A Appendix

Here we provide some additional information on our study design and dataset (Appendix B), when under/overconfident advice can help (Appendix C), our human behavior model (Appendix D), crowd-sourced data collection (Appendix E), the survey shown to empirical study participants (Appendix F), and additional results using step functions for the modified advice (Appendix G).

B Study Design and Data

B.1 Study design

In Figure 6, we visualize the study design described in Section 2. A human first completes a task on their own (Response 1). They then are given advice from an AI algorithm and are allowed to change their answer (Response 2).

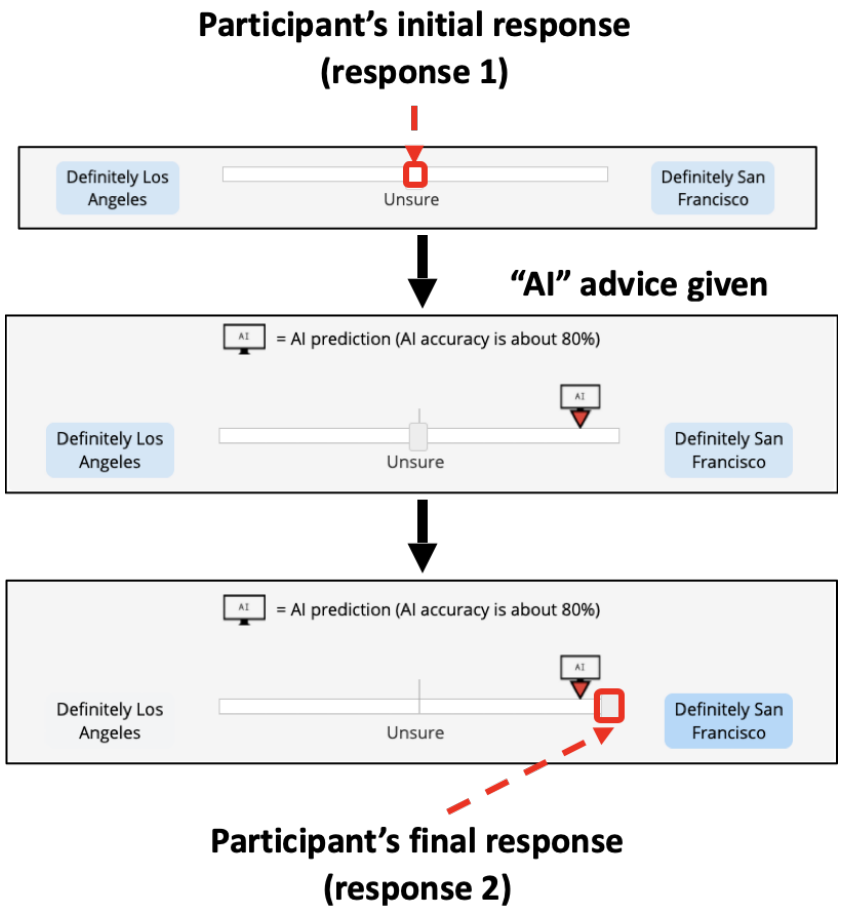
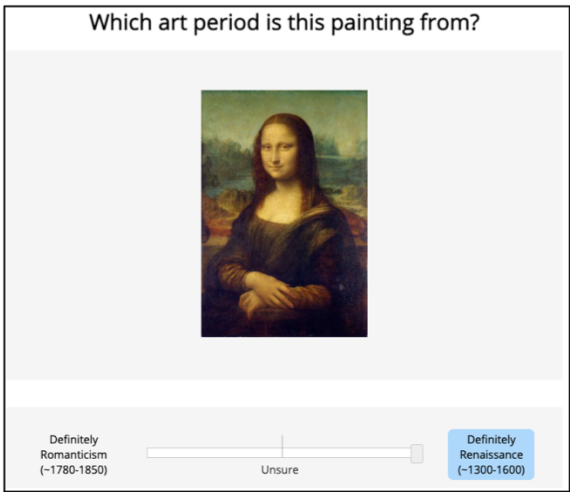


Figure 6: Visualization of our experimental setup.

