Artificial Work: AI Use and Moral Hazard in Online Labor Markets

Tara Kuruvila[*1], Suresh Naidu[†2], and Lena Song[‡3]

[1] *Columbia University*
[2] *Columbia University*
[3] *University of Illinois Urbana-Champaign*

Preliminary draft - please do not circulate
September 6, 2026

Abstract

Unobservable AI use by workers has become a threat to data integrity on platforms used by both industry and research. We build and validate a suite of AI detection tools, including a lab-generated "ground-truth" surrogate index, that reliably detect both human AI use and automated agentic survey responses. We then run a high-powered (N $\approx$ 6000), multi-platform suite of experiments looking at economic determinants of AI use in data provided by online workers. Randomizing bonuses and monitoring, we find a substantial role for moral hazard, where respondents paid a higher incentive, or monitored less, are more likely to use AI. These effects are not driven by higher effort overall, as higher incentives increase effort, and monitoring does not. Accuracy increases with AI use, and not with effort. Decreasing the manual ease of using AI reduces AI use and accuracy, while increasing effort. Randomizing base pay, we replicate previous residual labor supply elasticity estimates from online labor markets. We also find some evidence of adverse selection, as higher base pay attracts more automated AI responses on a platform where automated agents are prevalent. There is little evidence of significant interactions between treatments. Reweighting our data by predicted AI use, we show that the outputs of data labeling tasks and elicited behavioral measures would both be altered if the probability that a response is generated by a human is taken into account.

Keywords: Moral hazard, Data markets, Artificial Intelligence
JEL Codes: C83, C93, D82, D86, J41, L86, O33

[*]Email: tk3039@columbia.edu
[†]Email: sn2430@columbia.edu
[‡]Email: lenasong@illinois.edu

1 Introduction

Modern AI systems are extremely data demanding. The market for training data has tracked with the expenditure on compute, with data-labeling platforms like Scale AI, Mercor, Prolific, and CloudResearch generating billions of dollars in revenue. Human-generated data as an input, whether thought of as labor-like flows (Arrieta-Ibarra et al. 2018) or capital-like stocks (Veldkamp and Chung 2024), has expanded to historically unprecedented shares of business expenditure.

At the same time, the rapid adoption of AI by users, including online workers, has changed the market for data. AI is increasingly being used by respondents to complete online tasks, including answering surveys fielded by firms, political organizations, and researchers. Agentic AI can automatically complete a large number of tasks on data platforms (Westwood 2025). At the same time, threats of model collapse (Shumailov et al. 2024; Xing et al. 2025) mean that documenting human-vs-machine provenance of data, including recent innovations like watermarking images and text, is becoming increasingly important.

Motivated by the concerns from both researchers and industry about AI use on online platforms, we build and validate a suite of AI detection tools in online surveys. These rely on both question-specific measures as well as question-independent features like mouse and keystroke measures that are usable by others. We use a controlled laboratory environment to generate a "ground-truth" measure, and show that we can reliably detect both human AI use and automated agentic survey responses, measuring AI use on a broad set of tasks. Our measures correlate well with both platform-specific measures of AI use, where available, and other measures developed for text, like Pangram.

We then run a high-powered ($N \approx 6000$), multi-platform suite of experiments looking at economic determinants of AI use in online markets. We extend a well-known design in economics (Karlan and Zinman 2009) that randomizes pay at both the ex-ante contracting stage and again in the ex-post task completion stage. As in other markets with asymmetric information (Einav and Finkelstein 2023), the ex-ante variation in purchaser prices identifies both market power and adverse/advantageous selection, while the ex-post variation identifies moral hazard. We augment this design with interventions that increase perceived monitoring, making the detection probability and cost of being caught more salient. Finally, we benchmark our interventions to a treatment arm that increases the friction of using AI, by disabling copy-pasting and triggering pop-up windows when workers switch to other tabs.

We find a substantial role for moral hazard, where respondents paid a higher incentive, or monitored less, are more likely to use AI. These effects are small but economically meaningful: a 10x increase in incentives raises our preferred AI use indicator by 2 percentage points, over a base rate of roughly 29%. The magnitude of the effect of telling respondents they will be sanctioned for AI use, or of making our AI detection salient, is about twice as large. These effects are similar to the reduction in AI use when we add a set of features designed to make AI use more difficult. Bonuses increase both actual human effort as well as AI use, while sanctions reduce AI use without increasing effort.

Turning to the effects of randomizing base pay, we replicate previous residual labor supply elasticity estimates from online labor markets, but find little evidence of either adverse or advantageous selection into assisted AI use. While high ex-ante pay does draw in workers with higher reservation wages and greater skill, it does not draw in workers with any different propensity to use AI, either stated or measured. On the other hand, we find adverse selection into automated AI use only on the platform where the prevalence of automated agents is high, and no effect of base pay on platforms with stricter rules governing automated responses.

We explore interaction between treatments and heterogeneity using pre-specified demographics.

While we find little evidence of significant interactions between treatments, we find a pattern of heterogeneity where men react more to bonuses with AI-use.

We offer guidance to researchers worried about AI contamination. Reweighting our data by predicted AI use, we show that the outputs of data labeling tasks and elicited behavioral measures would both be altered if the probability that a response is generated by a human is taken into account. Multiple-choice questions on public opinion or political attitudes, on the other hand, are unaffected, which is what we would expect if respondents rarely use AI assistance on questions that have no right answer. As models improve in performance, AI-detection is likely to be stochastic at best, so we view reweighting the data by predicted AI-use, as in DiNardo, Fortin and Lemieux (1996), as a way to form distributions of variables that mimic, as best possible, those that would come from purely human sources.[1]

Finally, we interpret our results through the Holmstrom-Milgrom model of a multi-task principal-agent model (Holmstrom and Milgrom 1987, 1991). The output from the (human) agent is a data point that has both human effort and AI-output in it. The principal gets both a signal of output as well as a signal of data provenance (AI or human). We show that our experimental results map on to parameters of the agent's decision. In the discussion, we then relate our estimates to the optimal contract for data buyers on online platforms.

A recent crisis in AI use in online labor markets has led behavioral science researchers to debate whether widely used online platforms like Prolific and CloudResearch are being contaminated by AI. Consider behavioral surveys, where often the goal is to measure human biases (like failures of Bayesian rationality or social preferences like zero-sum thinking). AI use in online platforms is likely to skew measurements of these human biases, to the extent that LLMs do not exhibit them in the same manner as humans.[2]

Our paper contributes to the emerging literature on data contamination by AI use on online platforms. Recent research shows that modern LLMs are capable of completing online surveys (Westwood 2025; Van der Stigchel et al. 2026), that a large fraction of online workers use them (Veselovsky et al. 2025; Veselovsky, Horta Ribeiro and West 2023; Zhang, Xu and Alvero 2025; Velutharambath et al. 2026), and that AI use can be detected from text and behavior (Jabarian and Imas 2025; Claassen et al. 2026; Kay and Vlasceanu 2026; Rilla et al. 2026) and prevented (Çelebi et al. 2026; Fernandez et al. 2026). The general view is that agentic bots concentrate on a few platforms while human-assisted LLM use is more pervasive (Gordon et al. 2026; Zhang et al. 2026), a pattern our detection indices reproduce. This matters because even 3 to 7 percent contamination is enough to distort estimates (Panizza, Kyrychenko and Roozenbeek 2026). We add to this literature by treating AI use as an economic choice, varying incentives and monitoring to estimate how much each one moves AI use.

More broadly, our paper contributes to the literature on imperfections in online labor markets. Earlier work such as Horton and Chilton (2010) characterizes the reservation wages and labor supply of crowd workers, and we replicate the low residual elasticity of labor supply found on Amazon Mechanical Turk by Dube et al. (2020). Our interventions map closely to work on monitoring and incentivizing the quality of online work (Mason and Watts 2009; Ho et al. 2015; Liang et al. 2023; Holtz, Scult and Suri 2022; Pallais 2014). We add to this work by accounting for the influx of AI use on online platforms, and for how the unobservability of AI use may contaminate measures of quality in the resulting work.

Our third contribution speaks to what happens to models, and to research, when their inputs

[1] For AI training, however, a "Detect-and-Discard" approach may be preferable, as training efficiency falls with the share of non-human responses.

[2] For evidence on behavioral biases in LLMs, and their similarity and differences, see Araujo and Uhlig (2026) and Pate et al. (2026).

stop being human. Shumailov et al. (2024) show that training on recursively generated data causes irreversible loss of the tails of the original distribution; Alemohammad et al. (2024) document the same self-consuming dynamic in generative image models, Dohmatob et al. (2024) recast it as a change in scaling laws, and Xing et al. (2025) survey the resulting "autophagy" literature, offering provenance and watermarking as defenses. Our paper studies the decision of humans to use AI in data generation in response to economic incentives.

The rest of the paper proceeds as follows. Section 2 sets out a simple framework for the worker's decision to use AI, and Section 3 builds and validates our detection indices; Sections 4 and 5 describe the experimental design and estimation, Section 6 reports the results, and Section 7 concludes.

2 Conceptual Framework

In this section, we offer a simple variant of the Holmstrom and Milgrom (1991) model of multitasking to frame our experiment and interpret the results. The key intuition is that AI use by a data worker (the agent) is difficult to monitor and yet affects the output that the data buyer (the principal) receives, and so high-powered incentives may backfire, depending on principal objectives.

Consider a data worker with reservation utility u_i who chooses two actions, quality effort $t_q \geq 0$ and AI reliance $a \in [0, 1]$, at private cost

$$C(t_q, a) \;=\; \tfrac{1}{2}k_q t_q^2 \;+\; \tfrac{1}{2}k_a a^2, \qquad k_q, k_a > 0.$$

The buyer observes two signals, an accuracy score x_1 and an AI detector x_2,

$$x_1 = t_q + \lambda a + \varepsilon_1, \qquad x_2 = -\delta a + \varepsilon_2, \qquad \varepsilon \sim N(0,\ \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We assume the data quality is increasing in worker effort $t_q > 0$ but also in AI use, so $\lambda > 0$, along with an idiosyncratic shock ε_1.

We assume that the buyer only noisily observes AI use a with parameter δ and i.i.d. noise ε_2. We set the variances of the two noise terms to 1 without loss of generality, as the precision of AI use a will be parameterized by δ. The errors are correlated, $\rho \neq 0$, because both signals are computed from the same response data. A very high quality response may be flagged as AI-written (or not).

Suppose the buyer/experimenter offers the linear contract $w = \alpha + \beta_1 x_1 + \beta_2 x_2$. Suppose also that the worker has CARA preferences with risk aversion r, while the buyer is risk neutral, with a marginal value of b_q on data quality t_q, and a value γ on AI use a. The buyer may want a "clean" human-provided signal, with $\gamma < 0$. We think this case is more likely for both behavioral science researchers and AI-training labs, given concerns about data-provenance discussed above. But there may be cases where the data buyer only cares about accuracy x_1 regardless of provenance.

As AI gets better, λ rises (AI substitutes for quality effort in the score), k_a falls (AI is cheap for the worker), and δ falls while σ_2^2 rises (detection becomes harder).

