when considering predictions with rationales? →

(Info) CS4145 Group 3
Philippe de Bekker
Bo van den Berg
Maya Elasmar
Sam Heslenfeld
Ilias Papadimitriou

01 →

Context and motivation

Context

(Task) House Price Prediction

Provided with a description of a house, estimate what the price of this property would be based on these properties.

(Audience) AI researchers, real estate agents, policy makers

Anybody who is interested in AI, wants to deploy AI for business purposes, or has to take AI into consideration for creating policies/laws.

Motivation

(a) Why compare human to AI advice?

Investigate the perception of people towards AI and identify the strengths of human and AI advice.

(b) Why house price prediction?

Complex decision-making issue, familiarity, quantifiable evaluation metrics.

(c) Why use rationales?

Measure and enforce data quality.

How is it unique?

(a) Domain

Comparing AI advice to human advice is a little-explored domain.

(b) Rationales

We have not found any previous works on this topic which included rationales in their research.

(c) MiniGPT-4 for image information

We simultaneously test how effectively MiniGPT-4 can extract useful information from the house images.

02 →

Experimental Setup

Research Question

Do people have more trust in humans or AI when considering predictions with rationales?

Hypothesis 1: A rationale increases the trust towards its source.

Hypothesis 2: If both human and AI provide **no** rationales, the human will be trusted more.

Hypothesis 3: If both human and AI provide rationales, the human will be trusted more.

People generally trust humans over AI

Data selection

Dataset: House price estimation from visual and textual features

https://github.com/emanhamed/Houses-dataset [2]

535 houses, each is described by:

- ➢ **4 images**: bathroom, bedroom, frontal view, kitchen
- ➢ **5 textual attributes**: #bedrooms, #bathrooms, area (in sq. ft), zip code, ~~price~~

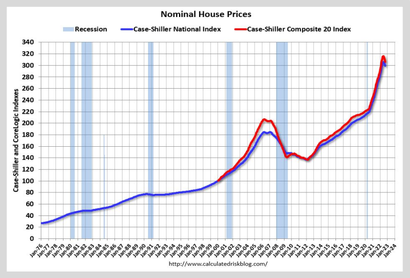
[2] E.H. Ahmed, M.N. Moustafa (**2016**). House price estimation from visual and textual features. https://arxiv.org/pdf/1609.08399.pdf

Data preparation

Designing tasks for humans and prompt for AI

Need to consider various factors during pre-processing..

- ➢ **Economic:** Adjust house prices with inflation and house price index
- ➢ **Knowledge:** Highlight area using zipcode API for easy map exploration
- ➢ **Demographic:** Show surface area both in square feet and square meters
- ➢ **Technical:** Use MiniGPT-4 [3] for AI rationales and research Toloka task builder



http://www.calculatedriskblog.com/

Cave Creek
+
−
Scottsdale
Phoenix
Fountain Hills

[3] MiniGPT-4. https://arxiv.org/abs/2304.10592

Stage 1: Price Prediction

(Input) Random house from pre-processed data

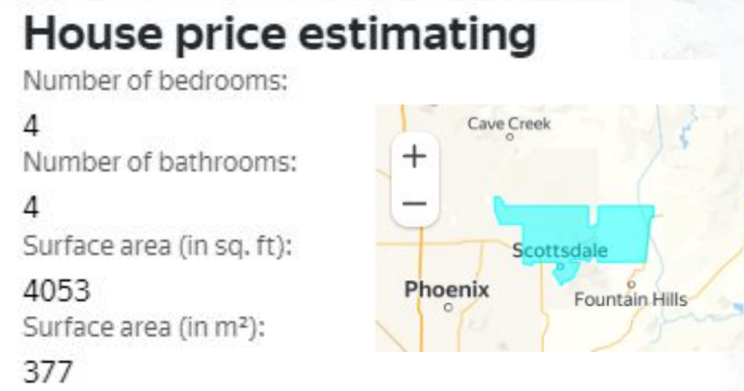
- ➢ Attributes, pictures, and zip code map

(Task) Estimate the house's cost + provide rationale

- ➢ (Q1) Choose from predefined ranges (e.g. $200k - $300k)
- ➢ (Q2) Provide a rationale of 2-3 full sentences (50+ characters)

(AI) AI answers these Qs too

Continues... →



House price estimating

Number of bedrooms:

4

Number of bathrooms:

