Human vs. AI: Exploring Trust in Predictions with Rationales

PHILIPPE DE BEKKER, BO VAN DEN BERG, MAYA ELASMAR, SAM HESLENFELD, and ILIAS PAPADIMITRIOU
Delft University of Technology, The Netherlands

Abstract. Artificial Intelligence (AI) has been evolving rapidly over the last few years, and it is being applied to a wide range of applications and domains. Trust plays a critical role in decision-making processes. This paper explores and compares the extent to which people trust humans versus AI when considering predictions with rationales. This study adopts a two-stage approach. In the first stage, crowd workers and AI are tasked with predicting the price of a house. In the second stage, crowd workers are asked to indicate their amount of trust in the price advice provided by humans and AI in addition to a final verdict, i.e. human, AI or no preference. A diverse sample of individuals from various demographic backgrounds and professional fields is recruited to ensure a representative dataset. Ultimately, the results of the last stage indicate no preference without rationales, preference for the rationale if only one is provided and most interestingly, AI when both provide rationales. A hypothesis (requiring future work) is the possibility of AI being trusted more due to having more digestible and convincing rationales compared to the human rationales, even though the accuracy was actually lower.

CCS Concepts: • **Human-centered computing**; • **Computing methodologies** → *Artificial Intelligence*;

Additional Key Words and Phrases: house price prediction, rationales, crowd computing, artificial intelligence, image context, trust

ACM Reference Format:
Philippe de Bekker, Bo van den Berg, Maya Elasmar, Sam Heslenfeld, and Ilias Papadimitriou. 2023. Human vs. AI: Exploring Trust in Predictions with Rationales. In *CS4145-Q4-2023: Crowd Computing, TU Delft*. 13 pages.

1 INTRODUCTION

A lot of significant advancements have occurred in the last years in the field of Artificial Intelligence (AI) — AI systems generate more accurate and coherent responses, making them viable sources of advice and information [1–4]. They have been widely adopted in various aspects of our lives and process vast amounts of data in very little time. On the other hand, humans can easily capture contextual information and are exceptionally wired in this regard [5–9].
Seeking advice is closely related to trust. Advice from people can be trusted for various reasons, such as the ability to empathize or to easily perceive human reasoning. On the other hand, AI can perform complex calculations, taking into account patterns and factors that humans cannot analyze, but are still widely considered as black boxes, which cannot provide accurate reasoning that justifies their prediction. Therefore, it is extremely valuable to compare the trust of people in AI when predictions include rationales.
The task for which trust will be evaluated is also crucial. House price prediction is a difficult and non-trivial question for both humans and AI. It depends on many factors, some of which are interdependent and non-linear. This task captures the human's ability to perceive situational context, as most people can differentiate between poor and wealthy homes and they can recognize luxurious interior design.

Manuscript submitted to ACM

This problem is addressed by combining crowd computing with AI. The central research question in this paper is formulated as follows:

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

Resulting from this research question, three hypotheses have been defined:

H 1. A rationale increases trust towards its source. Advice including a rationale will be preferred over one without it.
H 2. If both human and AI do **not** provide rationales, then the human will be trusted more.
H 3. If both human and AI do provide rationales, then the human will be trusted more.

2 RELATED WORK

Various research already exists in the domain of exploring trust in AI technology. These works provide some valuable insights and inspiration for this research. The following categories of related work are identified:

2.1 Trust in Artificial Intelligence

Trust in artificial intelligence is a complex issue that depends on many factors, such as user education, past experiences, user biases, and perception towards AI [10]. Previous research has shown that trust in artificial intelligence does not only depend on the reliability of the results it provides, but also tangibility and transparency [11]. In particular, people resist the black-box nature of AI [12–14]. Hence, AI is a highly intriguing topic that requires exploration across various genres of tasks. It is important to investigate the generalizability of each task, i.e. whether or not the task is suitable to extract information from to assess the trust in AI advice. Characteristics of the decision task, such as task risk, required expertise, and decision objectivity, can determine the task's generalizability [15]. Some sources report that humans tend to utilize the advice from AI more compared to advice from peers, though other findings claim a contrasting view - it boils down to the stakes involved in the operation. House price prediction is a task that belongs to the sector of business and finance, and therefore is considered a high-stakes task. In this task, there exists a ground truth and therefore quantitative measures of performance are suitable, which enables the objective evaluation of the human and AI predictions. Since house price prediction is considered an objective prediction task of high stakes, it makes for a highly generalizable task. Subsequently, the results of comparing human and AI predictions with rationales can provide a valuable contribution to understanding trust in this domain.