Agent. The certainty equivalent is $\alpha + \beta_1(t_q + \lambda a) - \beta_2 \delta a - C(t_q, a) - \frac{r}{2}\beta'\Sigma\beta$, so the agent's choices are given by:

$$t_q^* = \frac{\beta_1}{k_q}, \qquad a^* = \frac{\lambda\beta_1 - \delta\beta_2}{k_a}. \tag{1}$$

Experiment Comparative Statics: Our experimental design varies α (base pay), β_1 (bonus), $\delta\beta_2$ (sanctions/monitoring/salience) and k_a (cost of using AI). Equation (1) shows that increased incentives for accuracy will increase both effort t_q and AI use a, as well as overall accuracy. Increased

cost of using AI, k_a will lower AI use. Increased salience of monitoring and sanctions, $\delta\beta_2$ will also lower AI use. Base pay does not appear as a determinant of either effort or AI use. For any individual worker, the extensive margin of labor supply is separable from the intensive margin of effort. However, across workers, if reservation utility u_i is correlated with either cost parameters k_q, k_a or signal structure, there may be effects of base pay α on effort and AI use a. We return to this possibility when discussing the possibility of adverse/advantageous selection and the role of monopsony power below.

3 How to Detect AI?

In this section, we define two types of AI use in online tasks, AI-assistance and AI-automation, identify a suite of behavioral markers that can be tracked in an online survey, and combine those markers into indices for each type of AI use. To learn how these markers look with and without AI use, and to fit the indices that require labels, we collect ground truth from an in-person lab study and an online benchmark in which participants are directly instructed whether to use AI, and a synthetic sample constructed using automated agents. We report how well the indices separate AI from human responses out of sample, and then describe how they allow us to infer AI use in experimental samples across platforms, where no labels are available.

3.1 Defining AI Use

We conceptualize two types of AI use in online labor markets. Responses are *AI-assisted* when a worker uses generative AI for part or all of the task but remains involved. In the online survey setting, this means the worker uses AI on some or all of the questions, whether editing what the model returns or transferring it verbatim, but is still clicking through the survey.[3] Responses are *AI-automated* when the worker fully delegates the task, so that it is completed without human involvement.

We apply these definitions to the setting of an online survey. A response is AI-automated when the worker hands the survey to an AI agent that completes it end to end, so that the agent, rather than the worker, is the one going through the survey. On the other hand, a worker who pastes model output without reading it is AI-assisted because they are still completing the survey themselves. The distinction is whether the worker is involved in completing the task rather than how much of the answer is from an AI model.

By definition, AI-automation is a property of the whole survey. We similarly define our main measure of AI-assistance at the survey level, since no treatment outlined in Section 4.3 varies within a survey.

3.2 Detection Variables

We use a set of detection variables drawn from the literature (Çelebi et al. 2026; Rilla et al. 2026) and our own testing. While which variables are available varies with the nature of the survey question (for example, keystroke and text-based measures apply only to text-based responses), the variables are general enough that any online survey can adopt them, and some can be collected for any type of question. The main tools are JavaScript, which we use to collect our own measures,

[3] Using tools to answer surveys when they are not permitted is not new. What distinguishes generative AI from earlier tools (such as search engines) is that a complete answer is much easier to obtain, and therefore a closer substitute for quality effort.

and existing text-based or behavior-based AI detectors. Many of these variables leverage behavioral differences similar to CAPTCHAs and other existing AI screeners.[4]

Appendix Tables A.1 and A.2 provide the full set of detection variables for AI-assisted and AI-automated responses, respectively. To detect AI-assisted responses, we use variables spanning five broad categories: clipboard activity (e.g., copy and paste events), keystroke behavior (e.g., typing speed), page engagement (e.g., tab switching), cursor and scroll behavior (e.g., mouse movements), and an AI detection tool for text responses.[5]

For automated AI detection, we use a similar set of variables, with a few modifications. First, we exclude page-engagement variables that indicate browsing other pages (e.g., tab-switching and window-blurring), as automated agents do not need to navigate browser tabs. Second, we add respondent-level checks for failure on questions that are difficult for automated agents, such as reCAPTCHA. Third, we flag suspicious IP addresses. For both assistance and automation, all variables are oriented so that higher values indicate a higher probability of an AI response.

These categories of AI detection variables capture different dimensions on which human and AI responses differ. For example, clipboard activity and page engagement are evidence of traffic between the survey and an external model. Keystroke behavior measures what the worker is doing on the keyboard and where the entered text comes from. These variables detect AI use through inputs to a survey response. On the other hand, a text-based detector such as Pangram detects AI use from the output.

3.3 Index Construction

The individual detection variables carry correlated signal about AI use. As there are many measures across detection variables and across pages within a survey, we aggregate these variables into surrogate indices of AI-assisted and AI-automated responses. The identifying assumption is that the true outcome of the experiment is orthogonal to treatment status conditional on the surrogate index and covariates (Athey et al. 2026). That is, any effect of the experimental assignments on the true outcome would be picked up by the effect on the surrogate.

We build three indices, with the last relying on the correlation between AI use/automated AI in our auxiliary training data. The first index, the pre-specified index, is rule-based and calculates a binary variable using the rules specified in our pre-analysis plan. It does not require sample information, which also means that it does not account for new information about how people use AI over time. The other two indices are data-driven, and differ in what they require from the data. The second index, the PCA index, applies principal component analysis to the joint distribution of the detection variables to reduce dimension and calculate a continuous measure of AI use. The third index, the LASSO index, in addition requires a ground-truth training set in which we know whether AI was used, and predicts AI use out of sample.

The pre-specified binary index is an indicator equal to one if a participant meets at least one of several detection criteria. For AI-assisted responses, these criteria are: behavior consistent with pasting text into a response field, copying the full prompt of a question, or Pangram classifying the response as partially or completely AI-generated. For AI-automated responses, these criteria are: no editing on any question, mouse movement or keystroke behavior that is hard for humans to achieve, or leaving the survey at questions where AI agents struggle to answer (such as reCAPTCHA).

[4]Recently, companies have begun efforts to watermark AI output. This has been successful in image generation, where infusing invisible signals into pictures works well, but less so in the domain of copy-pasted text, as the watermarks have to live at the level of the distribution of words, and hence can be broken by rewriting.

[5]We use Pangram, which has the lowest false positive and false negative rates among leading commercial and open-source detectors, including on short passages (Jabarian and Imas 2025).

For the data-driven indices, we use data from ground-truth samples where we instruct workers to use AI or not to use AI, and a synthetic sample where we build and run our own AI agents to complete the survey, as detailed in Section 3.4. This gives us labeled samples where we know whether a response is human-only, AI-assisted, or AI-automated. Within each ground-truth sample, we standardize variables using the sample mean and standard deviation, reorient each variable so that higher values reflect greater AI use, and impute missing values using the sample mean.

The PCA index uses the training samples to construct weights. We take the loadings on the first principal component of the standardized variable matrix in each training sample and apply them out of sample. As the sign of a principal component is arbitrary, we align each PCA index with the pre-specified binary to ensure that higher values always correspond to more AI use.

We construct the LASSO index using the fitted value from an ℓ_1-penalized linear probability model of the AI-use label on all standardized detection variables, with the penalty parameter selected by cross-validation to minimize out-of-sample prediction error. While the PCA coefficient has no direct interpretation, a LASSO coefficient gives the change in the predicted probability that a response is AI-assisted or automated associated with a one standard deviation increase in that variable.

Agreement across these three indices is informative because each relies on different information. The LASSO index uses the labels from the labeled samples and can therefore adapt to AI-assisted human behavior and AI agent behavior at the time of the study. The pre-specified binary cannot have been overfitted. The PCA index is data-driven but does not use the labels or perform selection.

We also build variants of these primary indices that exclude certain variables from the feature set, to help interpret the treatment effects of the interventions described in Section 4.3. One of the interventions we study raises the cost of using AI, in part by disabling copy-pasting within the survey and displaying a pop-up when the respondent switches tabs. In the no-copy-paste variant of the index, we exclude the copy and paste indicators along with the keystroke-log variables that are mechanically affected when copy-pasting is blocked (empty keystroke log, input-jump count, keystroke event count, and keylog-reconstructed similarity) and variables recording tab-switching behavior. This allows us to capture changes in AI use beyond what the intervention mechanically prevents. To disentangle AI use from worker effort, we build a no-effort variant of the index that excludes variables that could also change when a worker puts in more effort, such as the number of keystrokes, inter-keystroke timing, time spent on the question, and mouse behavior (e.g., scroll direction switches and cursor movement patterns) and editing behavior (e.g., backspaces).

3.4 Labeled Samples and Ground Truth

For the data-driven indices, we construct three labeled samples.[6] To understand how the detection variables look for AI-assisted responses, we use an in-person lab sample and an online survey sample. For the in-person sample, we recruited undergraduate students from Columbia and Barnard through the Columbia Experimental Laboratory for Social Sciences (CELSS), and the students completed the survey under monitored conditions. In a within-subject design, each student answered one version of each question type with AI assistance and one version without, randomizing both the order of conditions and which question within each type was assigned to the AI condition. In-person monitoring ensures high compliance, so the AI/no-AI assignment serves as ground truth.

The second sample, which we refer to as the training sample, draws participants from Prolific. In a between-subject design, participants were randomly assigned to one of two conditions, either

[6]Collecting labeled samples is not a prerequisite for using our detection methodology. Researchers can apply the index weights we estimate here directly to their own surveys, or use our labeled samples to estimate weights for their online survey.

instructed to use AI to complete the survey or instructed not to use AI. We use both an in-person and an online labeled sample because there is a trade-off between compliance and representativeness. In-person monitoring ensures high compliance in the CELSS labels, but the sample is drawn from a population of undergraduate students who complete the survey on lab computers. In the training sample the population is the one we study in the field and participants use their own devices, but because we randomize the instruction rather than the behavior, participants told not to use AI may use it anyway, and vice versa, leaving some responses mislabeled. In Section 3.5, we test how well AI-assisted and human responses can be separated.

To understand how AI-automated responses look, we build AI agents to complete our survey and construct a synthetic sample. These agents can be viewed as digital twins of the participants in the CELSS and Prolific studies who completed the survey without AI assistance, as we include the demographic information of the participant in the prompt provided to the corresponding agent. To allow for different levels of agent sophistication, we collect responses both from off-the-shelf agents that use current models and are simply told to answer the survey, and from more sophisticated agents that incorporate a layer of code imitating human mouse movements. For the data-driven indices for AI-automated responses, we use data from this synthetic sample and from the CELSS and training samples under the conditions in which participants were instructed not to use AI.

3.5 Detector Performance

Before we apply the indices to samples in which we do not know who used AI, we test how well they separate AI from human responses out of sample, using the two labeled samples, CELSS and training. In this cross-sample test, we fit each index on CELSS and score it on Training, and vice versa.

Table 1 shows that the indices perform well out of sample. The full LASSO reaches an AUC of 0.98 when fitted on the training sample and scored on CELSS, and 0.80 in the other direction, against ten-fold cross-validated values of 0.98 within CELSS and 0.83 within the training sample.[7] The PCA index, which uses no labels, transfers at 0.94 and 0.79. The no-copy-paste and no-effort variants transfer similarly well. This provides evidence that what we detect is AI use rather than a feature of a particular sample.

[7]The CELSS sample uses a within-subject design, so the same respondent supplies both the AI and the no-AI observation, and in-sample separation is correspondingly high. The Training sample uses a between-subject design, does not directly enforce compliance, and therefore separates its own classes less sharply.