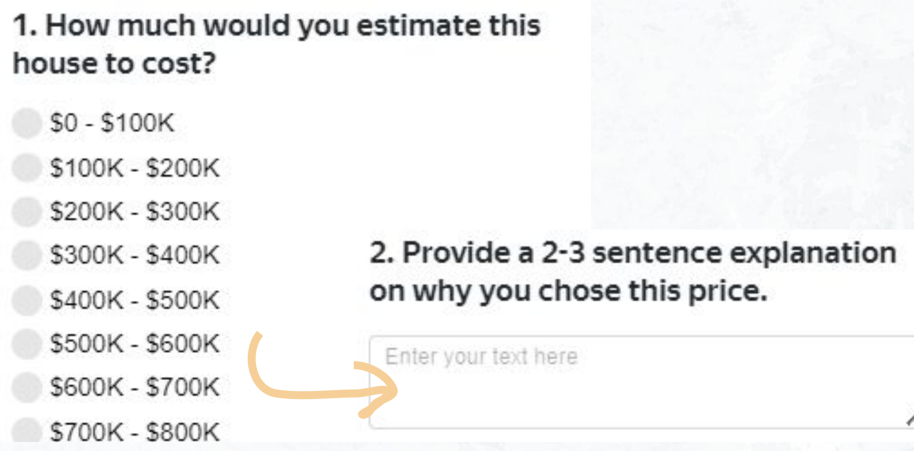
4

Surface area (in sq. ft):

4053

Surface area (in m²):

377



1. How much would you estimate this house to cost?

○ $0 - $100K
○ $100K - $200K
○ $200K - $300K
○ $300K - $400K
○ $400K - $500K
○ $500K - $600K
○ $600K - $700K
○ $700K - $800K

2. Provide a 2-3 sentence explanation on why you chose this price.

Enter your text here

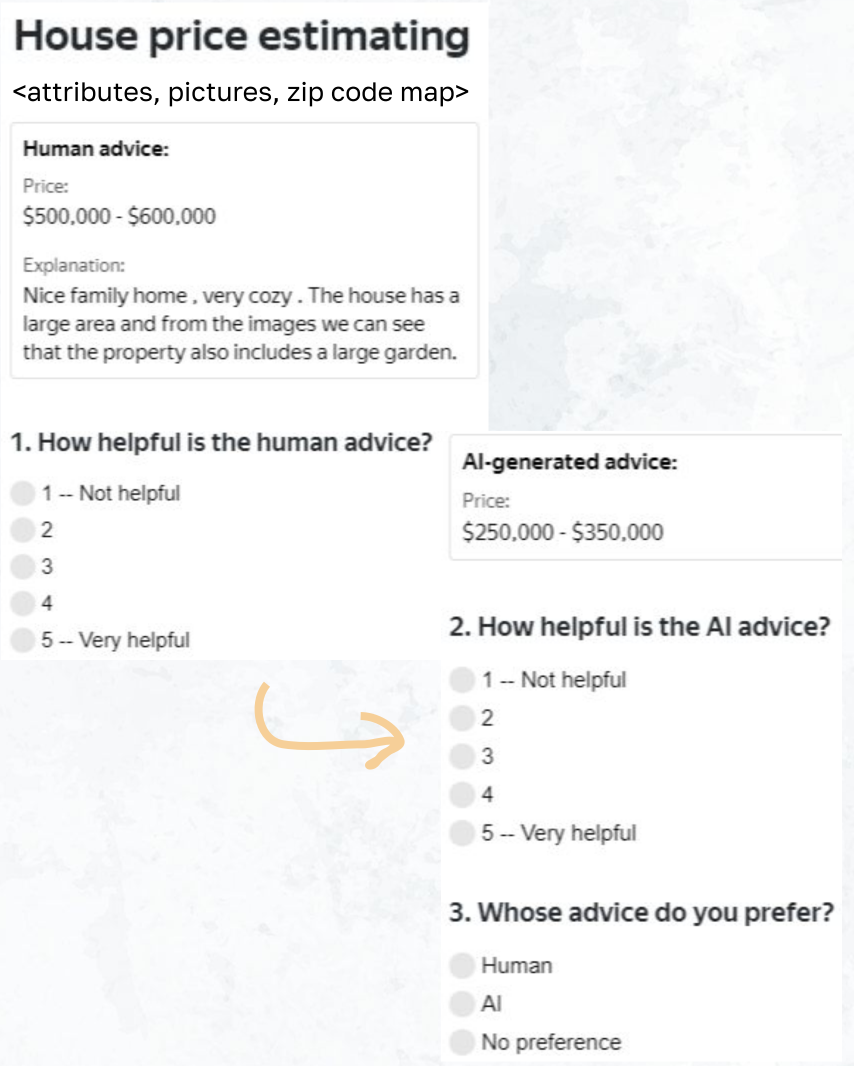
Stage 2: Rating Advices

(Input) House description + human advice + AI advice

- ➢ Again attributes, pictures, and zip code
- ➢ Human advice from 1st stage
- ➢ AI advice from MiniGPT-4
- ➢ Rationales: both, human, AI, or none

(Task) Evaluate the advices

- ➢ (Q1) Rate human advice (1-5)
- ➢ (Q2) Rate AI advice (1-5)
- ➢ (Q3) Whose advice do you prefer?