2.2 Providing Rationales

The rationale was first introduced as a way to increase the performance of machine learning for text classification [16]. It has been utilized in various annotation tasks, such as information retrieval relevance, sentiment analysis, or text classification. The rationale has proven to encourage more thoughtful decision-making and enable crowd verification [17]. It has also been proved to discourage cheating and provide additional domain value [18]. In this research, the additional domain value is the information that mentions which criteria were important for the crowd worker in order to estimate the price of a property and subsequently serves for great context and verification of the decision.

2.3 Framework of Similar Experiments

Several prior studies have explored the trustworthiness of advice from both human and AI sources [19–21]. A two-stage process is notably the most prevalent framework to compare trust. The first stage includes predictions made by both human and AI, and the second stage involves a verdict given by people to determine which advice is more trusted.

Unlike other studies, this study also explores experimenting with randomly in- and excluding rationales for the human and AI in order to gather more statistical evidence and evaluate whether any prior belief without rationales is skewing the distribution of verdicts with all rationales included.

3 SYSTEM DESIGN

One of the main objectives of our task was to ensure an appropriate comparison between a human and an AI. For example, the same input needs to be processed by both of them. To summarize all these requirements, a list of functional and non-functional requirements is provided below. Recall that the system is a twofold of prediction and comparison as explored in Section 2.3 and elaborated on more in-depth when discussing the experimental setup in Section 4. Ultimately, these requirements form a blueprint for the system that will be deployed during the experiment.

3.1 Functional Requirements

FR 1. The model representing the AI must be able to process images and answer questions about them.
FR 2. When providing crowd workers with input to form a prediction, additional context of the area from the property should be presented by providing a map with the surrounding area highlighted.
FR 3. When comparing human and AI predictions, the Graphical User Interface (GUI) must show the crowd workers the complete input first and then both human and AI outputs in randomized order to prevent any bias.
FR 4. Respondents should be able to provide feedback on the helpfulness of the human and AI predictions.
FR 5. The comparison task should allow respondents to pick a preference between a human and an AI prediction.

3.2 Non-Functional Requirements

NFR 1. During prediction, both human and AI need to receive the same attributes and require to answer in a similar format or select a fixed range of options.
NFR 2. The human and AI advice (prediction + rationale) should be checked for accuracy and reliability.
NFR 3. Each rationale should be free from errors and misleading information.
NFR 4. The crowdsourcing platform should allow worker selection based on selected qualities.
NFR 5. The crowd-working task must contain an attention check and require certain properties such as completion time and the number of characters in open responses to be above a certain limit.
NFR 6. The crowd workers must be paid a fair hour salary.
NFR 7. The crowd workers should remain anonymous to ensure their privacy and keep results unbiased.
NFR 8. The crowd workers must be able to understand English proficiently.
NFR 9. The GUI should be user-friendly, intuitive, and visually appealing.

4 EXPERIMENTAL SETUP AND EVALUATION

The experiment was deployed using a two-stage approach in order to explore trust in predictions with rationales. Both stages were performed by crowd workers using Toloka surveys [22]. In addition, the second stage required AI input which was prompted via MiniGPT-4 [23] (see Appendix B). This tool is free, however, using OpenAI's GPT-4 [1] for prompting might be more promising. The used dataset, obtained via Ahmed and Moustafa [24], contains 535 individual houses. Each house is described by four images (frontal view, kitchen, bathroom, bedroom), as depicted in Figure 1, and five textual attributes (number of bedrooms, number of bathrooms, surface area in square feet, zip code, price). Generating the AI advice with the free tool (Section 4.3) was time-consuming, so the choice was eventually made to

continue with only 21 houses. This reduction in house quantity was compensated by adding more overlap in both stages. Moreover, the prices for each task suite were calculated using the US minimum salary of $6 per hour (as advised by Toloka), resulting in a total experiment cost of $141.

The remaining part of this section discusses the data preparation (Section 4.1) needed to enable the first stage, divided by the human prediction (Section 4.2) and AI prediction (Section 4.3), and the second stage (Section 4.4). As quality control is universal throughout these sections, it is discussed separately in Section 4.5. An example of each normal and control task can be found in Appendix A.

Fig. 1. Illustrative example of images provided per house in the dataset [24]. Displays a frontal view, kitchen, bathroom and bedroom.

4.1 Data Preparation

To ensure the data used in this experiment was suitable for task deployment, analysis and comparison, the data underwent several modifications via the use of scripts written in Python.