Table 1: Cross-Sample AUC for the AI-Assisted Indices

	CELSS in-sample	CELSS-trained on Training	Training in-sample	Training-trained on CELSS	CELSS 10-fold CV	Training 10-fold CV
PCA (full)	0.948	0.790	0.807	0.936	0.941	0.806
PCA (no-cp)	0.959	0.755	0.750	0.952	0.953	0.750
PCA (no-effort)	0.964	0.791	0.793	0.962	0.959	0.792
LASSO penalized (full)	0.989	0.804	0.863	0.980	0.984	0.828
LASSO post-sel (full)	0.992	0.803	0.883	0.973	0.986	0.820
LASSO penalized (no-cp)	0.986	0.777	0.850	0.971	0.983	0.815
LASSO post-sel (no-cp)	0.988	0.774	0.862	0.957	0.985	0.809
LASSO penalized (no-effort)	0.984	0.807	0.864	0.975	0.981	0.819
LASSO post-sel (no-effort)	0.986	0.812	0.866	0.951	0.984	0.821

NOTES: This table reports the area under the ROC curve for the AI-assisted family of indices across its feature-set variants, one row per variant. Each index is fitted on one labeled sample and scored on the other, so the sample supplying the labels is different from the sample supplying the evaluation. The cross-sample columns indicate whether a model generalizes beyond the sample it was fitted on; the in-sample columns are optimistic by construction and serve only as a reference point. The final two columns give the ten-fold cross-validated AUC within each sample, which is the appropriate within-sample benchmark for the cross-sample columns.

The indices aggregate across a set of detection variables. As shown in Table 2, most of these detection variables separate the AI and no-AI arms of each labeled sample in which AI assistance is randomly assigned. This suggests that, for detecting AI assistance, no single detection variable carries the measurement on its own, and that it is useful to pool a set of signals that capture AI use through different markers. As a placebo, we see that performance on attention checks and survey completion do not separate the two arms.

Table 2: Index Features Separating the No-AI and AI Arms, by Family

	CELSS		Training	
Feature family	Features	Separating	Features	Separating
Copy and paste	22	18	22	18
Keystroke dynamics	40	20	41	19
Navigation and focus	27	23	36	21
Response timing	18	6	18	6
Text detector	2	2	3	3
Attention and missingness	3	0	1	0
Total	112	69	121	67

NOTES: The table reports, for each family of detection features, how many features enter the index and how many separate the no-AI from the AI arm of a labeled sample. The unit of observation is a feature, not a respondent. "Features" counts the features in that family that enter the index for the sample concerned, and "Separating" counts those whose difference in means across the two arms is significant at the five percent level in a two-sided t-test. The two samples are CELSS and the Training sample. The feature counts differ across the two samples because a feature enters only where the question it is measured on was fielded and the underlying variable was collected. No p-value here is adjusted for multiple testing, so the counts describe how widely signal is spread across the feature set rather than constituting a joint test.

We note two caveats about what these variables and indices can and cannot detect. First,

the indices should not be read directly as the probability that a response is AI. Recovering levels requires the reweighting outlined in Section 5.2, and at the respondent level we use the indices to reweight rather than to classify and drop individual respondents. Second, our labels come from agents we built ourselves using models available at the time of the study, and other agents may behave differently.

3.6 Detection in New Samples

In new samples without labels, we construct indices as described in Section 3.3. In the main sample of participants from Prolific, Table 3 shows that all three of our indices, the pre-specified index, PCA, and LASSO trained on either CELSS (C) or Training (T), take significantly different values for responses that Prolific flags as AI than for those it does not. Appendix Table A.3 shows that every index takes a substantially higher value among responses Pangram flags as AI-generated than among those it does not.

Table 3: Our Measures Against Prolific's Own AI Flags

	Prolific bot			Prolific LLM		
	Not flagged	Flagged	Diff	Not flagged	Flagged	Diff
Prespec	0.343	0.519	0.175*	0.309	0.973	0.663***
PCA-C	0.490	0.710	0.220***	0.474	0.895	0.421***
LASSO-C	0.208	0.449	0.241***	0.185	0.661	0.477***
PCA-T	0.491	0.713	0.222***	0.474	0.899	0.425***
LASSO-T	0.325	0.428	0.104***	0.308	0.649	0.341***
Pangram	0.072	0.222	0.150***	0.050	0.464	0.414***
Self	0.029	0.000	-0.029	0.023	0.121	0.098***
Attn: video	0.890	0.926	0.036	0.888	0.896	0.007
Attn: IMC	0.633	0.704	0.070	0.640	0.462	-0.178***
N (respondents)	2693	27		2773	182	

NOTES: The table reports the mean of each of our AI-use measures separately by Prolific's own bot and LLM flags. The unit of observation is a respondent. The sample is Prolific only, since no comparable platform flag exists on MTurk or CloudResearch. N differs across rows because the flag is missing for some respondents. Stars are on the difference in means between flagged and unflagged respondents: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4 reports the correlation matrix across the three indices, Pangram, self-reported use, and the two attention checks. The two PCA indices correlate almost perfectly with one another, and the behavioral indices correlate positively with Pangram, most strongly for the LASSO index. The pre-specified, PCA, and LASSO indices are moderately correlated with one another, indicating that they capture different dimensions of AI use. Self-report correlates only weakly with any of the behavioral measures, consistent with existing evidence that AI use is not something a researcher can simply ask about (Ling, Kale and Imas 2026). Passing an attention check using instructional manipulation is negatively correlated with AI use, suggesting that respondents with AI assistance were also less attentive.

Appendix Figure A.1 plots the distribution of the LASSO predicted probability for human versus AI-assisted responses. There is no clear threshold, and this spread is what motivates the reweighting in Section 5.2 rather than the classification of individual responses as AI or human.

Table 4: Correlations Among the AI-Use Measures

	Prespec	PCA-C	LASSO-C	PCA-T	LASSO-T	Pangram	Self	Attn: video	Attn: IMC
Prespec	1.00								
PCA-C	0.54	1.00							
LASSO-C	0.39	0.50	1.00						
PCA-T	0.56	0.96	0.51	1.00					
LASSO-T	0.65	0.64	0.63	0.66	1.00				
Pangram	0.56	0.35	0.36	0.37	0.62	1.00			
Self	0.25	0.07	0.08	0.08	0.21	0.30	1.00		
Attn: video	-0.00	0.04	-0.01	0.04	-0.01	-0.01	-0.07	1.00	
Attn: IMC	-0.21	-0.10	-0.13	-0.10	-0.22	-0.25	-0.23	0.08	1.00

NOTES: The table reports pairwise Pearson correlations among the AI-use measures and two attention checks. The unit of observation is a respondent who completed the main survey, pooled across all platforms (Prolific, CloudResearch, and MTurk). "Prespec" is the pre-specified binary rule; "PCA-C" and "LASSO-C" are the two indices trained on the CELSS labeled sample, and "PCA-T" and "LASSO-T" the two trained on the Training sample; "Pangram" is the AI-text detector score on open-response answers; "Self" is the respondent's own statement that they used AI; "Attn: video" and "Attn: IMC" are the video comprehension check and the instructional manipulation check, each coded 1 for a correct answer. Correlations involving Pangram are computed on respondents with at least one scored open-response answer. Only the lower triangle is shown.

3.7 Base-Rate Correction

Our indices cannot be interpreted directly as a probability of AI use. While our primary analysis focuses on the treatment effects of our interventions and therefore differences out any under- or over-estimation in levels, results stated in levels require a base-rate correction.

To see this, let X denote the observed features that describe each response. Our labeled sample provides, for each response, the features X and a ground-truth AI label, while a new field sample provides X but no label. From the labeled sample we train a detector $s(X)$, a scalar summary of the features. Reading the detector as a posterior probability $\Pr(\text{AI} \mid X)$ requires two ingredients: a likelihood (what AI and human answers look like in X-space) and a prior (how common AI use is). The prior in the labeled sample differs from the unknown prior in a new field sample. Using the raw index assumes that the two priors are equal, which biases every predicted probability toward the AI rate in the training sample (close to 50 percent), no matter how accurate the detector is.

We adjust for differences in the base rate across samples in two steps. First, we estimate the overall rate of AI use in the new sample. The observed distribution of scores, $q(s)$, is a mixture of the two stable components:

$$q(s) = \pi\, g_1(s) + (1 - \pi)\, g_0(s) \tag{2}$$

where g_1 and g_0 are the densities of $s(X)$ among AI and among human responses, identified in the labeled sample, q is observed in the field sample, and the mixing proportion π is the only unknown.

The key assumption we make is measurement stability. We assume that the distribution of detector scores among AI responses, and among human responses, is the same in the labeled sample as in the new sample, and what differs across samples is only the share of responses of each type.

Taking the mean of both sides of Equation (2) gives the black-box shift estimator:

$$\hat{\pi}_{\text{field}} = \frac{\mu_q - \mu_0}{\mu_1 - \mu_0} \tag{3}$$

where μ_1 and μ_0 are the mean detector scores among known AI and known human responses in the labeled sample and μ_q is the mean score in the new sample. Identification requires $\mu_1 \neq \mu_0$: the detector must respond differently to AI vs. human responses.

Second, we use this estimated rate to re-scale each respondent's predicted probability of AI use. The log-odds of the predicted probability decompose into a term that depends only on the response's features, which by the measurement stability assumption holds fixed across samples, and a term that depends only on the base rate. We therefore subtract the base rate of the labeled sample and add an estimate of the base rate in the new sample:

$$\operatorname{logit} p_i^{\text{field}} = \operatorname{logit} p_i - \operatorname{logit} \pi_{\text{lab}} + \operatorname{logit} \hat{\pi}_{\text{field}} \tag{4}$$

where p_i is the detector's calibrated predicted probability for respondent i and π_{lab} is the AI rate in the labeled sample.

There are two caveats to note for these results. The detector separation $\mu_1 - \mu_0$ is measured in-sample on the data the LASSO was fitted to, so it is optimistic. Since separation is the denominator in Equation (3), an inflated value biases $\hat{\pi}_{\text{field}}$ downward, and the estimate should be read as a lower bound on the field AI rate. The mean re-based posterior also does not land exactly on $\hat{\pi}_{\text{field}}$. The two coincide only when the underlying score is a calibrated probability, and ours is a penalized linear-probability prediction, so the gap is a diagnostic of that approximation rather than of the correction itself. We report both quantities in Appendix Table A.4 and alongside the corrected means.

The detector separates AI from human responses in the labeled samples, with a gap of 0.724 between the mean score among known AI and known human answers. The denominator of Equation (3) is therefore far from zero. Applying the base-rate correction to our field sample, we find an AI use rate of 16.6% in the Prolific field sample, substantially lower than the unadjusted mean score of 0.254. The bootstrap standard error on $\hat{\pi}_{\text{field}}$ comes from resampling the labeled and field samples jointly.

4 Experimental Design

4.1 Setting and Recruitment

We ran several experiments across multiple platforms to detect AI use and to test related interventions, as outlined in Section 4.3. We recruited 5580 participants (or over half of our sample) from Prolific, an online survey platform widely used in social science research as well as in data labeling. The inclusion criteria are: aged 18 to 64, living in the United States, and not having participated in any of our pilot studies. To recruit a sample representative in age and gender, we imposed recruitment quotas matching 2024 U.S. Census proportions. Once a listing reaches ten places, Prolific displays an estimated pay rate based on the time earlier participants spent on the survey rather than the listed rate. As treatment assignment could affect time spent on the survey, we capped each listing below ten participants and split recruitment into many small listings fielded across six waves, so that participants only ever saw the listed rate. This avoids prospective participants being influenced by the behavior of earlier participants, which would violate SUTVA.