B.2 Tasks

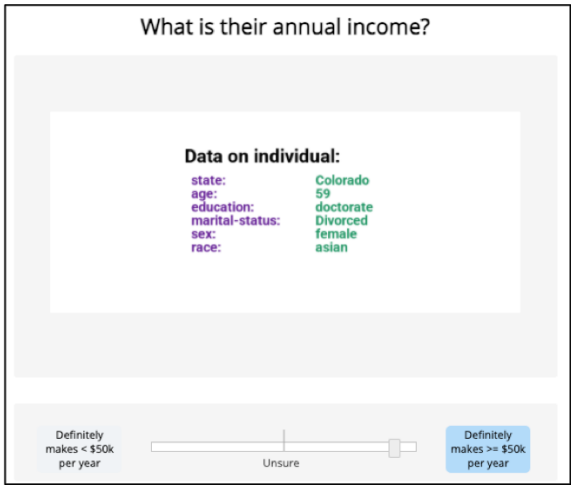
In Figure 7, we show example questions from each of the four tasks described in Section 2. For each task, questions were chosen to have a wide range of difficulties – in Figure 8, we show the probability density function of the advice across all four tasks. The mean advice across questions is indicated by the dotted vertical lines – it is similar across all four tasks.



(a) Art dataset

(b) Cities dataset

(c) Sarcasm dataset



(d) Census dataset

Figure 7: Example tasks for each of the four datasets we use.

B.3 Unmodified AI advice

In Figure 8, we plot the probability density function of the advice for each of the four tasks. Positive and negative values indicate correct and incorrect advice respectively. The magnitude of the advice indicates confidence. Notice that (1) the advice is skewed towards the right as the advice is roughly 80% accurate, (2) the average advice (indicated by the vertical dotted lines) is similar across all four tasks, and (3) the quality of advice given is fairly diverse, with each task having questions with incorrect and low-confidence advice.

In Figure 9, we visualize the calibration of advice confidence. As was mentioned in Section 2, the expected calibration error (ECE) is calculated to be 0.074 indicating the advice is well calibrated.

B.4 Summary of baseline performances

In Table 4, we summarize the baseline human and advice accuracies (i.e., before any interaction between the human and AI has occurred).

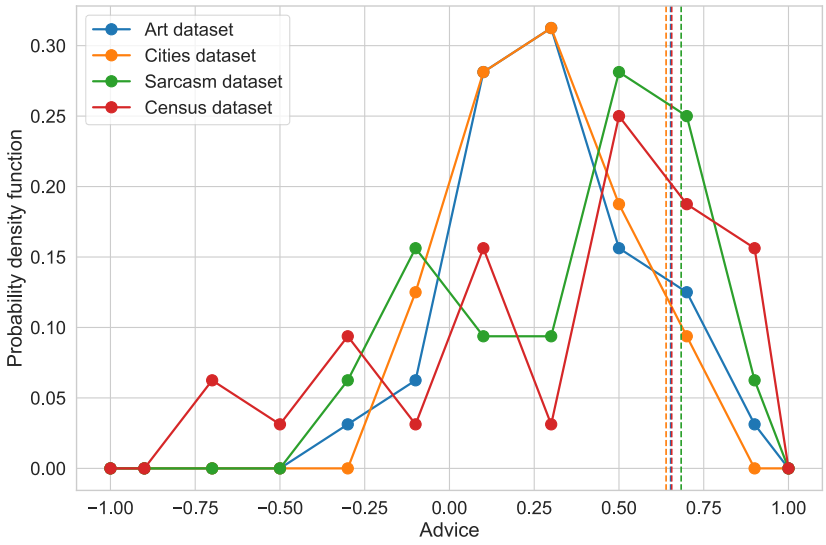


Figure 8: Distribution of advice confidence for all four tasks. Negative values indicate incorrect advice and positive values indicate correct advice. Dotted lines indicate mean advice across questions.