As the dataset originates from 2016, the prices need to be aligned to the respective economic landscape of the respondent in order to properly evaluate the predictions against the ground truth and enable fair comparisons between human and AI predictions. Human knowledge translates to current time, however, the underlying AI training data from LAION is not up to date and has knowledge up until 2021 [25]. Utilising a web tool built around statistics from the U.S. Bureau of Labor Statistic allows for adjustment using inflation rates and a house price index [26]. The data indicates an adjustment factor for the human perspective of 1.30 from 2016 to 2023. For the AI perspective, factor 1.26 from 2016 to 2022 resulting in approximately 1.057 from 2022 to 2023.

In order to enhance the contextual understanding of the dataset for humans (especially crowd workers from all across the world), several measurements were taken. A Python zip-code package *uszipcode* based on US governmental data allowed for highlighting the respective areas on an interactive map provided in the Toloka survey (see Figure 6), facilitating easy exploration and visualization of the geographical distribution of the houses [27–29]. For example, users can check whether the house is located in an expensive part of the city or in a distant neighbourhood with no supermarkets nearby. In addition, some training data is also provided before doing any predictions to form a better mental model of the data, which also serves as quality control as mentioned in Section 4.5.

Furthermore, the surface area is only provided in square feet. Catering to diverse user preferences and understanding, the surface area attribute is also converted from square feet to square meters - a widely recognized format - in order to show both.

Lastly, as covered in the upcoming sections, the data needed to be translated into Toloka survey tasks built on JSON data, input prompts for MiniGPT-4 and results needed to be post-processed.

4.2 Stage 1a: House Price Estimation by Humans

In the first stage, crowd workers were asked to provide an estimation of the price of the house based on pictures of the frontal view, the bathroom, the bedroom, and the kitchen. They were also provided with numerical information about the number of bedrooms and bathrooms, and the location of the zip code in a map. From this map, someone can deduct information about whether the house is in a rural or urban area and if any significant landmarks are nearby. The estimations of the price are presented as predefined ranges, such as $200K - $300K. The crowd workers are also asked to provide a short rationale of 2-3 sentences, in which they justify the reason for which they provide this price range. Each stage 1 task had an overlap of 3 reviewers. Before the task begins, two examples are given where the price of the apartment is shown, in order to familiarize the crowd worker with both the user interface and some example prices.

4.3 Stage 1b: House Price Estimation by Artificial Intelligence

In order to calculate the AI estimations, we used the MiniGPT-4 API [23]. MiniGPT-4 is an AI tool for image processing, which can then be used to then extract information from it. After processing the images of a house, this API was used for two purposes, the first being the prediction of the house price. This prediction was asked to be made, based on the images and properties of the house, and in the same price ranges that we included in the first stage of the experiment. In addition, the API was used to generate a rationale for the predicted price range. This rationale should make the AI reasoning more clear and more understandable, as this reasoning is often considered to be a black box. The exact prompt that has been used to collect the MiniGPT-4 results, can be found in Appendix B. The generated house price estimates already account for inflation. The tool also considers prices of similar houses in the zip code.

4.4 Stage 2: Advice Evaluation

For the second stage of the experiment, crowd workers were asked to compare the estimations provided by humans and AI. They were provided with the information that is provided in the first stage, as well as the human estimation and rationale and the AI estimation. In order to measure the rationale's impact, each individual explanation randomly included or excluded the rationale. Each stage 2 task had an overlap of 7 reviewers.

4.5 Quality Control

This subsection outlines the quality control measures taken during the experiment. These resulted in approval rates of 81% for stage 1 and 48% for stage 2. The relatively lower acceptance rate for stage 2 might indicate that multiple-choice forms are more hastily completed and that workers are more focused on images than text.

QC 1. Each task suite contains one control task, where the desired answer is clearly provided. This question functions as an attention check.
QC 2. The first-stage rationales have an enforced minimum length of 50 characters to avoid very small generic answers.
QC 3. Submissions were manually reviewed and were rejected if the submission time was unreasonably low, the control task was answered incorrectly, or if the rationales were insufficiently poor or not unique.
QC 4. Users were automatically banned after multiple fast responses, rejections, or failed control tasks.
QC 5. Only the top 20% of users quality-wise were allowed to participate in the tasks.

5 RESULTS AND DISCUSSION

In this section, the findings of our experiments and analysis are presented. The results of both stages will be visualized and analyzed. In the first stage section, the accuracy of the human and AI predictions is compared and the dataset is divided in expensive and more affordable properties, in order to provide an intuition about what properties tend to be underestimated or overestimated by humans or AI. In the second stage section, statistical tests take place, aiming to investigate the hypotheses that were earlier introduced, i.e. whether a rationale increases trust towards its source, and whether the human will be trusted more, in case both humans and AI do or do not provide rationales.