We fielded the main survey on two additional platforms, Amazon Mechanical Turk and CloudResearch Connect. These three platforms differ in their worker pools, reputation systems, screening options, and costs, so we adjust our recruitment and survey accordingly, as outlined in Section 4.4. We also compare across platforms to understand whether the market failures we document are specific to particular platforms or extend to online survey markets more generally.

4.2 Survey Overview

Participants across platforms completed a main survey consisting of a screener and consent form, demographic questions, a battery of tasks presented in randomized order, a module on attitudes toward AI, and a post-task section that elicited self-reported AI use. The task battery was designed to reflect the range of activities common in behavioral and social science surveys and in data labeling for machine learning.

The battery includes four types of tasks. The first are questions commonly used by psychologists and behavioral economists to measure behavioral biases. In particular, we use questions from Bordalo et al. (2026) involving probability reasoning, such as balls-and-urns problems. These questions have a correct answer, which allows us to measure performance. In the Prolific survey, we asked both an easy and a difficult probability question, so that we can examine how question difficulty affects AI use. The second are beliefs and attitudes, which are common in public opinion and behavioral science surveys where participants share their opinion and there is no right or wrong answer. In particular, we use questions measuring zero-sum thinking following Chinoy et al. (2026). The third involve writing, where we ask participants to summarize a short passage or respond to an open-ended prompt in their own words, producing text-based responses that allow us to apply text-based AI detection tools. The fourth are data labeling tasks, including text classification and entity identification representative of those used in machine learning pipelines, where there is a right answer. We also include questions from Anthropic's training dataset, where participants choose between two chatbot responses. These questions require judgment and have no right answer. We use a diverse range of tasks because how well AI can be detected, and its consequences for data quality, plausibly vary with task type.

4.3 Experimental Interventions

For the Prolific main survey, we cross-randomized four pre-specified interventions. Base payment was randomized at the listing level at the time of recruitment, before participants started the survey. Within the survey, we randomized three interventions at the individual level, each with equal probability: Bonus (no bonus, low bonus, or high bonus), Sanction (with or without sanction), and Prevention (with or without prevention). This yields a $2 \times 3 \times 2 \times 2$ factorial design with 24 unique experimental conditions.

Prolific participants were assigned to either a low or a high base payment displayed in the study listing.[8] To ensure that participants in the low and high base payment conditions are comparable, we randomized base payment at the listing level. Across six waves, we fielded one listing per single year of age (18–64) in every wave. Within each wave, we took each pair of adjacent ages and randomly assigned one to low and one to high base pay. Although we do not control who receives the invitation to our study on Prolific, this ensures that participants in the low and high base pay conditions are similar ex ante.

For the Bonus intervention, participants are randomly assigned to receive additional payment conditional on task performance. We randomized participants with equal probability into No bonus, Low bonus ($0.25 for a correct or high-quality response), or High bonus ($2.50, ten times the low bonus). The bonus was revealed after the screener and baseline questions, before the tasks.

For Prevention, we randomly assigned participants either to a version of the survey that only tracks detection variables, or to a version that, in addition to tracking, incorporates tools that raise the manual difficulty of using AI. The latter version blocks copy-pasting in text-entry fields

[8] Across experiments, base pay was set at $8 and $15 per hour. For a survey advertised at fifteen minutes, base pay was $2.00 in the low arm and $3.75 in the high arm.

and displays a pop-up window reminding participants to stay on the page whenever they try to switch browser tabs. These tools are not visible to participants who never attempt to copy, paste, or switch tabs.

For Sanction, we randomly assigned participants to either receive or not receive a message that combines moral persuasion, emphasizing the importance of providing authentic human responses, with an explicitly stated penalty for detected AI use. In addition, participants in the Prevention arm were asked, "Will you complete this survey on your own, without using AI tools like ChatGPT?" The intervention therefore combines an increase in the salience of the penalty for detected AI use with a potential increase in gift exchange.[9]

We also randomize the location of the question block asking about the respondent's AI use and attitudes, to either before or after the tasks. This varies the salience of AI and of potential monitoring of AI use, without explicitly stating a penalty or encouraging people not to use AI.

4.4 Cross-Platform Replication and the Measurement of Take-up

We followed a similar design for CloudResearch with a subset of the interventions in order to maximize power on the main treatment arms. We randomized base payment at the worker level, and within the survey we randomized Prevention (with or without prevention), Sanction (with or without sanction), and Bonus, keeping only the No bonus and High bonus arms. This yields a $2 \times 2 \times 2 \times 2$ factorial design with 16 unique experimental conditions. For MTurk, we fielded the full set of interventions, so the design matches Prolific.

On both CloudResearch and MTurk, we randomized base payment at the individual worker level rather than at the listing level. To do this, we started by fielding a short pre-survey that collected worker IDs for a large pool of participants, along with their demographics, platform tenure, and baseline AI attitudes. Using this pool and the platforms' worker-level access controls (qualifications on Mechanical Turk and allowlists on CloudResearch), we randomized base payment across individual workers, stratified on age, race, gender, and platform tenure, and delivered the main survey to each worker at their assigned base pay. This two-step procedure allows us to measure take-up and to describe who accepts at each base pay level.

We recruited for the main survey in waves from the pool of pre-survey completers. Within each wave, we posted the listings with low and high base pay simultaneously. As on Prolific, the other interventions are randomized within the survey at the worker level.

4.5 Outcomes

Our main outcomes are AI use, effort, performance, take-up, and completion. We construct AI use using the indices described in Section 3.3. We also report self-reported AI use, elicited toward the end of the survey after the worker completes the main tasks.

To measure effort, we use a set of behavioral markers. We capture typing effort (number of keystrokes), pauses and cognitive effort (median inter-keystroke interval), revision (backspaces and deletions), time on task (seconds spent on the question), and page engagement (scroll direction switches and cursor movement patterns).[10] We measure performance as the share of tasks for which the participant gives the correct answer, on the probability reasoning and data labeling tasks.

[9] We bundle these channels in our design to maximize power, so the estimate should be read as the effect of a treatment that combines several elements, each of which could impact AI use on its own.

[10] Because these variables also enter the indices for AI use, we drop them in the no-effort variant of the AI use index.

We measure completion among participants who start the survey. Differential attrition across treatment arms, such as sanction, could indicate that participants who intended to use AI select out of the survey when faced with monitoring and a penalty. For CloudResearch and MTurk, where we observe the pool of participants eligible for the main survey, we also measure take-up.

4.6 Follow-up Survey

On Prolific, we also conducted a follow-up survey one to three weeks after the main survey, inviting all workers who were randomized in the main survey. We use this follow-up survey to examine whether the interventions in the main survey affected who chose to return, a form of differential attrition relevant for panel and repeated-measurement designs.

5 Empirical Strategy

5.1 Estimating Equations

Our design cross-randomizes four interventions: base payment B_i, bonus payment A_i, prevention P_i, and sanctions S_i. Let Y_i denote an outcome for participant i (e.g., AI index or performance). Let $\mathbf{X}_i$ denote a vector of pre-treatment covariates (age, race, gender, income, and baseline AI attitudes) included to improve precision. All specifications include wave fixed effects to absorb calendar timing and platform-side changes between waves, and pooled specifications also include platform fixed effects. Our primary estimating equation is:

$$Y_i = \alpha + \beta B_i + \delta A_i + \sigma S_i + \pi P_i + \mathbf{X}_i'\gamma + \varepsilon_i \tag{5}$$

where standard errors are clustered at the listing level. Because base payment is randomized at recruitment, the coefficient β captures adverse selection, whether higher pay attracts a systematically different composition of workers, including AI-automated responses and AI-assisted participants. Because bonus payment is randomized within the survey after participants have enrolled, the coefficient δ captures moral hazard, whether participants already in the survey are more likely to use AI under higher-powered incentives. The coefficients σ and π capture the effects of sanctions and prevention, respectively.

We also estimate a fully saturated specification that replaces the linear bonus term with indicators for each bonus level j and adds the randomized placement of the AI-use block, R_i, equal to one when that block was shown before the tasks rather than after:

$$Y_i = \alpha + \beta B_i + \sum_j \delta_j \mathbf{1}[A_i = j] + \sigma S_i + \pi P_i + \rho R_i + X_i'\gamma + \varepsilon_i \tag{6}$$

where j indexes the levels of bonus pay (No Bonus, Low Bonus, High Bonus). Saturating in the bonus lets its effect vary freely across levels rather than imposing linearity, and lets us assess whether moving from no bonus to a low bonus differs qualitatively from moving from a low to a high bonus. The coefficient ρ captures the effect of salience. Asking participants about their AI use and attitudes before they begin the tasks could raise the perceived probability that AI use is being monitored, so ρ isolates the effect of monitoring salience alone, without explicitly telling participants not to use AI and holding the perceived penalty fixed.

For MTurk and CloudResearch, we also estimate take-up on the pool of participants who were invited to the main survey. Because this pool includes workers who never started the survey, and only base payment is randomized at the time of recruitment, we focus on the effect of base payment on take-up, controlling for covariates collected in the pre-survey and wave fixed effects.

We are primarily interested in the main effects of each treatment, and we examine interactions between the treatment indicators in exploratory analyses.

5.2 Reweighting for Human-Only Statistics

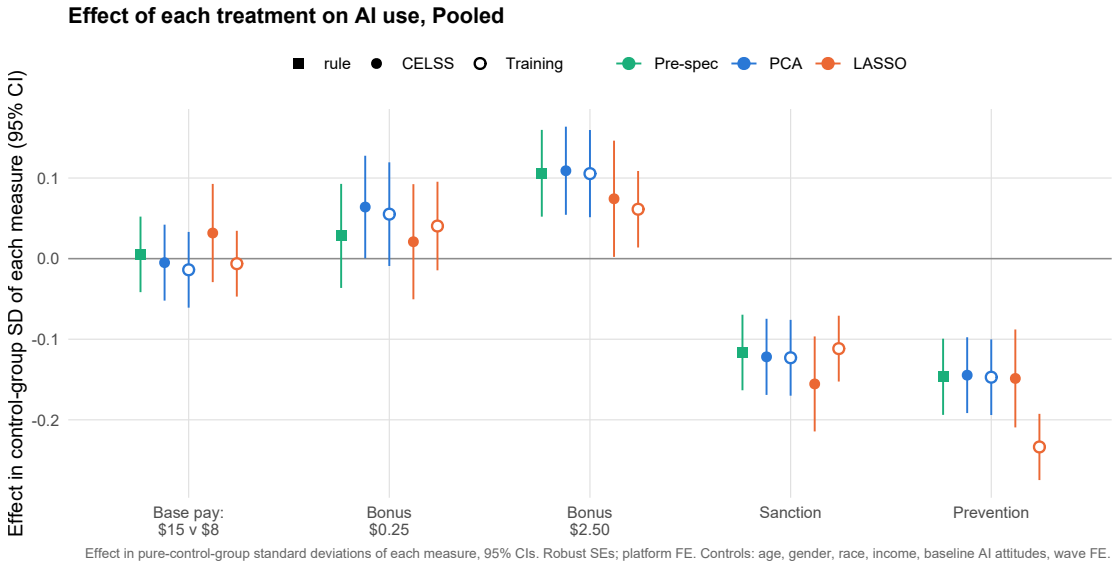
We also ask how much AI use changes the data, comparing the unweighted results with results reweighted by the estimated probability of AI use. Let p_i^{field} be the probability that respondent i used AI, re-based to the field as in Section 3.7. We apply a weight $w_i = 1 - p_i^{\text{field}}$ to each observation, and calculate the corrected mean of any outcome as the weighted average of that outcome using these weights. We bootstrap standard errors by resampling the labeled and field samples jointly to account for the sampling error in $\hat{\pi}_{\text{field}}$. The difference between this reweighted mean and the original mean tells us how the results would look if all answers had come from human respondents. This lets us account for AI use without setting a threshold and labeling individual responses as human or AI.