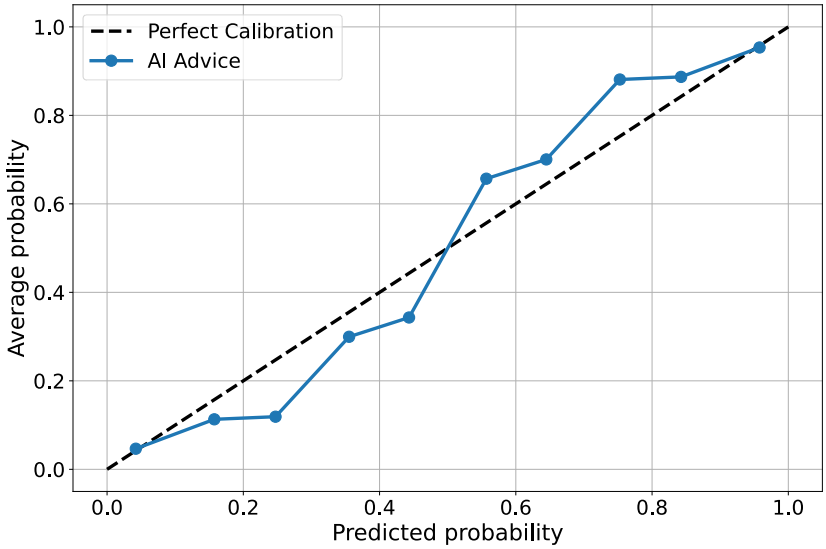


Figure 9: Calibration plot of the AI advice, aggregated across all four tasks. We note that the advice is roughly calibrated with an expected calibration error of 0.074.

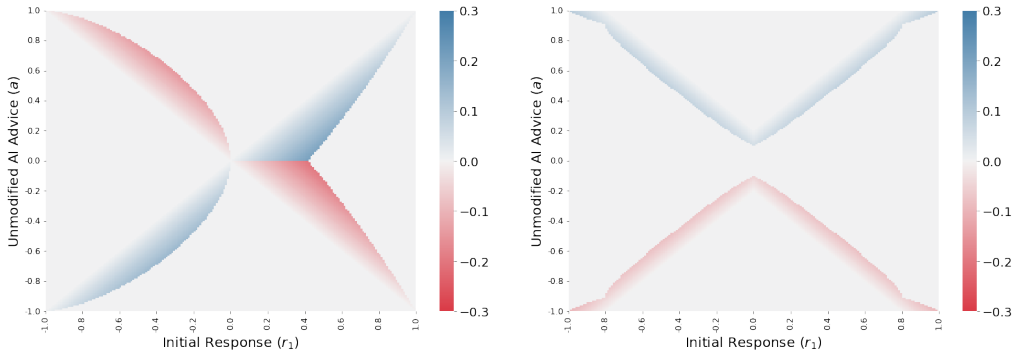
C Under- and Overconfident Advice

Here we provide additional examples for when under/overconfident advice is beneficial. We make the same assumptions from Section 4.2; they are repeated below:

1. Our goal is to minimize the logistic loss of the human response after receiving AI advice (we consider only binary classification problems here).
2. Humans either use their initial response or follow the given advice exactly. This is a constrained version of the activation-integration model.
3. We have an oracle for human behavior and know exactly the calibration of the human and the AI advice.

Table 4: Overview of baseline human and AI accuracies (before human-AI interaction occurs).

Task	Human Accuracy	AI Accuracy
Art	63.7%	90.1%
Cities	70.7%	87.5%
Sarcasm	72.7%	78.1%
Census	70.1%	78.1%



(a) Miscalibrated human (b) Biased human

Figure 10: Optimal human-calibrated AI output for two settings. Heatmap shows delta change in advice compared to calibrated advice: red and blue values indicate where the AI should be under- and overconfident respectively in order to achieve the best human-AI system output.

C.1 Miscalibrated human

Consider the case where $p_{r_1} = r_1^2$ – that is, the human is miscalibrated (in particular, underconfident for Label 0 but overconfident for Label 1 predictions). We show in Figure 10a the optimal modification to a calibrated AI advice to correct for the human's miscalibration. This modified advice is optimal in the sense that it minimizes the logistic loss of the human's final response given (potentially modified) advice, assuming the aforementioned model for human behavior.

The left region indicates the advice should as underconfident to compensate for the human's underconfidence in this region. This results in a decrease in confidence when the advice is predicting Label 1 and an increase in confidence when the advice is predicting Label 0. Recall that in the assumed setting, the human selects to follow the advice exactly whenever it is perceived to be more confident than the human is in their initial response. This results in symmetric behavior across the label boundary.

Similarly, the right region indicates the advice should be overconfident to compensate for the human's overconfidence in this region.

C.2 Biased human

Consider the case where the human is biased when valuing their response compared to the advice. In particular, $p_{\text{activation}} = \mathbb{1}[\bar{r_1} + \epsilon < \bar{a}]$ (here $\bar{x} = \max\{x, 1-x\}$), with $\epsilon = 0.1$. Here, the human is overvaluing their own response – they only use the advice when the advice confidence exceeds their own confidence by the margin, ϵ.

In Figure 10b, we show how to modify the advice to compensate for this bias. As the human is undervaluing the AI advice, the modified advice only increases the confidence of the AI in certain conditions (this corresponds to a positive delta when the predicted labels is "1" and a negative delta when the predicted label is "0").