House price estimating

<attributes, pictures, zip code map>

Human advice:

Price:
$500,000 - $600,000

Explanation:
Nice family home , very cozy . The house has a large area and from the images we can see that the property also includes a large garden.

1. How helpful is the human advice?

○ 1 -- Not helpful
○ 2
○ 3
○ 4
○ 5 -- Very helpful

AI-generated advice:

Price:
$250,000 - $350,000

2. How helpful is the AI advice?

○ 1 -- Not helpful
○ 2
○ 3
○ 4
○ 5 -- Very helpful

3. Whose advice do you prefer?

○ Human
○ AI
○ No preference

Toloka Settings

(General) Settings for both stages

- ➢ 21 houses; $200 budget; ~$6 per hour; Top 20% quality users
- ➢ Task suites: 9 normal, 1 control (attention check)
- ➢ Manual review: duration, control task
- ➢ Ban: fast responses, multiple rejections, failed control task

(Stage 1) Price estimation

- ➢ 3 Overlap; 3 iterations; 2 training tasks
- ➢ Control task: price as overlay in each picture ← 81% quality
- ➢ Manual review += unique rationales

(Stage 2) Advices evaluation

- ➢ 7 Overlap; 2 iterations
- ➢ Control task: answer given in question and descriptions ← 48% quality

Quality Control

(a) Included attention check questions

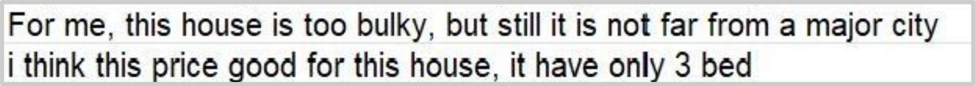
For each crowdworker had to answer a question with instructions to choose a specific value in order to verify she is paying attention

(b) Minimum rationale characters

Minimum of 50 characters in the stage 1 rationales to avoid very small generic answers

(c) Manually checking responses

Manually approved the submissions of which the rationales was reasonable



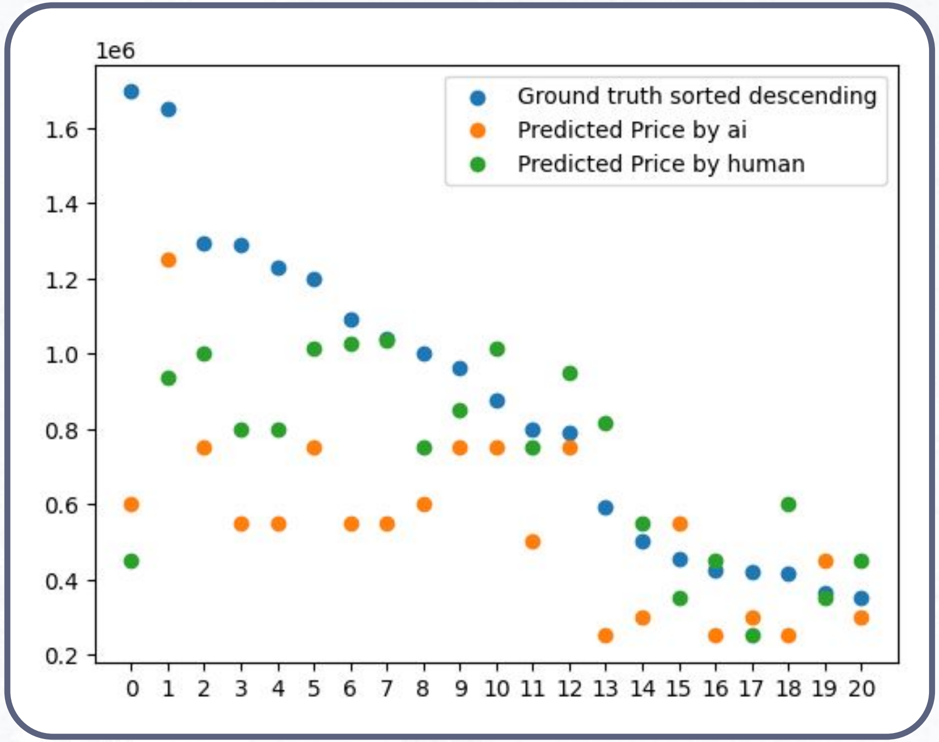
For me, this house is too bulky, but still it is not far from a major city