6 Results

6.1 Effect on AI-Assisted Use

We start by looking at how the experimental interventions affect AI assistance. Figure 1 plots the treatment effects on the AI-assistance indices from the saturated specification in Equation (6), pooled across the three platforms with platform fixed effects. See Appendix Table B.1 for the corresponding table with the bonus entering linearly in dollars, and Appendix Table B.2 for the saturated specification with indicators.

Figure 1: **Higher-powered incentives raise AI-assisted use, while prevention and sanction lower it and base pay does not move it.**



NOTES: The figure reports treatment effects on each AI-assistance measure, with 95% confidence intervals. Effects are in standard deviations of the pure control group's distribution of each measure, which is the axis the figure is drawn on, so they are not directly comparable to the 0–1 point coefficients in Table B.1. Estimates are pooled across Prolific, CloudResearch, and MTurk with platform fixed effects, and the indices are constructed from questions common to all platforms. Marker color denotes the index method (PCA or LASSO), and marker fill denotes the labeled sample the index was trained on (CELSS or Training). The bonus enters as its two dollar tiers, \$0.25 and \$2.50, against no bonus, and base pay as \$15/hr against \$8/hr. Controls are age, gender, race, income, and the four baseline AI attitudes, with wave fixed effects. Standard errors are robust.

Paying higher-powered incentives increases AI-assisted use. The Bonus raises the pre-specified index by 0.020 points per dollar ($p = 0.000$) against a control rate of 0.418, and the effect is positive and significant on both LASSO indices (0.008 CELSS-trained, $p = 0.038$; 0.004 Training-trained, $p = 0.028$) and both PCA indices (0.011, $p = 0.000$; 0.011, $p = 0.000$). While the pre-specified index applies a set of detection rules and the two data-driven indices exploit different variation in the data, with each trained on two different labeled samples, we find that these indices all move together and point to higher bonuses leading to higher AI-assisted use.

Self-reported AI use diverges from our surrogate indices. The Bonus moves self-reported AI use by 0.004 points ($p = 0.184$), indistinguishable from zero. Respondents offered more money for correct answers use more AI and do not report doing so.

Base Pay, by contrast, has little detectable effect on AI use. The coefficient on high Base Pay (\$15/hour vs \$8/hour) is statistically indistinguishable from zero on all measures, including the pre-specified rule (0.003, $p = 0.828$), both LASSO indices (0.009, $p = 0.308$; -0.001, $p = 0.757$), and both PCA indices. Base Pay was posted in the listing and could only operate by changing who started the survey, while the Bonus was revealed after starting and could only operate on people already in the survey. Higher pay does not attract a more AI-prone pool, while higher-powered incentives change the behavior of those who have already started. We therefore find evidence of moral hazard, but no evidence of adverse or advantageous selection into AI-assisted use. Section 6.3 shows that Base Pay does change the composition of the sample on other margins.

We can benchmark these effects against the effect of Prevention. Blocking copy-paste and displaying a pop-up after tab-switching lowers the pre-specified rate by -0.073 points ($p = 0.000$)

from a control rate of 0.418, and lowers both the LASSO and PCA indices. Against this benchmark, the high Bonus moves AI use by roughly two-thirds to three-quarters as much on the pre-specified rule and the two PCA indices, though by less on the two LASSO indices.

Sanction similarly reduces AI-assisted use, providing further evidence of moral hazard, as the Sanction raises both the perceived probability of monitoring and the cost of being caught. It lowers the pre-specified rate by -0.058 points ($p = 0.000$) and moves all four indices. Sanction also lowers self-reported use (-0.026, $p = 0.000$), while the reduction under Prevention is only marginal ($p = 0.069$).

The main results are robust to alternative constructions of the indices, as shown in Appendix Figures B.1 and B.2. In the no-copy-paste variants, which drop the features that Prevention mechanically affects, Prevention and Sanction stay negative and significant and Bonus stays positive, at magnitudes close to the main estimates. The no-effort variants, which drop keystroke counts, timing, time on page, and mouse behavior, give similar results. This suggests that the indices are picking up AI use rather than effort.

Appendix Table B.3 summarizes which detection variables each arm moves. For example, Prevention mechanically reduces the behaviors it blocks, such as copying and pasting, but it also reduces other indicators of AI-assisted use.

We also randomize the salience of monitoring, through the order of the AI use question block. The block asking about AI use and attitudes was randomized to appear either before or after the tasks, and seeing it first could increase the perceived probability that AI use is being monitored. We find that even questions about AI can move behavior substantially. As Table 5 shows, respondents who receive the AI use block before the tasks are less likely to provide an AI-assisted response, with an effect of -0.039 points on the pre-specified rule ($p = 0.026$), and similar effects on both LASSO indices (-0.039, -0.020), both PCA indices (-0.031, -0.030), and Pangram (-0.027). Self-report is the only exception, and is unmoved by the order of the block ($p = 0.928$). Simply making detection salient, with no explicitly stated penalty and no change in payoffs, produces a reduction of the same order as raising the manual cost of using AI. Beliefs about being monitored therefore play an important role in workers' decisions to use AI assistance online.

Table 5: Treatment Effects on AI Use, with AI-Use Block Placement

	(1) Pre-spec	(2) PCA (CELSS)	(3) LASSO (CELSS)	(4) PCA (Training)	(5) LASSO (Training)	(6) Self-Report	(7) Pangram
Base Pay	0.00562	0.00317	0.00141	0.00369	-0.00342	-0.00424	-0.00443
	(0.0175)	(0.00950)	(0.00966)	(0.00931)	(0.00575)	(0.00571)	(0.00874)
LowBonus	0.0260	0.0179	0.0100	0.0135	0.00814	0.00149	0.00377
	(0.0206)	(0.0120)	(0.0124)	(0.0122)	(0.00681)	(0.00698)	(0.0110)
HighBonus	0.0642***	0.0299**	0.0392***	0.0284**	0.0185**	0.00935	0.0108
	(0.0205)	(0.0118)	(0.0124)	(0.0118)	(0.00747)	(0.00779)	(0.0110)
Sanction	-0.0598***	-0.0316***	-0.0351***	-0.0334***	-0.0231***	-0.0435***	-0.0195**
	(0.0167)	(0.0107)	(0.00989)	(0.0107)	(0.00606)	(0.00584)	(0.00876)
Prevention	-0.0789***	-0.0236**	-0.0599***	-0.0252**	-0.0445***	-0.0135**	-0.0507***
	(0.0174)	(0.0105)	(0.00952)	(0.0104)	(0.00587)	(0.00601)	(0.00890)
AI-use block before tasks	-0.0394**	-0.0316***	-0.0388***	-0.0305***	-0.0201***	-0.000530	-0.0273***
	(0.0176)	(0.0109)	(0.0103)	(0.0108)	(0.00634)	(0.00573)	(0.00943)
Control mean	0.354	0.497	0.235	0.501	0.348	0.054	0.119
N	2955	2955	2955	2955	2955	2955	2955
R-squared	0.057	0.077	0.098	0.078	0.095	0.068	0.066

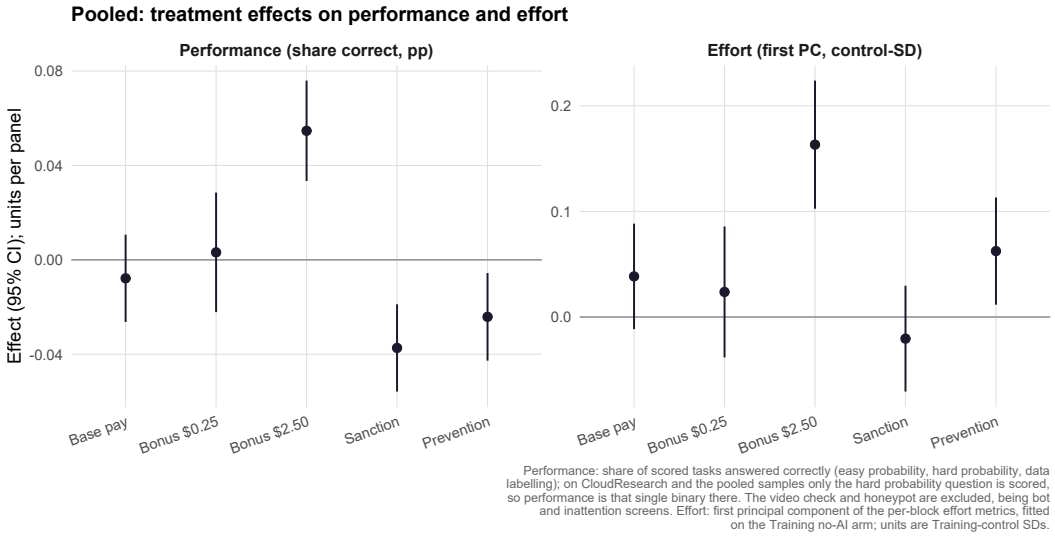
NOTES: The table reports treatment effects on each AI-use measure, adding the randomized placement of the AI-use question block to the four interventions. The unit of observation is a respondent who completed the main survey. The sample is Prolific only, because placement was randomized there alone. The placement coefficient is the effect of seeing the AI-use block before rather than after the tasks, which raises a respondent's estimate of the probability that AI use is being measured without changing incentives or the technical environment. All AI-use measures are on a common 0–1 scale: the PCA indices are percentile ranks, so their coefficients are rank points, while the pre-specified binary, the LASSO predicted probabilities, Pangram and self-report are probability or rate points. Controls, not displayed: age, gender, race, income, the four baseline AI attitudes, and wave fixed effects. Standard errors clustered at the listing level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

6.2 Effect on Effort and Performance

Next, we look at the effect of the interventions on effort and performance. As shown in Figure 2, performance moves closely with AI use in response to our interventions, while effort does not always move in the same direction as AI use. Under Prevention, when copy-paste is blocked, respondents increase their effort, typing more and faster. We see a similar increase in effort under the Bonus, as predicted by the model: respondents offered more for performance do work harder, as tracked by mouse and keyboard activity and time spent. The higher Bonus raises the return to performance, and respondents pursue that return through both more effort and more AI use.

Performance, on the other hand, moves in the same direction as AI use. Larger bonuses raise accuracy on the hard probability question by 0.022 points and the overall share correct by 0.013. The gain is concentrated on the question where a language model has the clearest advantage over a human respondent. Prevention and Sanction, which both reduce AI use, also reduce accuracy. Prevention lowers the overall share correct by 0.031 points, and Sanction lowers accuracy on the hard probability question by 0.037 points and the overall share correct by 0.013.

Figure 2: **The bonus raises effort and performance; prevention raises effort but lowers performance.**



NOTES: The figure reports treatment effects on the two summary indices, with 95% confidence intervals. Estimates are pooled across Prolific, CloudResearch and MTurk with platform fixed effects; robust standard errors. Performance is the fraction of scored tasks answered correctly, in percentage points; on CloudResearch and on the pooled sample the one commonly scored task, the hard probability question, is used, so performance is a binary there. Effort is the first principal component of the per-block effort metrics (keystrokes, backspace and delete keystrokes, inter-keystroke interval, time on page, scroll-direction switches, mouse jerk), fitted on the no-AI arm of the training sample and expressed in that arm's standard deviations. Mouse jerk is retained here, unlike in the AI-use indices, because the reference arm is a mixed-device field sample, so the device profile is part of the fitting data. Appendix Table B.4 reports the same estimates as numbers.