5.1 First Stage Results

In the first stage of the experiment, the crowd workers achieved higher accuracy in their predictions than AI. As we observe in Figure 2, both of them provide relatively accurate predictions. The Mean Absolute Percentage Error (MAPE) was calculated by comparing the prices to the mean of each range of predictions and was found to be 24.2% for the crowd workers and 36.3% for the AI. Particularly, as depicted in Figure 3a, the prices of the houses that cost over 700 thousand dollars got significantly underestimated by the AI and slightly underestimated by humans.

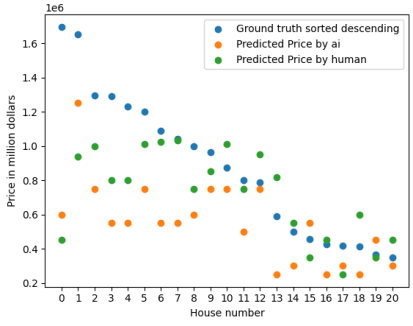
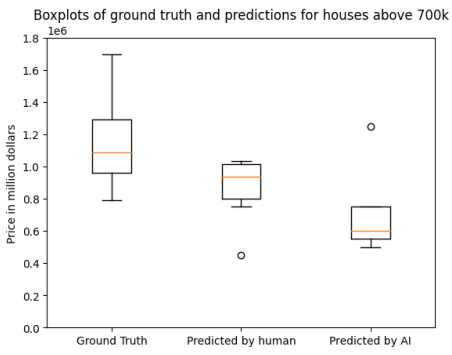
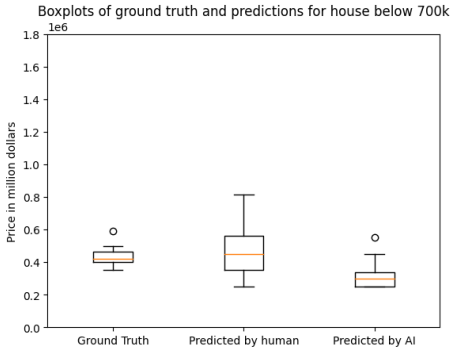


Fig. 2. Stage 1 results of ground truth price sorted descending compared to the predicted prices.



(a) Predictions compared to ground truth in houses over $700k.

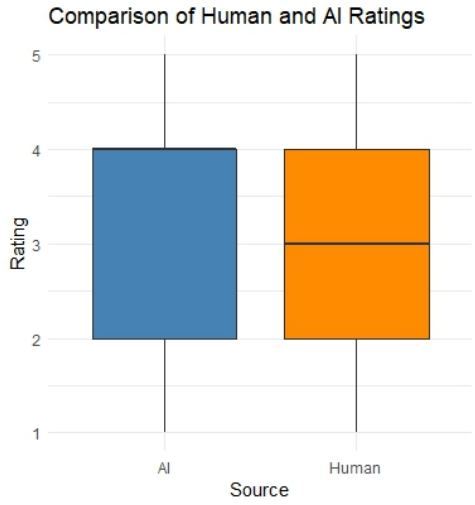


(b) Predictions compared to ground truth in houses under $700k.

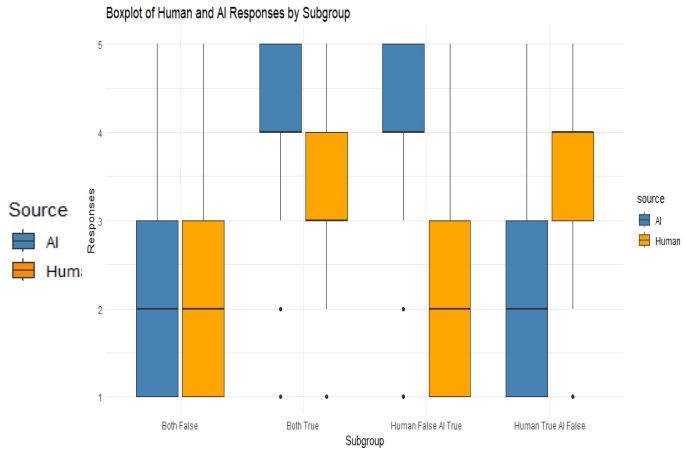
Fig. 3. Predictions compared to ground truth boxplots.