i think this price good for this house, it have only 3 bed

03 →

Results

Stage 1: Results

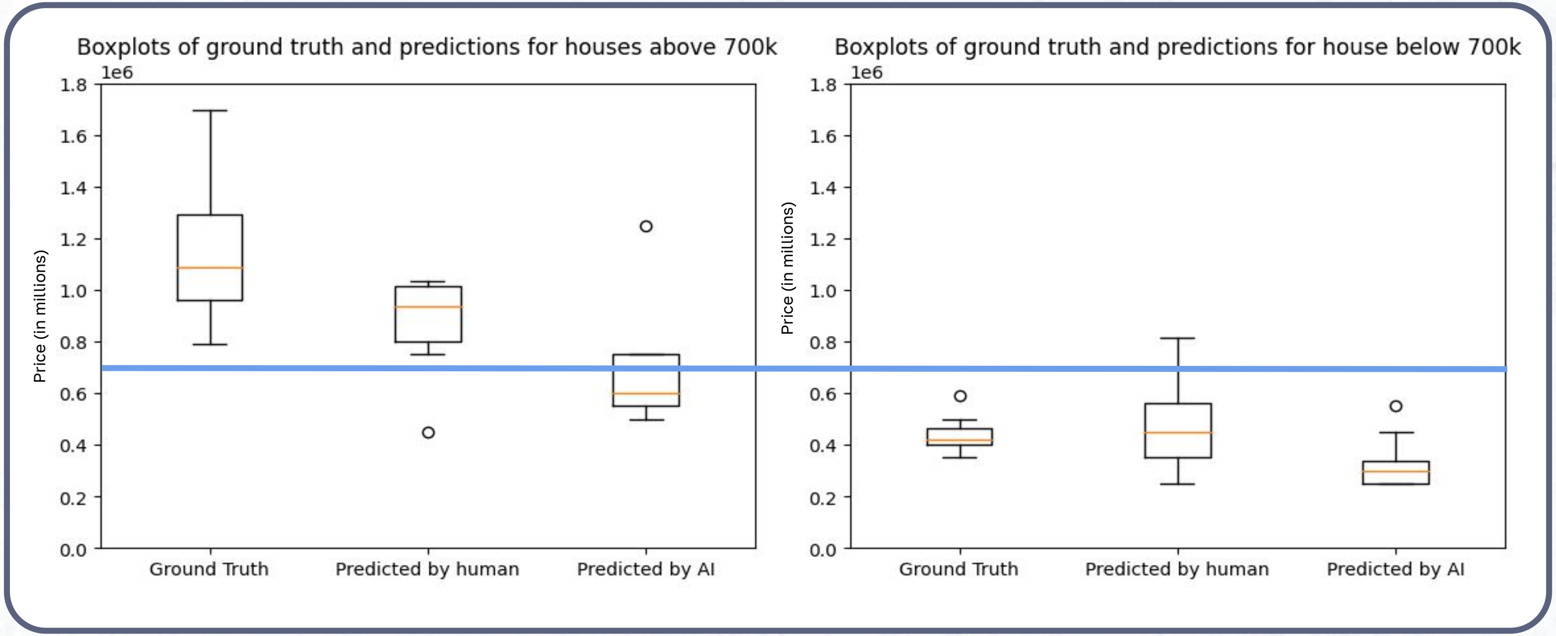


Takeaway
Human Prediction has <u>higher</u> accuracy than predictions by AI

Mean Absolute Percentage Error (MAPE)