The Bonus effect is concentrated among men, who are also more responsive to Sanction (Appendix Table B.7). Respondents with higher income are more responsive to the Bonus (Appendix Table B.8), and we find little heterogeneity by age (Appendix Table B.9). The interventions also show little interaction with one another.

6.3 Selection, Attrition, and Recruitment Time

Base Pay does not move AI-assisted use, but it does change who takes the study and how quickly they take it. Because Base Pay is randomized at recruitment and posted in the listing a worker chooses to accept, its effect on anything measured after enrollment is a composition or take-up response rather than a behavioral one. Table 6 collects the five outcomes it moves, all on Prolific, the platform where base pay varies at the listing level and so gives the cleanest test. Higher Base Pay does not move the reservation wage workers report, but it draws workers who care more about their platform reputation and who carry substantially more prior approvals.

Appendix Table B.10 reports treatment effects on the two attention checks. Higher Base Pay raises attentiveness on both, either because it attracts more attentive workers or because it leads those who enroll to pay closer attention, and the design cannot separate the two.

Table 6 also shows that higher Base Pay fills a listing faster on both margins of speed: the posting draws its first respondent sooner, and the workers who do take it start sooner after it goes live.

Table 6: Effects of Base Pay on Selection and Take-Up, Prolific

	(1) Reservation Wage	(2) Reputation	(3) Approvals	(4) Take-up speed (log hrs)	(5) Time to First Response (hrs)
Base Pay: $15 v $8	0.0794	0.0375***	690.5***	-0.740***	-0.265***
	(0.236)	(0.0123)	(150.9)	(0.0652)	(0.0632)
Sample	Respondents	Respondents	Respondents	Workers who took up	Study cells
N	2954	2955	2955	3789	282
R-squared	0.001	0.004	0.016	0.108	0.095

NOTES: The table reports the effect of base pay on who selects into the study and on how fast the listing fills. Base pay is randomized at recruitment and posted in the listing, so every coefficient reflects selection and take-up rather than a behavioral response; this is the counterpart to the null base-pay effect on AI use in Figure 1. Every column is estimated on the binary base-pay arm, $15/hr against $8/hr, so the coefficients are directly comparable, and the sample is Prolific throughout, the platform where base pay is randomized at the listing level and shown in the listing a worker chooses to accept. The unit of observation differs across columns and is recorded in the Sample row. Columns (1)–(3) are measured on respondents who completed the main survey: the reservation wage in dollars, the importance of platform reputation on a three-point scale where 1 is not at all and 3 is very, and the worker's lifetime approved-submission count from the Prolific API. Approvals is a measure of platform tenure rather than a survey response and exists only on Prolific, with no analogue on MTurk or CloudResearch; it is the sharpest of the three composition outcomes, with higher base pay drawing workers with roughly 690 more prior approvals against a low-base-pay mean of about 1,810. Column (4) is the log of the hours between the posting going live and a worker's start, over every worker who took up, so the coefficient implies roughly a 52 percent reduction in time to take-up; it is the arm-difference form of the elasticity reported for all three platforms in Table 7, the same estimate rescaled by log(15/8). Column (5) is one observation per Prolific study cell and measures the hours from the study's publish time to its first started submission, both in UTC, against a low-base-pay mean of 0.44 hours; the other two platforms do not export a comparable publish timestamp. Wave fixed effects are included throughout, and no demographic controls, because those are themselves post-selection. Standard errors are clustered at the listing level in columns (1)–(3) and on the study cell in column (4), and are heteroskedasticity-robust in column (5). $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Platforms also differ in how strongly take-up responds to price, as Table 7 shows. Regressing take-up speed on log base pay, the elasticity is largest in magnitude on Prolific and smallest on MTurk.

Table 7: Take-Up Speed Against Log Base Pay

	(1) Prolific	(2) CloudResearch	(3) MTurk	(4) All
Log hourly base pay	-1.177***	-0.359**	-0.247*	-0.805***
	(0.104)	(0.0935)	(0.0940)	(0.0962)
Mean hours, extdollar 8/hr	6.98	21.04	79.02	23.00
Postings	282	6	4	292
N	3789	1732	1146	6667
R-squared	0.108	0.351	0.021	0.271

NOTES: The table reports how fast a posting fills as a function of the pay it advertises. The unit of observation is a worker who took the study, and the outcome is the log of the hours between the posting going live and that worker's start, so a negative coefficient means higher pay is taken up faster. Base pay is the posted rate the worker saw, $15/hr against $8/hr on every platform, so the coefficient is the arm difference scaled by log(15/8) and reads as an elasticity. Columns (1)–(3) are the three platforms separately and column (4) pools them. Wave fixed effects are included within platform, and platform-by-wave fixed effects in the pooled column. Posting-to-start is measured from the Prolific publish time, the MTurk HIT creation time and the CloudResearch study launch time, all converted to UTC before differencing, since Prolific reports UTC, MTurk Pacific and CloudResearch Europe/Berlin. Standard errors clustered on the posting in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

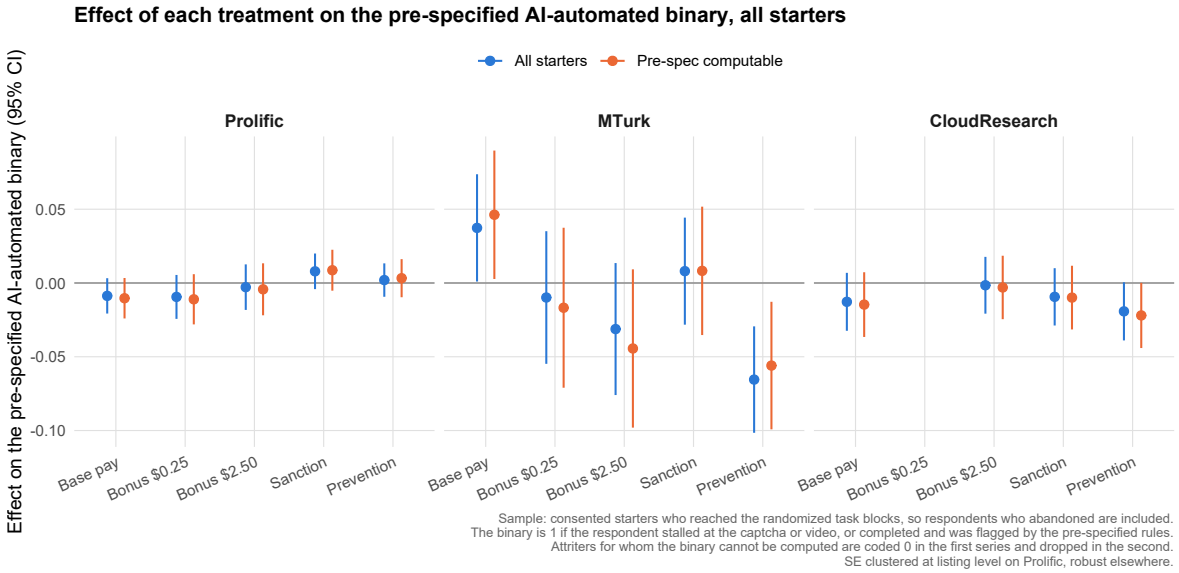
Participants are also somewhat more likely to complete the survey under higher Base Pay, as shown in Appendix Table B.11.

Appendix Table B.13 shows that the interventions had little effect on completion of the Prolific follow-up survey, with high Base Pay and the largest Bonus each retaining slightly fewer respondents.

6.4 Effect on AI-Automated Use

We now turn to the effect of the interventions on AI-automated responses, that is, surveys answered by bots. As Figure 3 and Appendix Table B.14 show, the treatments move automated responses mainly on MTurk, the platform with by far the highest baseline rate: about 20 percent of control-arm starters, against 7 percent on CloudResearch and 3 percent on Prolific. On MTurk, higher Base Pay raises automated responses and Prevention lowers them, plausibly because the interface makes it harder for an agent to complete the survey. On Prolific we find no detectable effects, and on CloudResearch only Prevention moves automated use, and only marginally.

Figure 3: **Automated responses respond to treatment mainly on MTurk, where higher base pay raises them and prevention lowers them.**



NOTES: The figure reports the effect of each treatment on the pre-specified AI-automated binary, with 95% confidence intervals, estimated separately by platform. The unit of observation is a consented starter who reached the randomized task blocks, so respondents who abandoned the survey are included; this is the sample the question requires, because an automated respondent that stalls at a screen never appears among completers. The binary is 1 if the respondent stopped at the captcha, stopped at the video, or completed the survey and was flagged by the pre-specified rule. The rule needs the three core typed tasks and so cannot be computed for abandoners who did not stall at a screen: the "All starters" series codes them 0 and the "Pre-spec computable" series drops them, and the two series bracket the answer. The three panels share a vertical scale, so effects are directly comparable across platforms even though the underlying rates are not: the control-arm rate of automated responses is about 3 percent on Prolific, 7 percent on CloudResearch and 20 percent on MTurk. Coefficients are linear-probability effects. Base pay is entry-stamped on Prolific, assigned by intention to treat on MTurk, and taken from the assignment lists on CloudResearch, which did not field the $0.25 tier. Standard errors are clustered at the listing level on Prolific and robust elsewhere. Appendix Table B.14 is the regression version of this figure.

6.5 Effects on Behavioral Evidence and Model Training

We next ask whether AI use is consequential for the data a researcher or firm ends up with. Using the procedure in Section 5.2, we reweight responses so that the reweighted mean approximates what the human respondents alone would have produced.

Table 8: Raw Against Human-Only Means

	Unweighted	Human-only	Difference	% change	N
Performance					
Probability (easy)	0.601	0.585	-0.016	-2.7	2955
		(0.009)			
Probability (hard)	0.129	0.102	-0.027	-21.3	2955
		(0.006)			
Data labelling	0.643	0.631	-0.012	-1.8	2955
		(0.010)			
Performance (share correct)	0.458	0.439	-0.018	-4.0	2955
		(0.005)			
Zero-sum beliefs (1–5)					
Zero-sum: ethnic	2.645	2.647	0.002	0.1	2955
		(0.023)			
Zero-sum: citizenship	2.575	2.571	-0.004	-0.1	2955
		(0.023)			
Anthropic labelling (0–100)					
Anthropic label 1	28.019	27.365	-0.654	-2.3	2955
		(0.521)			
Anthropic label 2	23.193	22.873	-0.320	-1.4	2955
		(0.521)			

NOTES: The table reports each outcome's raw mean against the mean the human respondents alone would have produced. The unit of observation is a respondent who completed the main survey, and the sample is Prolific only; the scalars quoted in the text below are the Prolific ones, which is why this rather than the pooled version carries the paragraph. The human-only column reweights every respondent by $w_i = 1 - p_i^{\text{field}}$, the CELSS-trained no-effort LASSO probability re-based to the estimated field AI rate by the procedure in Section 5.2; the full sample is retained and no respondent is dropped, so no threshold is imposed. Bootstrap standard errors in parentheses, from 500 replications resampling both the labeled and the field sample. Zero-sum items are coded 1 = strongly disagree to 5 = strongly agree. Anthropic labeling items are a 0–100 rating in which 0 is option A, the most honest and helpful option and the one chosen by a human labeler in the original training data, and 100 is option B.