5.2 Second Stage Results

For the second stage, the crowd workers were asked to rate how helpful the human and AI were on a scale from 1 to 5. In R studio, a t-test was performed on these scales to compare humans with AI. The t-test assumes the data has a Gaussian distribution [30]. The t-test tests the null hypothesis: there is no difference between the mean of the human scales and AI scales. This hypothesis is rejected because the p-value < 0.05, which means there is a significant difference between Human and AI scales. The mean difference is 0.3819672 in favour of the AI. Figure 4a shows that AI ratings have indeed a higher mean.



(a) Stage 2: Comparison between the overall verdict distribution between human and AI ratings.



(b) Stage 2: The boxplots of the four subgroups. True and False on the y-axis indicate the presence of a rationale. Each time, AI is mentioned first, followed by human.

Fig. 4. The boxplots of the ratings from various perspectives. Ratings are from 1 to 5, indicating the helpfulness of the provided advice.

Furthermore, a Wilcoxon test was performed on the ratings of humans and AI. This test does not make any assumption about the data distribution. This test also rejected the null hypothesis and showed that there is a significant difference between Human and AI ratings with a p-value < 0.05. Additionally, the Wilcoxon test was performed on each task of the second stage. Results of these tests are depicted in Table 1.

Table 1. Wilcoxon signed rank test with continuity correction

Human Rationale	AI Rationale	P-Value
✓	✓	9.817e-12
✓	×	2.2e-16
×	✓	2.2e-16
×	×	0.4255

The Wilcoxon test, when providing no rationales for both humans and AI, shows a p-value > 0.05 (no significant difference). For all other possibilities, i.e. only one rationale for both or both rationales, the test shows a p-value < 0.05,

which indicates a significant difference between the two rating groups. The Wilcoxon test does not show which source (Human or AI) has a higher mean for these subgroups. However, this can be read from the boxplots in Figure 4b. The figure shows that when the human rationales are provided, the mean ratings for humans are higher, with a mean difference of 2. The same holds for AI, when its rationales are provided, it has a higher mean. The mean difference for that case is 3. When both rationales are provided, the mean rating for AI is higher by 2. Based on these results, our first hypothesis: a rationale increases trust towards its source. Advice including a rationale will be preferred over one without it, is accepted. The second hypothesis: if both humans and AI do **not** provide rationales, then the human advice will be trusted more, is rejected. The final hypothesis: if both humans and AI do provide rationales, then the human advice will be trusted more, is rejected. Finally, it is shown in Figure 5 that people generally preferred AI over humans when considering (house price) predictions with rationales.

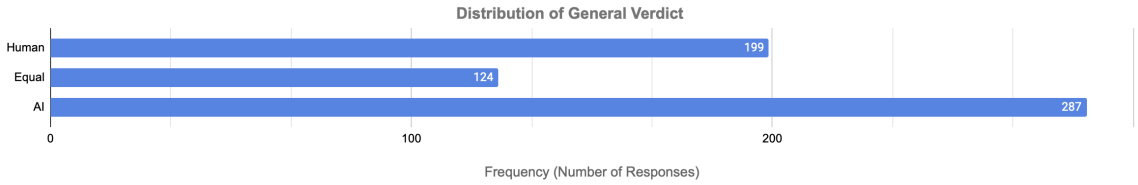


Fig. 5. Summarized preferences of people regarding (house price) predictions. Verdict can either be human, AI or equal.

6 CONCLUSION & FUTURE WORK

The aim of this study was to investigate whether people have more trust in humans or AI when considering (house price) predictions with rationales. After conducting a two-stage experiment, results indicate that people generally trust AI more than humans, even though this is not fully backed up by the accuracy compared to the ground truth. When crowd workers were not provided with any rationales, they rated human and AI advice evenly. When only one rationale was shown, that advice was rated substantially more useful. Essentially, this follows direct intuition. However, the mean difference was higher when only AI rationales were provided compared to only human rationales. This already shows that, even when only human rationales were provided, people thought AI was relatively promising. Eventually, when both rationales were provided, our study shows that people rate AI higher than human advice instead of equal as was the case with both not providing rationales. Summarizing all the groups, AI is cumulatively rated the highest and preferred over human advice.