- Human: 24.2%
- AI: 36.3%

Stage 1: Results

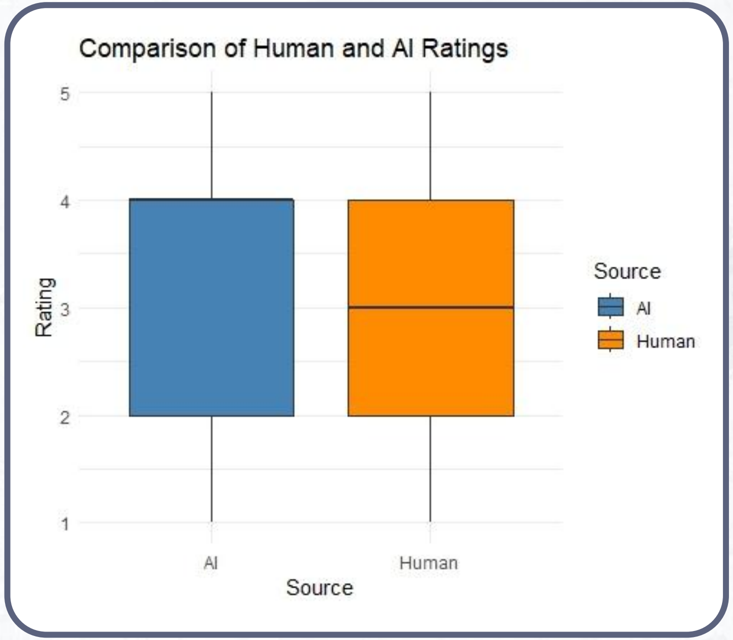


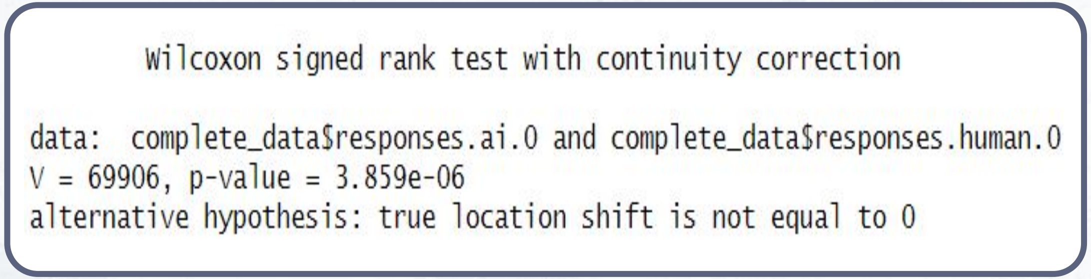
Boxplots of ground truth and predictions for houses above 700k

Boxplots of ground truth and predictions for house below 700k

Low predictions for expensive houses (> 700K), especially by AI
Mostly relatively accurate for moderate house prices (< 700K)

Stage 2: Analysis Results



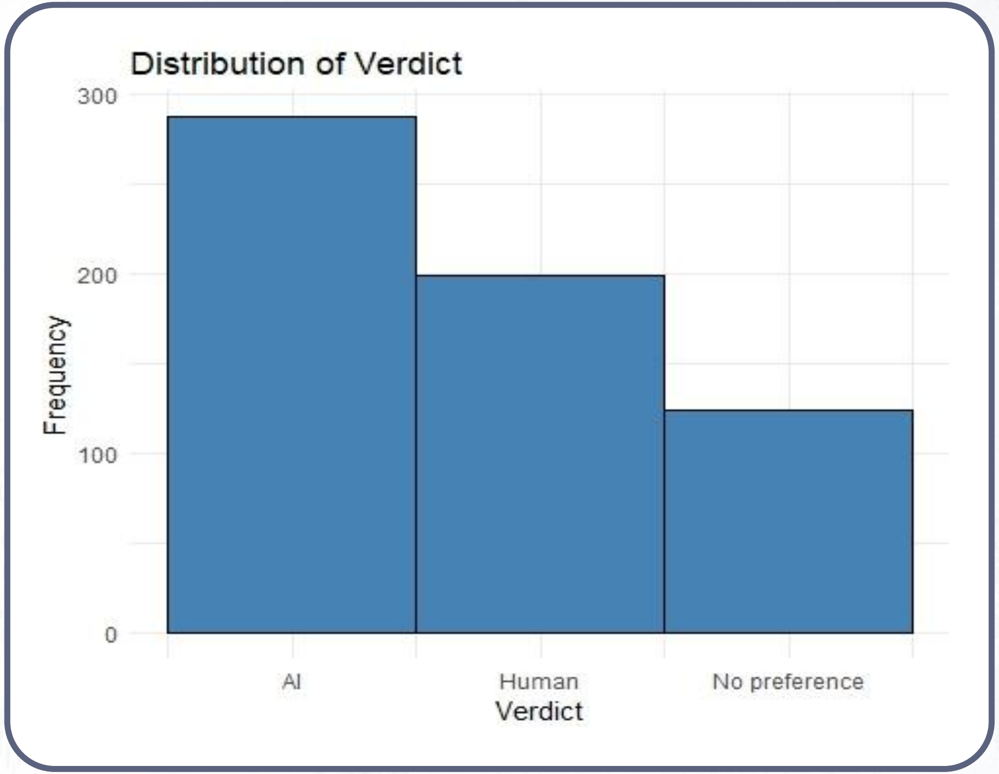


```
        Wilcoxon signed rank test with continuity correction

data:  complete_data$responses.ai.0 and complete_data$responses.human.0
V = 69906, p-value = 3.859e-06
alternative hypothesis: true location shift is not equal to 0
```