Note there is a gap between $0.45 - 0.55$ for the unmodified advice where no modification is recommended. This gap results from being unable to decrease the logistic loss when the advice is

sufficiently unsure. For the advice to have an impact, it must be presented as more confident than the human by at least ϵ; however, for regions where the advice truly has low confidence, presenting the advice as overconfident results in higher logistic loss.

C.3 More complex examples

Now consider taking together the two modifications from the above settings – $p_{r_1} = r_1^2$ and $p_{\text{activation}} = \mathbb{1}[\bar{r_1} + 0.1 < \bar{a}]$ for $\bar{x} = \max\{x, 1-x\}$. This gives us the scheme for modified advice shown in Figure 3b.

C.4 Uncalibrated AI Assumption

In our analysis, we assumed the AI was calibrated. This is necessary – if the advice is not calibrated, we cannot optimize the advice as we do not know the value of p_a.

However, if we are given uncalibrated advice but are able to calibrate that advice, it is possible to remove this assumption. In particular, the optimal modified advice will be a composition of a calibration function and the modification based on that calibrated advice.

C.5 Optimization procedure

We optimize each advice-initial response pair separately. For each advice-initial response pair, we compute the optimal modified advice using a line search method with a step-size of 5×10^{-3}.

In general, the candidate for the optimal value of modified advice is the advice with the smallest change in advice confidence that results in the human being activated. If following this modified advice results in a higher logistic loss, then the optimal modified advice will be to make no modification.

D Human Behavior Model

D.1 Features for the human behavior model

Here we detail the features used for our activation and integration models ($f_{\text{activation}}$ and $f_{\text{integration}}$). Features and a description are given in Table 5.

Table 5: Description of input features for $f_{\text{activation}}$ and $f_{\text{integration}}$ models.

Feature #	Value Range	Description
1	$[0, 1]$	Response 1 Confidence
2	$[0, 1]$	Advice Confidence
3	$\{-1, +1\}$	$+1$ if Advice and Response 1 agree on label, otherwise -1
4	$[-1, 1]$	Feature 1 $\cdot$ Feature 3
5	$[-1, 1]$	Feature 2 $\cdot$ Feature 3
6	$[-1, 1]$	Survey question measuring human's perception of AI performance on the given task
7	$\mathbb{R}_{\geq 18}$	Age
8	$\{0, 1\}$	Sex
9	$\{0, 1\}$	Does the person have prior experience with computer programming?
10	$[1, 10]$	Self-reported socioeconomic status
11	$[0, 1]$	Survey question assessing perceived presence of AI in person's life
12	$[1, 8]$	Self-reported education level

D.2 Human behavior model training procedure

The activation and integration models are three-layer fully-connected neural networks with ReLU activations. The layer sizes are (1) 12×24, (2) 24×12, (3) 12×1. We trained our models using

the empirical data described in Section 2. We optimized our models using the Adam optimizer with learning rate 1×10^{-3} and no weight decay. We used early stopping to avoid overfitting.

E More Details on Crowd-Sourced Data Collection

Here we include additional details on the crowd-sourced data collection procedure we run.

E.1 Demographic information collection

Our experiments were run on Prolific [23]. The demographic information used by our activation-integration model was provided by Prolific. Education level is defined on a scale from 1 to 8, and should be interpreted as:

1 – Don't know / not applicable

2 – No formal qualifications

3 – Secondary education (e.g. GED/GCSE)

4 – High school diploma/A-levels

5 – Technical/community college

6 – Undergraduate degree (BA/BSc/other)

7 – Graduate degree (MA/MSc/MPhil/other)

8 – Doctorate degree (PhD/other)

Socioeconomic status is defined on a scale from 1 to 10 and was assessed by asking participants to answer the question shown in Figure 11.