Turning to performance, easy tasks are barely affected: the easy probability question moves 2.7% and data labeling 1.8%. Hard tasks move a great deal. Accuracy on the hard probability question falls from 0.166 to 0.128 once responses are discounted by their probability of being AI-assisted, a drop of 22.5%. A researcher fielding this instrument and reporting the raw number would overstate human performance on the hard question by about a quarter.

This has several consequences for researchers and firms that use online platforms for data labeling.

First, measured task performance is inflated selectively. Because the distortion is concentrated where the task is hard, AI use does not simply add noise to a performance measure; it compresses the variation researchers want to study by affecting the harder tasks disproportionately.

Second, AI prevalence affects inferences about human behavioral biases. The hard probability question is a ball-and-urn problem used to illustrate the base rate fallacy. With $P(\text{Jar A}) = 0.25$, $P(\text{green} \mid A) = 0.80$ and $P(\text{green} \mid B) = 0.20$, Bayes gives $P(A \mid \text{green}) = 0.57$, so a respondent answering 25 has anchored on the base rate and ignored the evidence, while one answering 80 or 20 has anchored on a likelihood and ignored the base rate. Reweighting to humans cuts the correct-answer rate from 16.6% to 12.8% and simultaneously raises base-rate anchoring from 17.1% to 18.7% (Appendix Table B.15). The AI-assisted responses err differently, updating on evidence where human respondents anchor, so a study of the base-rate fallacy run on contaminated data would understate the prevalence of the bias.

Attitude and opinion questions are a different matter. The treatments barely move the zero-sum thinking questions, and reweighting to humans barely moves their means, which is what we would expect if respondents rarely reach for a model on questions that have no right answer.

Finally, the labels that feed machine learning pipelines shift. The Anthropic labeling questions ask respondents to choose the more honest and helpful of two chatbot responses drawn from a training dataset, scored 0–100, with 0 the option chosen by a human labeler in the original data. Reweighting to humans moves the first question by 3.8% and the second by 0.5%, both toward the training-data answer, so AI-assisted respondents pull the aggregate label away from the answer human annotators would have given. Models trained without any adjustment for AI use would therefore be trained partly on model output relabeled as human judgment, which defeats the purpose of collecting it.

Overall, the moral hazard we document has consequences beyond the survey itself. It changes measured performance, it changes the inferred structure of human reasoning, and it changes the content of data used to train models.

Appendix Table B.16 summarizes the level of AI use across platforms, adjusted by the reweighting described in Section 3.7.

7 Conclusion

Online labor markets have become almost irreplaceable sources for human-generated data. To manage heterogeneity in work quality, researchers have developed a wide variety of practices and incentives to ensure high-quality work. The traditional tools for ensuring quality responses, namely reputation, screening, and incentives, have been widely used by both human behavior and computer science researchers. However, as AI use proliferates on these platforms, these tools may have to be rethought.

We find that incentives for high-quality responses can backfire in the presence of AI, as respondents are more likely to use AI when told that correct answers will be highly rewarded. However, the contaminating effects of incentives are not so large that they cannot be undone by relatively cheap interventions appealing to social preferences or raising the cost of using AI. This may speak to a larger role for organizations to vertically integrate their data-collection efforts in order to ensure compliance, via control of devices and organizational norms, with internal rules on the appropriate use of AI in producing data.

Our results show that AI-assisted responses (e.g., AI answers to questions fed by human respondents) are ubiquitous, but AI-automated responses (e.g., bots) are only common on Amazon Mechanical Turk, perhaps a reason why it has been decommissioned beginning at the end of July 2026. Nonetheless, as AI capabilities develop and become widely available, we expect that increased AI-automation on the other platforms will become more of a threat, and our methods may offer some guidance on how to screen for automated respondents. We also think that detection methods

will be probabilistic at best, and so reweighting the resulting data, as we have done, may be a solution that extracts the most "human-generated" signal from the data.

References

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. 2024. "Self-Consuming Generative Models Go MAD." In *Proceedings of the 12th International Conference on Learning Representations*. Vienna: ICLR. https://arxiv.org/abs/2307.01850.

Araujo, Douglas K. G., and Harald Uhlig. 2026. "How Does AI Distribute the Pie? Large Language Models and the Ultimatum Game." National Bureau of Economic Research Working Paper 34919. https://www.nber.org/papers/w34919.

Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E Glen Weyl. 2018. "Should we treat data as labor? Moving beyond "free"." Vol. 108, 38–42, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2026. "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely." *Review of Economic Studies*, 93(4): 2284–2312.

Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer. 2026. "How People Use Statistics." *Review of Economic Studies*, 93(1): 250–285.

Çelebi, Can, Christine Exley, Sören Harrs, Hannu Kivimäki, Marta Serra-Garcia, and Jeffrey Yusof. 2026. "Mission Possible: The Collection of High-Quality Data." CESifo Working Paper 12535. https://www.ifo.de/en/cesifo/publications/2026/working-paper/mission-possible-collection-high-quality-data.

Chinoy, Sahil, Nathan Nunn, Sandra Sequeira, and Stefanie Stantcheva. 2026. "Zero-Sum Thinking and the Roots of US Political Differences." *American Economic Review*, 116(3): 1052–1096.

Claassen, Joshua, Jan Karem Höhne, Ruben Bach, and Anna-Carolina Haensch. 2026. "Identifying Bots Through LLM-Generated Text in Open Narrative Responses: A Proof-of-Concept Study." *Social Science Computer Review*. Advance online publication. https://doi.org/10.1177/08944393251408022.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." *Econometrica*, 64(5): 1001–1044.

Dohmatob, Elvis, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. 2024. "A Tale of Tails: Model Collapse as a Change of Scaling Laws." In *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235 of *Proceedings of Machine Learning Research*, 11165–11197. PMLR. https://proceedings.mlr.press/v235/dohmatob24b.html.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. 2020. "Monopsony in Online Labor Markets." *American Economic Review: Insights*, 2(1): 33–46.

Einav, Liran, and Amy Finkelstein. 2023. "Empirical Analyses of Selection and Welfare in Insurance Markets: A Self-Indulgent Survey." *Geneva Risk and Insurance Review*, 48(2): 167–191.

Fernandez, Kianté A., Xiaozhi Yang, Stephanie M. Smith, and Ian Krajbich. 2026. "Defending Against AI Agents with Webcam Eye Tracking." PsyArXiv Preprint. https://osf.io/preprints/psyarxiv/ecrdw.

Gordon, Andrew, David M. Rothschild, Felipe M. Affonso, Justin Sulik, David J. Hauser, Karine Pepin, and Simon Jones. 2026. "AI Agent Prevalence and Data Quality Across Multiple Online Sample Providers." PsyArXiv Preprint. https://osf.io/preprints/psyarxiv/pvdjr.

Ho, Chien-Ju, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. "Incentivizing High Quality Crowdwork." In *Proceedings of the 24th International Conference on World Wide Web*. 419–429. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2736277.2741102.

Holmstrom, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica*, 55(2): 303–328.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7: 24–52.

Holtz, David M, Liane Scult, and Siddharth Suri. 2022. "How much do platform workers value reviews? An experimental method." 1–11.

Horton, John J., and Lydia B. Chilton. 2010. "The Labor Economics of Paid Crowdsourcing." In *Proceedings of the 11th ACM Conference on Electronic Commerce*. 209–218. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1807342.1807376.

Jabarian, Brian, and Alex Imas. 2025. "Artificial Writing and Automated Detection." National Bureau of Economic Research Working Paper 34223. https://www.nber.org/papers/w34223.

Karlan, Dean, and Jonathan Zinman. 2009. "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment." *Econometrica*, 77(6): 1993–2008.

Kay, Cameron S., and Madalina Vlasceanu. 2026. "A Scale for Detecting LLM-Generated Responses in Online Survey Research." PsyArXiv Preprint. https://osf.io/preprints/psyarxiv/4p7ns.

Liang, Chen, Jing Peng, Yili Hong, and Bin Gu. 2023. "The hidden costs and benefits of monitoring in the gig economy." *Information Systems Research*, 34(1): 297–318.

Ling, Yier, Alex Kale, and Alex Imas. 2026. "Underreporting of AI Use: The Role of Social Desirability Bias." In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–22. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3772318.3791073.

Mason, Winter, and Duncan J. Watts. 2009. "Financial Incentives and the 'Performance of Crowds'." *ACM SIGKDD Explorations Newsletter*, 11(2): 100–108.

Pallais, Amanda. 2014. "Inefficient hiring in entry-level labor markets." *American Economic Review*, 104(11): 3565–3599.

Panizza, Folco, Yara Kyrychenko, and Jon Roozenbeek. 2026. "Survey-Taking AI Tools Surpass Human Abilities. Here's What We Can Do About It." *Nature*, 650(8101): 293–295.

Pate, Neeley, Adiba Mahbub Proma, Hangfeng He, James N. Druckman, Daniel C. Molden, Gourab Ghoshal, and Ehsan Hoque. 2026. "Replicating Human Motivated Reasoning Studies with LLMs." arXiv Preprint 2601.16130. https://arxiv.org/abs/2601.16130.

Rilla, Raluca, Tobias Werner, Hiromu Yakura, Iyad Rahwan, and Anne-Marie Nussberger. 2026. "Recognising and Mitigating LLM Pollution in Online Behavioural Research." *Nature Communications*, 17(1): 5578. https://doi.org/10.1038/s41467-026-74621-9.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. "AI Models Collapse When Trained on Recursively Generated Data." *Nature*, 631(8022): 755–759.

Van der Stigchel, Stefan, Christoph Strauch, Beleke de Zwart, and Leendert Van Maanen. 2026. "Will Online Behavioral Research Follow the Fate of Online Survey Research?" *Proceedings of the National Academy of Sciences*, 123(8): e2535585123.

Veldkamp, Laura, and Cindy Chung. 2024. "Data and the aggregate economy." *Journal of Economic Literature*, 62(2): 458–484.

Velutharambath, Aswathy, Neele Falk, Sofie Labat, Tarun Tater, and Amelie Wührl. 2026. "Can Crowdsourcing Survive the LLM Era? A Community Survey on Human Data Collection." arXiv Preprint 2606.04924. https://arxiv.org/abs/2606.04924.

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West. 2023. "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks." arXiv Preprint 2306.07899. https://arxiv.org/abs/2306.07899.

Veselovsky, Veniamin, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David M. Rothschild, and Robert West. 2025. "Prevalence and Prevention of Large Language Model Use in Crowd Work." *Communications of the ACM*, 68(3): 42–47. https://doi.org/10.1145/3685527.

Westwood, Sean J. 2025. "The Potential Existential Threat of Large Language Models to Online Survey Research." *Proceedings of the National Academy of Sciences*, 122(47): e2518075122.

Xing, Xiaodan, Fadong Shi, Jiahao Huang, Yinzhe Wu, Yang Nan, Sheng Zhang, Yingying Fang, Michael Roberts, Carola-Bibiane Schönlieb, Javier Del Ser, and Guang Yang. 2025. "On the Caveats of AI Autophagy." *Nature Machine Intelligence*, 7(2): 172–180.

Zhang, Grace, Robert Walatka, Stephanie Y. Chen, Oleg Urminsky, Kianté Fernandez, Andrea Low, Jonathan E. Bogard, and Craig R. Fox. 2026. "Estimating the Threat of AI-Agent Responding Across Online Survey Platforms." PsyArXiv Preprint. https://osf.io/preprints/psyarxiv/xcg26.

Zhang, Simone, Janet Xu, and AJ Alvero. 2025. "Generative AI Meets Open-Ended Survey Responses: Research Participant Use of AI and Homogenization." *Sociological Methods & Research*, 54(3): 1197–1242.