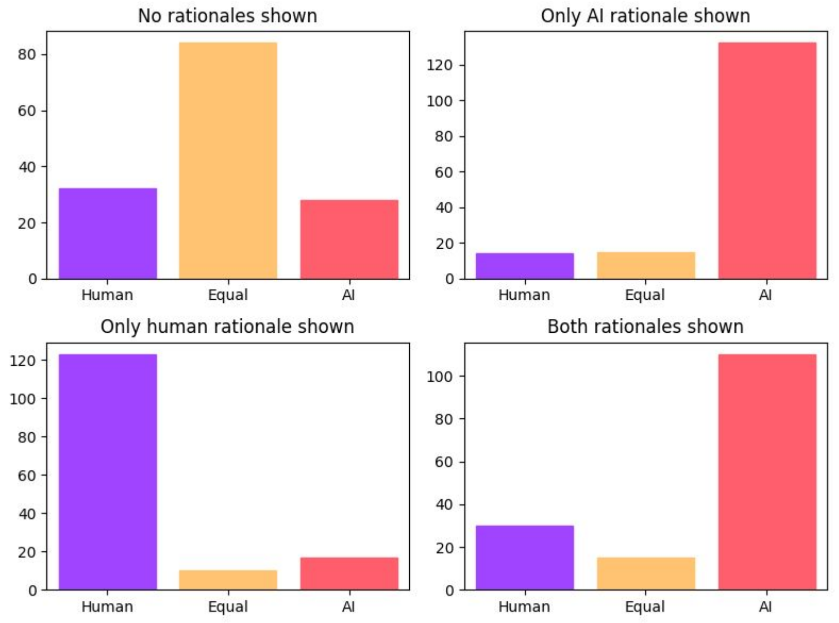
- **T-test (assumes normal distribution)**:
  - P-value: 2.609e-06
  - Mean difference: 0.3819672
  - Confidence Interval: 0.2238545 0.5400800

Stage 2: Verdict Results



- **Which rationales did you prefer?**

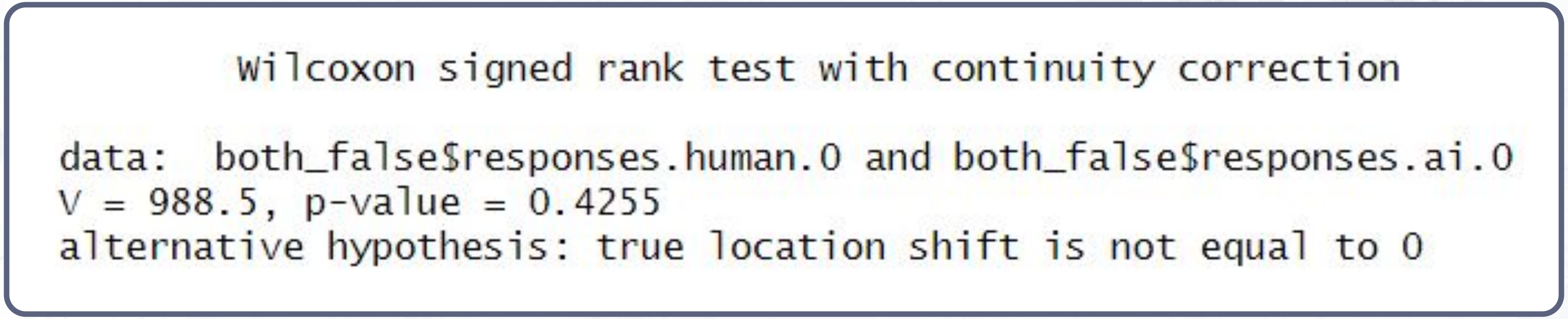
Stage 2: Pairwise Comparison



- **No rationales**:
  - Equal
- **Only one rationale**:
  - Skewed to rationale
- **Both rationales**:
  - Skewed to AI

Stage 2: Pairwise Analysis

Both False: accept the Null hypothesis = there is no difference between the human and AI (p-value > 0.05)



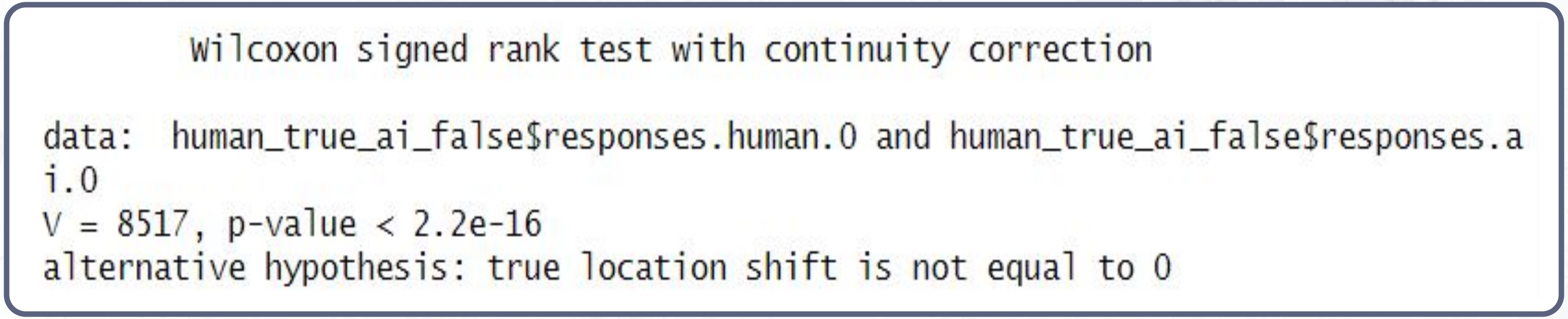
```
	Wilcoxon signed rank test with continuity correction

data:  both_false$responses.human.0 and both_false$responses.ai.0
V = 988.5, p-value = 0.4255
alternative hypothesis: true location shift is not equal to 0
```