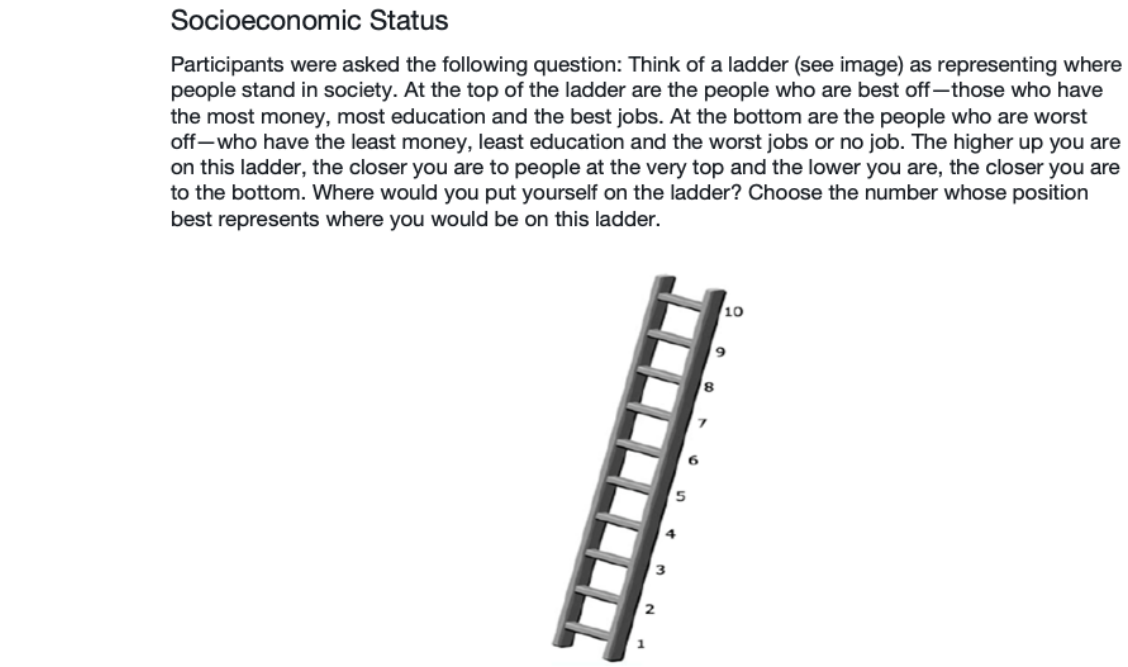


Figure 11: Socioeconomic status question.

E.2 Survey questions

Participants were asked two survey questions. We used their responses in our activation and integration models. The questions were:

1. Do you think the AI or the average person (without help) can do better on this task?
2. How often do you use AI systems to aid you in your everyday life and/or at work?

Participants responded to each question on a slider scale. The first question was intended to measure participant's perceived confidence in the AI prior to doing any tasks. The second question was intended to measure the participant's familiarity and comfort with AI.

E.3 Participant compensation

We compensated participants at $\geq \$10.00$ per hour (depending on bonus pay) as per the recommended rates by Prolific. We informed participants of the possibility for bonus pay. Bonus was calculated as:

$$\text{bonus} = \begin{cases} 0 & \text{if } S < 0.3 \\ S * 0.3 & \text{otherwise} \end{cases}.$$

Here, S is the average performance, computed as (the average of)

$$sign(\text{correct response}) \cdot \text{response}_1,$$

where $-1 \leq \text{response}_1 \leq 1$.

F Participant Instructions

Our study was designed using standard methods [28]. We implemented our study as a simple web app using jsPsych [4] with a Python-based web server. This was necessary to (1) customize our survey design and (2) deploy our survey to crowd-sourced survey participants.

Here we describe the survey shown to participants. In the "example-survey-slides" folder of the supplementary material, we include a set of PDF documents showing screenshots of our survey for the Art dataset for a participant. Each "page" in the directory corresponds to a separate web page. Filenames are numbered in the order a participant encounters them. Clicking "continue" / "submit" brings the participant to the next web page.

A brief description of the survey. Any content referring to the art data was substituted with appropriate language for the given task.

1: Participant enters a unique ID assigned to them through Prolific [23], the crowdsource platform we use to recruit participants.

2: Instructions specific to the task and advice source participants will receive. Note that this page is seen by participants as a single, continuous web page.

3: Additional information on the advice source.

4: Information on bonus payment.

5: Manipulation check. Participants who got the wrong answer were sent back to Page 2.

6: Pre-survey question to assess prior belief in AI advice

7: Example task: recording Response 1.

8: Example task: recording Response 2.

9: After completing all tasks, Participants are shown this screen.

10-14: Additional survey questions.

15: Check for errors in survey.

16: Debrief slide.

17: Bonus payment and survey submission slide.

G Additional Results

G.1 Other AI modification functions

We additionally ran experiments to compare other modification functions to the sigmoid-like $g_{\alpha,\beta}$ function we used to modify the AI in the main experiments. In particular, we consider two variants of a step function:

$$g_{step,\lambda}(x) = \begin{cases} \lambda & \text{if } x \geq 0 \\ -\lambda & \text{if } x < 0 \end{cases}$$

for some $\lambda \in (0, 1]$. For our experiments, we chose $\lambda \in \{0.50, 0.95\}$. We visualize these transform functions in Figure 12.