Based on certain limitations, we deem the following as valuable for future work. The experiment could be extended by using residents nearby or from the region itself as crowd workers/respondents (as opposed to distant ones in the performed research). Enhanced contextual understanding of factors that influence the price of a property, such as the neighbourhood's reputation, could affect trust due to personal experience in this region.It would be interesting to make a controlled experiment that measures the effect of personal experience to the results. Another similar approach could be adding another stage where information is collected about critical indicators of the region, such as the reputation or the proximity to amenities (schools, hospitals, shopping centers), allowing for more precisely composed responses and thus results. Furthermore, other AI models should be compared as well for a broader perspective on the study. For example, using GPT-4 instead of MiniGPT-4 should increase the quality of the AI predictions and rationales. Lastly, it would be valuable to research if AI writing more digestible advice compared to humans influences the trust verdict positively, no matter the underlying accuracy.

REFERENCES

[1] OpenAI. Gpt-4 technical report, 2023.

[2] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[3] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.

[4] Namkee Oh, Gyu-Seong Choi, and Woo Yong Lee. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Annals of Surgical Treatment and Research*, 104(5):269, 2023. doi: 10.4174/astr.2023.104.5.269. URL https://doi.org/10.4174/astr.2023.104.5.269.

[5] JE Korteling, GC van de Boer-Visschedijk, RA Boswinkel, RC Boonekamp, Rubricering rapport Ongerubriceerd, Managementuittreksel Ongerubriceerd, Samenvatting Ongerubriceerd, Rapporttekst Ongerubriceerd, and Bijlagen Ongerubriceerd. Effecten van de inzet van non-human intelligent collaborators op opleiding and training [v1719]. *Report TNO 2018 R11654. Soesterberg: TNO defense safety and security*, 2018.

[6] J.E. (Hans) Korteling, Gillian van de Boer-Visschedijk, Romy Blankendaal, Rudy Boonekamp, and A. Eikelboom. Human- versus artificial intelligence. *Frontiers in Artificial Intelligence*, 4:622364, 03 2021. doi: 10.3389/frai.2021.622364.

[7] Ben Shneiderman. Design lessons from ai's two grand goals: human emulation and useful applications. *IEEE Transactions on Technology and Society*, 1(2):73–82, 2020.

[8] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6): 495–504, 2020.

[9] Yanyan Dong, Jie Hou, Ning Zhang, and Maocong Zhang. Research on how human intelligence, consciousness, and cognitive computing affect the development of artificial intelligence. *Complexity*, 2020:1–10, October 2020. doi: 10.1155/2020/1680845. URL https://doi.org/10.1155/2020/1680845.

[10] Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. Artificial intelligence and human trust in healthcare: Focus on clinicians. *J Med Internet Res*, 22(6):e15154, Jun 2020. ISSN 1438-8871. doi: 10.2196/15154. URL https://doi.org/10.2196/15154.

[11] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14 (2):627–660, 2020.

[12] Stanford Law School. Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know, 4 2019. URL https://law.stanford.edu/publications/artificial-intelligence-in-the-health-care-space-how-we-can-trust-what-we-cannot-know/.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

[14] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. doi: 10.1016/j.artint.2018.07.007. URL https://doi.org/10.1016/j.artint.2018.07.007.

[15] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: A survey of empirical studies, 2021.

[16] Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1033.

[17] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why is that relevant? collecting annotator rationales for relevance judgments. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 4(1):139–148, Sep. 2016. doi: 10.1609/hcomp.v4i1.13287. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13287.

[18] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69:143–189, 2020.

[19] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on ai systems, 01 2023.

[20] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions, 2021. URL https://arxiv.org/abs/2107.07015.

[21] Xinru Wang and Ming Yin. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Trans. Interact. Intell. Syst.*, 12 (4), nov 2022. ISSN 2160-6455. doi: 10.1145/3519266. URL https://doi.org/10.1145/3519266.

[22] Toloka AI. Powering ai with human insight - toloka ai, 2023. URL https://toloka.ai/.

[23] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. URL https://arxiv.org/abs/2304.10592.

[24] Eman Ahmed and Mohamed Moustafa. House price estimation from visual and textual features. *CoRR*, abs/1609.08399:1–7, 2016. URL http://arxiv.org/abs/1609.08399.

[25] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. URL https://arxiv.org/abs/2111.02114.

[26] Official Data Foundation / Alioth LLC. Housing price inflation calculator. https://www.in2013dollars.com/Housing/price-inflation/2016-to-2023?amount=1, 2023. [Accessed 7-Jun-2023].

[27] S Hu. Documentation uszipcode, 2022. URL https://uszipcode.readthedocs.io/.
[28] U.S. Census Bureau. Data python package uszipcode. URL https://data.census.gov/cedsci/table?q=94103.
[29] Toloka. Toloka template-builder map documentation. URL https://toloka.ai/docs/template-builder/reference/field.map/.
[30] Muhammad Usman. Power efficiency of sign test and wilcoxon signed rank test relative to t-test. 2015. Retrieved from https://core.ac.uk/download/pdf/234680288.pdf.