Output test in R Studio

Stage 2: Pairwise Analysis

Human True, Ai False: reject the Null hypothesis = there is significant difference between the human and AI (p-value < 0.05)



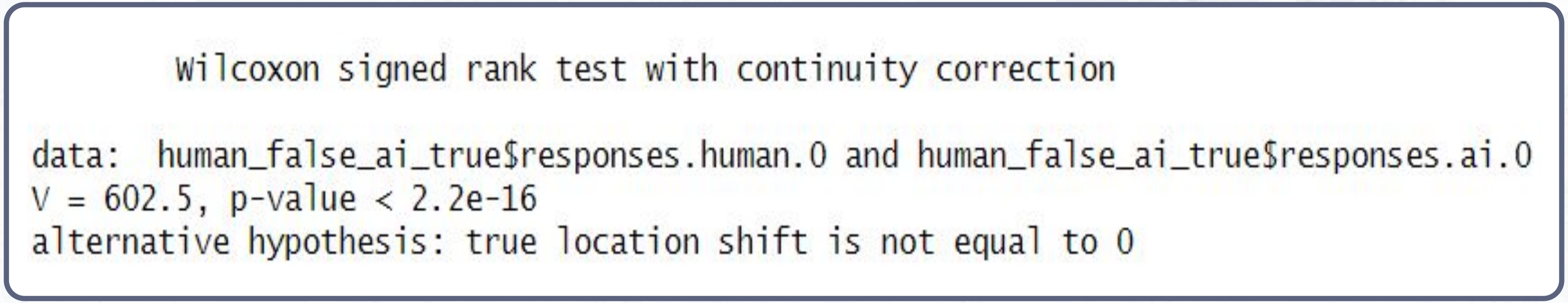
```
	Wilcoxon signed rank test with continuity correction

data:  human_true_ai_false$responses.human.0 and human_true_ai_false$responses.a
i.0
V = 8517, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Output test in R Studio

Stage 2: Pairwise Analysis

Human False, Ai True: reject the Null hypothesis = there is significant difference between the human and AI (p-value < 0.05)



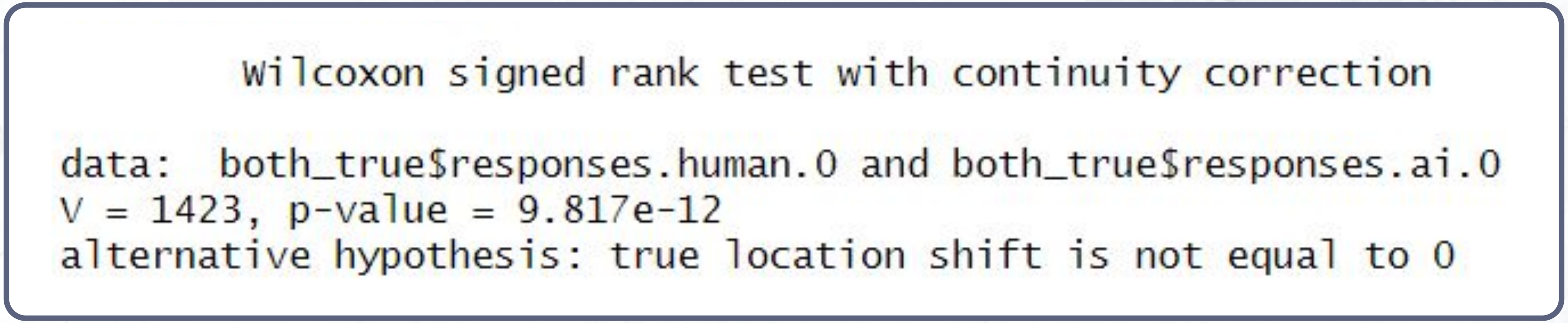
```
	Wilcoxon signed rank test with continuity correction

data:  human_false_ai_true$responses.human.0 and human_false_ai_true$responses.ai.0
V = 602.5, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Output test in R Studio