The idea of using a step function is to remove all information about advice confidence. We present the information in the same way as other experiments to remove possibility that the presentation affects human decision making. The two lambdas we selected allow us to compare showing whether showing some variation of confidence in the binary output of the AI affects human usage.

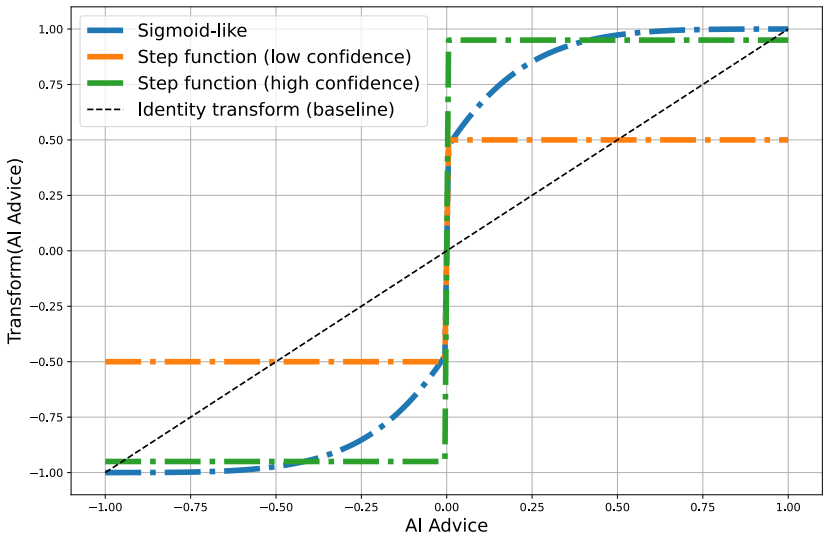


Figure 12: Visualization of the transform functions we evaluate in additional experiments.

We compare the performance of each of these functions by recruiting 100 adult US-based participants and randomly assigning each to a different modified advice. The experiment was conducted using the painting identification task. All other aspects of the recruitment process and study design are the same as used in the main study.

Results are summarized in Table 6. As there are variations between the people in each group (and their performance on the tasks before seeing advice), we use binned averages of the metrics. In particular, we place each human-AI interaction in a bin based on the initial response of the user, compute the metric value in each bin, and finally average across bins. We used 21 uniformly spaced bins for this evaluation. These binned metrics are shown in Table 7. We observed two key results:

1. Activation rate increases when using modified advice. This increase is highest when using the sigmoid-like function. It is similar for both step functions.

   Additionally, note that when looking at the non-binned activation rates, the high confidence step function activation rate actually decreases relative to the baseline. This suggests that the high confidence step function advice is only used by certain groups of people (i.e., based on their confidence in the tasks), and may not be as useful depending the distribution of users. This is not the case for the sigmoid-like modification function.
2. Accuracy and confidence increase when using the modified advice. This increase is similar for the sigmoid-like function and the high confidence step function, but is lower for the low confidence step function.

Table 6: Summary of results comparing step function transforms of AI advice. "before / after" refers to before and after advice. "Confidence" refers to the confidence in the correct label (which is negative if the user responds incorrectly). All metrics are averages across our population.

AI advice	# Participants	# Observations	Activation Rate	Accuracy (before / after)	Confidence (before / after)
Baseline	25	800	0.42	69.0% / 79.2%	0.33 / 0.44
Sigmoid-like	28	895	0.54	69.9% / 82.7%	0.35 / 0.57
Step function (low confidence)	23	736	0.57	63.2% / 79.8%	0.19 / 0.39
Step function (high confidence)	24	768	0.36	70.3% / 81.5%	0.37 / 0.56

Table 7: Summary of results using binning to reduce variation based on user performance across AI advice groups. We place each human-AI interaction in a bin based on the human's initial response, and average each metric across bins.

AI advice	Activation Rate	Accuracy Delta	Confidence Delta
Baseline	0.56	19.0%	0.19
Sigmoid-like	0.77	28.5%	0.39
Step function (low confidence)	0.69	26.5%	0.29
Step function (high confidence)	0.71	29.1%	0.39

Taken together, these results suggest that the sigmoid-like function is the best choice out of the 3 functions. It has comparable performance increases to the high confidence step function but also has the highest increase in activation rate, suggesting it is more useful for a broader category of human-AI interactions.