A TASK EXAMPLES

For stage 1, an example standard task looks like Figure 6.

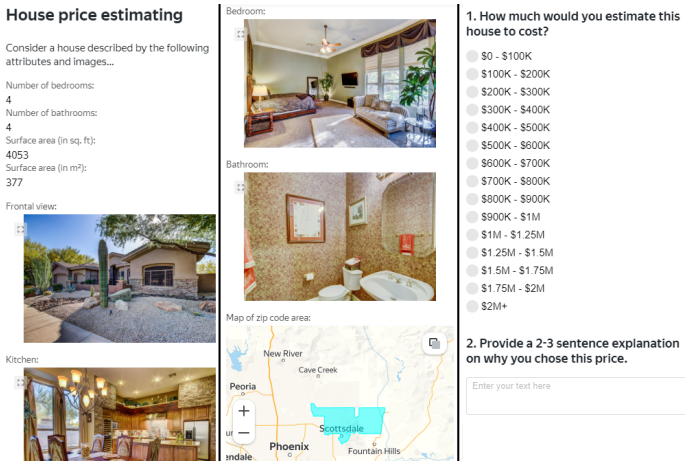


Fig. 6. Example of a normal task in stage 1.

For stage 1, an example control task looks like Figure 7. As validation of the control task, we check if the worker entered the price range $800K - $900K, as shown in the picture. The rationale provided in the control task is not validated due to its complexity.

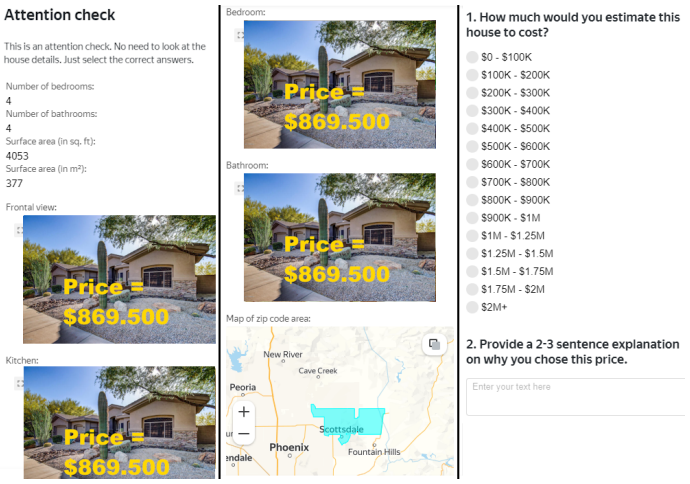


Fig. 7. Example of a control task in stage 1. The clear answer here is the range $800K - $900K. The rationale is not checked.

For stage 2, an example standard task looks like Figure 8, and an example control task looks like Figure 9. As validation of the control task, we check if the worker sequentially selected the answers '3', '1', and 'no preference', as indicated in the questions and advice.