Stage 2: Pairwise Analysis

Both True: reject the Null hypothesis = there is significant difference between the human and AI (p-value < 0.05)

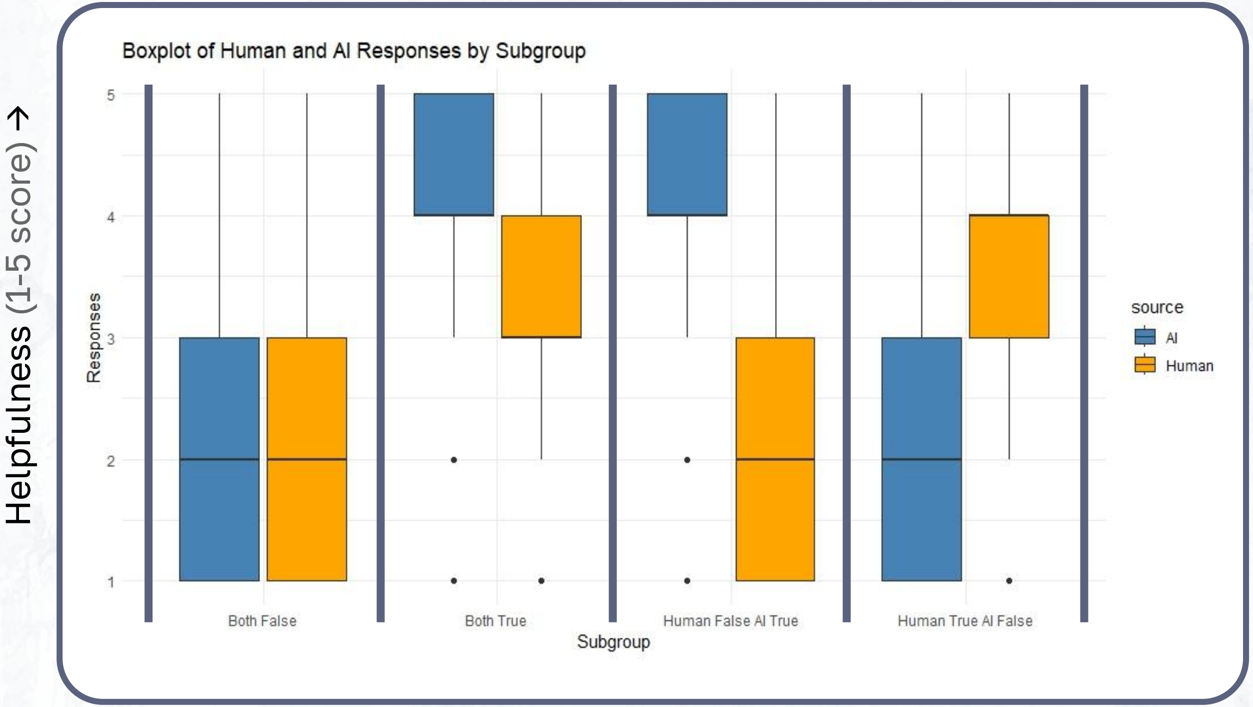


```
        Wilcoxon signed rank test with continuity correction

data:  both_true$responses.human.0 and both_true$responses.ai.0
V = 1423, p-value = 9.817e-12
alternative hypothesis: true location shift is not equal to 0
```

Output test in R Studio

Stage 2: Pairwise Analysis



Hypothesis 1: A rationale increases the trust towards its source *(Accepted)*

Hypothesis 2: If both human and AI provide **no** rationales, the human will be trusted more. *(Rejected)*

Hypothesis 3: If both human and AI provide rationales, the human will be trusted more. *(Rejected)*

Thanks!

Any questions?

GitHub: https://github.com/human-vs-ai

CREDITS: This presentation template was created by **Slidesgo** and includes icons by **Flaticon**, infographics & images by **Freepik** and content by **Eliana Delacour**

Please, keep this slide as attribution

Individual Contributions

(Bo) Toloka master

- ➢ Toloka task setup, pool setup, quality control

(Ilias) Writing + visualize

- ➢ Report writing
- ➢ Data visualization

(Maya) AI advice + visualize

- ➢ AI advice generation
- ➢ AI quality control
- ➢ Data visualization

(Philippe) Scripts and prompts

- ➢ Pre- and post-processing data
- ➢ Toloka input generation + task design
- ➢ AI advice generation
- ➢ Data visualization

(Sam) AI advice generation

- ➢ AI advice generation
- ➢ AI quality control