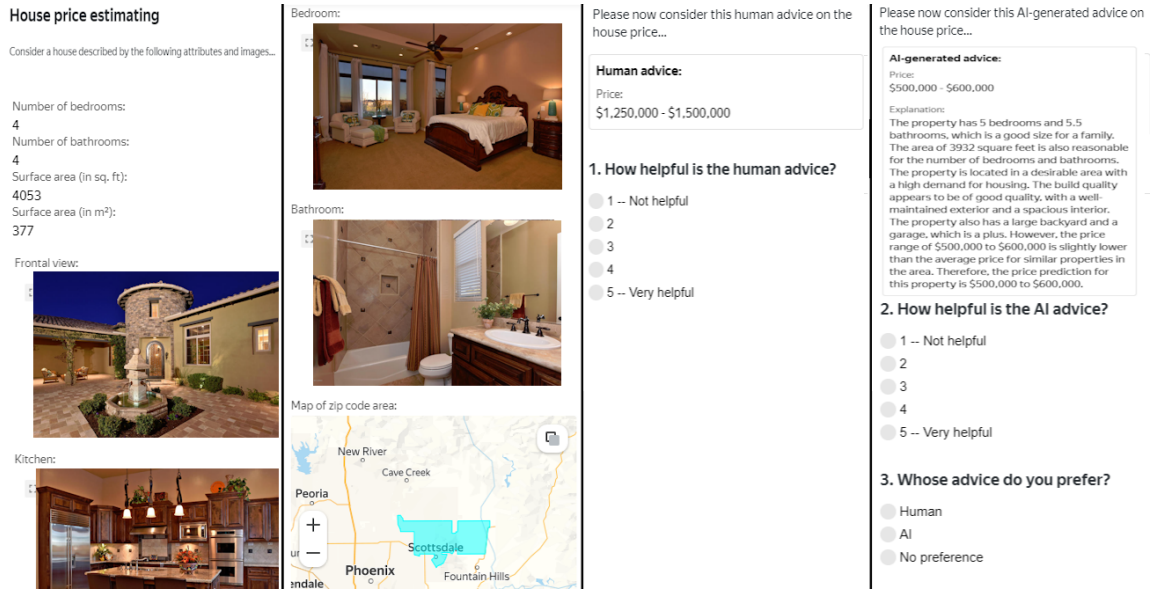


Fig. 8. Example of a normal task in stage 2.

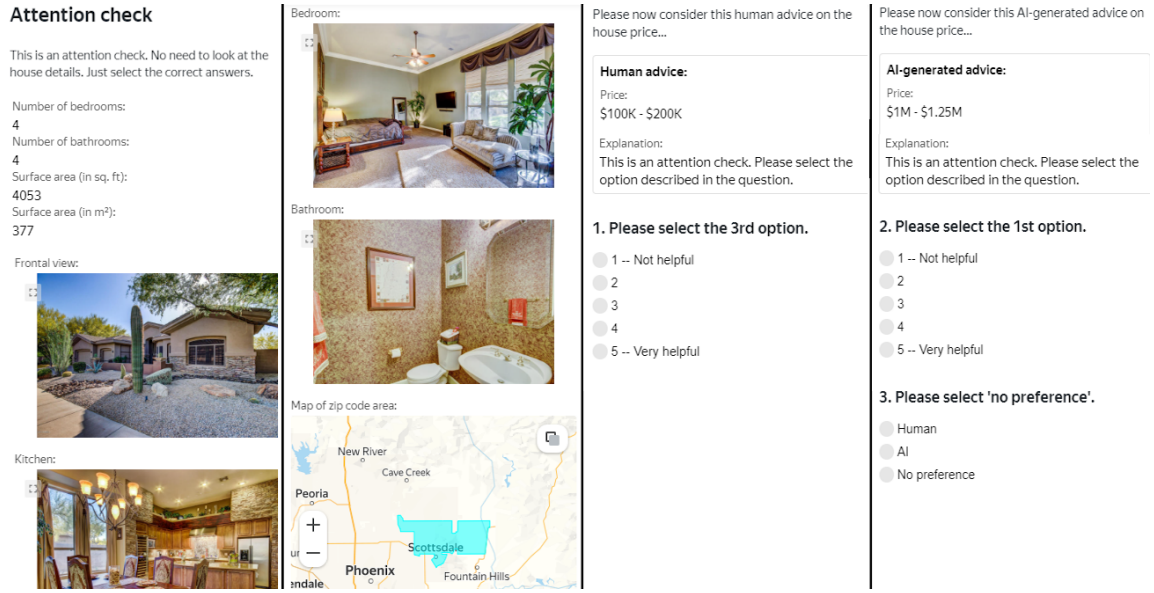


Fig. 9. Example of a control task in stage 2. The required answers are '3', '1', and 'no preference', as indicated in the questions.

B MINIGPT-4 PROMPTS

For the prediction of house prices and the generation of its corresponding rationale, the following prompt has been used (on the respective images of a certain house which were put in a 2 x 2 grid consisting of the frontal view, kitchen, bedroom and bathroom image), in which the data from the dataset was entered where necessary:

```
Please give a price prediction for the property in the provided image and textual attributes below.
Take into account the number of bedrooms, bathrooms, and areas are the most important features to
focus on.

The property has the following attributes:
- Number of Bedrooms: [number of bedrooms]
- Number of bathrooms: [number of bathrooms]
- Area: [area in sqft]
- Zipcode: [zipcode]

Furthermore, provide your rationale for the answer. Please refer in your rationale to the image,
e.g. by detecting build quality, garden, garage (if present), etc., and the importance of the
attributes.

Format your answer in the form of a JSON:
{
    “zipcode”: number,
    “price_prediction_lower_bound”: number,
    “price_prediction_upper_bound”: number,
    “rationale”: string
}

Note that price_prediction_upper_bound - price_prediction_lower_bound < 100000.
```

Received 23 June 2023