Appendices

Appendix A AI Detection

Supplementary material for Section 3.

Detection Variables

Table A.1 gives the variables used to detect AI-assisted responses and Table A.2 those used to detect AI-automated responses. The two sets overlap but are not identical: the AI-automated set drops the page-engagement variables that record browsing to another tab, since an automated agent has no reason to navigate one, and adds respondent-level screens that are hard for an agent to pass. In both tables the "Tasks" column records where a variable can be collected — "All" for any question, "Text" for free-response questions, and "Prompt" for questions whose text a respondent could copy, which excludes prompts delivered as an image or a video. Every variable is oriented so that a higher value indicates a higher probability of an AI response.

Table A.1: Detection Variables: AI-Assisted (Human Using LLM) Responses

Category	Variable	Description	Tasks
Clipboard Activity	Copy event	Indicator for any copy event	All
	Paste event	Indicator for any paste event	Text
	Copied text matches prompt	Indicator for copied text matching the question text or passage	Prompt
Keystroke Behavior	Empty keystroke log	Indicator for no keystroke activity recorded	Text
	Input jump count	Number of discrete input length increases exceeding a single keystroke, indicative of paste events	Text
	Keystroke event count	Total number of keystroke events recorded	Text
	Keylog-reconstructed similarity	One minus the Levenshtein distance between the keylog-reconstructed and final response, normalized by maximum text length	Text
	Median inter-keystroke interval	Median time between consecutive keystrokes (ms)	Text
	SD inter-keystroke interval	Standard deviation of inter-keystroke interval (ms)	Text
	Typing speed indicator	Indicator for median inter-keystroke interval below 75 ms	Text

Continued on next page

Table A.1 continued from previous page

Category	Variable	Description	Tasks
	Input-jump indicator	Indicator for a paste keystroke in the log or a single input jump of at least fifty characters	Text
Page Engagement	Page leaves	Number of times the respondent left the survey page or switched browser tabs	All
	Left page	Indicator for at least one page leave	All
	Window blurring	Indicator for the survey window losing focus	All
	No focus loss	Indicator for the page never being hidden or losing focus	All
	Time off page	Seconds the page was hidden or unfocused	All
	Focused time	Page time less time off page	All
	Off-page share	Time off page divided by total time on page	All
	Fast completion	Indicator for completion time in the bottom fifth percentile of the sample distribution	All
	Slow completion	Indicator for completion time in the top ninety-fifth percentile of the sample distribution	All
Cursor and Scroll Behavior	Cursor moved	Indicator for any cursor movement on the page	All
	Move count	Number of discrete cursor movement events	All
	Point count	Number of recorded cursor coordinates	All
	Stop count	Number of pauses in cursor movement	All
	Time to first move	Milliseconds from page load to the first cursor movement	All
	Velocity dispersion	Standard deviation of cursor velocity	All
	Angle dispersion	Standard deviation of movement direction; low values indicate a straight cursor path	All
	Jerk dispersion	Standard deviation of the third derivative of cursor position; low values indicate unnaturally smooth motion	All
	Reversal count	Number of direction reversals in the cursor path	All

Continued on next page

Table A.1 continued from previous page

Category	Variable	Description	Tasks
	Reversal rate	Direction reversals per one thousand recorded cursor points	All
	Scroll count	Number of scroll events	All
	Scroll direction changes	Number of scroll direction reversals	All
	Scroll switch rate	Scroll direction changes divided by scroll events	All
AI Text Detection	Pangram	Categorical AI-text indicator: 0 (fully human), 0.5 (partially AI-assisted), 1 (fully AI-generated)	Text

Notes: This table summarizes the variables used to detect AI-assisted responses. Variables at the task or page level can be recorded separately for each relevant task or page. "All" denotes variables that can be measured for all questions. "Text" denotes variables that can be measured for text-based responses. "Prompt" denotes questions that display text a respondent could copy, which excludes questions whose prompt appears in an image or video.

Table A.2: Detection Variables: AI-Automated (Bot) Responses

Category	Variable	Description	Tasks
Clipboard Activity	Copy event	Indicator for any copy event	All
	Paste event	Indicator for any paste event	Text
	Copied text matches prompt	Indicator for copied text matching the question text or passage	Prompt
Keystroke Behavior	Empty keystroke log	Indicator for no keystroke activity recorded	Text
	Input jump count	Number of discrete input length increases exceeding a single keystroke, indicative of paste events	Text
	Keystroke event count	Total number of keystroke events recorded	Text
	Keylog-reconstructed similarity	One minus the Levenshtein distance between the keylog-reconstructed and final response, normalized by maximum text length	Text
	Median inter-keystroke interval	Median time between consecutive keystrokes (ms)	Text
	SD inter-keystroke interval	Standard deviation of inter-keystroke interval (ms)	Text
	Typing speed indicator	Indicator for median inter-keystroke interval below 75 ms	Text

Continued on next page

Table A.2 continued from previous page

Category	Variable	Description	Tasks
	Input-jump indicator	Indicator for a paste keystroke in the log or a single input jump of at least fifty characters	Text
Page Engagement	No focus loss	Indicator for the page never being hidden or losing focus	All
	Time off page	Seconds the page was hidden or unfocused	All
	Focused time	Page time less time off page	All
	Off-page share	Time off page divided by total time on page	All
	Fast completion	Indicator for completion time in the bottom fifth percentile of the sample distribution	All
	Slow completion	Indicator for completion time in the top ninety-fifth percentile of the sample distribution	All
Cursor and Scroll Behavior	Cursor moved	Indicator for any cursor movement on the page	All
	Move count	Number of discrete cursor movement events	All
	Point count	Number of recorded cursor coordinates	All
	Stop count	Number of pauses in cursor movement	All
	Time to first move	Milliseconds from page load to the first cursor movement	All
	Velocity dispersion	Standard deviation of cursor velocity	All
	Angle dispersion	Standard deviation of movement direction; low values indicate a straight cursor path	All
	Jerk dispersion	Standard deviation of the third derivative of cursor position; low values indicate unnaturally smooth motion	All
	Reversal count	Number of direction reversals in the cursor path	All
	Reversal rate	Direction reversals per one thousand recorded cursor points	All
	Scroll count	Number of scroll events	All
	Scroll direction changes	Number of scroll direction reversals	All

Continued on next page

Table A.2 continued from previous page

Category	Variable	Description	Tasks
	Scroll switch rate	Scroll direction changes divided by scroll events	All
AI Text Detection	Pangram	Categorical AI-text indicator: 0 (fully human), 0.5 (partially AI-assisted), 1 (fully AI-generated)	Text

Notes: This table summarizes the variables used to detect AI-automated responses. Variables at the task or page level can be recorded separately for each relevant task or page. "All" denotes variables that can be measured for all questions. "Text" denotes variables that can be measured for text-based responses. "Prompt" denotes questions that display text a respondent could copy, which excludes questions whose prompt appears in an image or video.

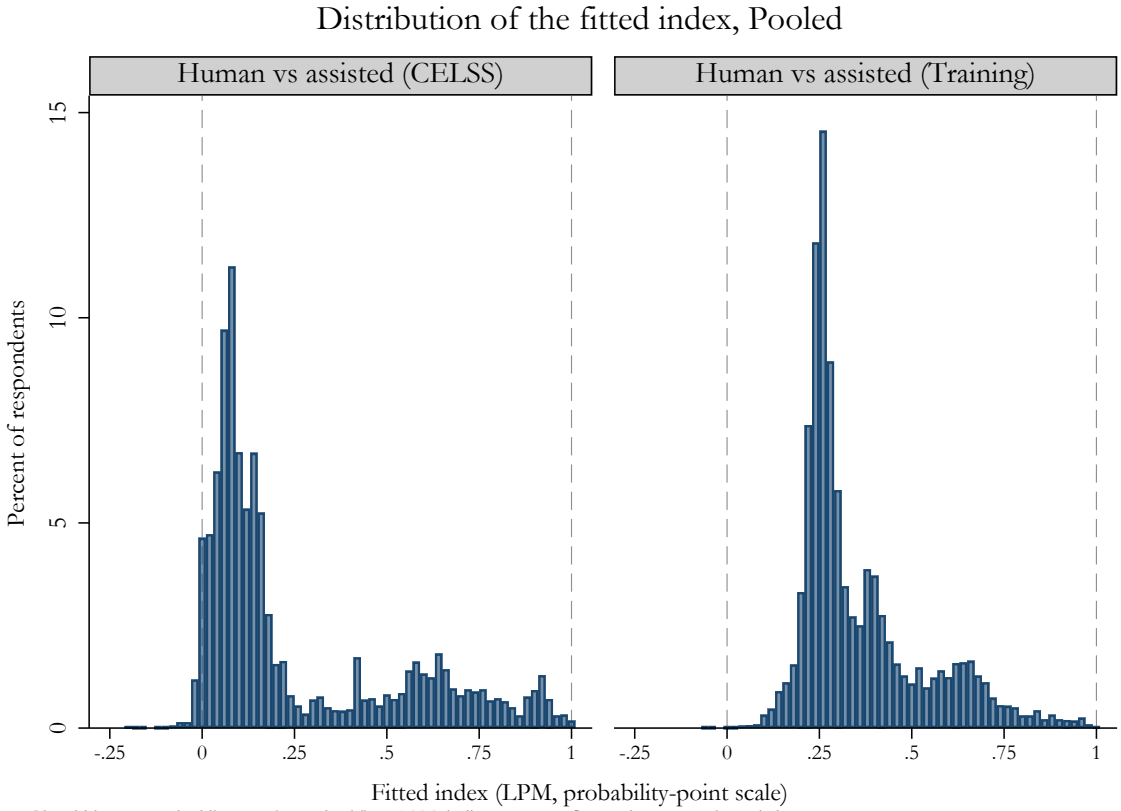
Table A.3: AI-Use Measures Against Pangram's AI-Text Flag

	Pangram: any AI			Pangram: fully AI		
	Not flagged	Flagged	Diff	Not flagged	Flagged	Diff
Prespec	0.274	1.000	0.726***	0.294	1.000	0.706***
PCA-C	0.414	0.691	0.276***	0.422	0.694	0.272***
LASSO-C	0.192	0.532	0.340***	0.201	0.535	0.334***
PCA-T	0.409	0.694	0.285***	0.416	0.697	0.281***
LASSO-T	0.312	0.600	0.288***	0.318	0.609	0.291***
Self	0.052	0.277	0.225***	0.056	0.287	0.231***
Attn: video	0.856	0.840	-0.016	0.854	0.845	-0.009
Attn: IMC	0.575	0.253	-0.323***	0.569	0.238	-0.331***
N (respondents)	4582	998		4713	867	

NOTES: The table reports the mean of each AI-use measure separately for respondents the Pangram AI-text detector does and does not flag. The unit of observation is a respondent who completed the main survey, pooled across Prolific, CloudResearch and MTurk, so unlike Table 3 this benchmark is available on every platform. The two column blocks are two thresholds on the Pangram score, which is 0 for fully human, 0.5 for partially AI-assisted and 1 for fully AI-generated: "any AI" flags a respondent scoring 0.5 or 1, and "fully AI" only those scoring 1, which is why fewer respondents are flagged in the second block. A respondent's score is the maximum across their scored open-response items, so any one flagged answer flags the respondent. Each block partitions the full pooled sample, so no respondent is dropped. "Diff" is the difference in means between flagged and unflagged respondents, and the stars are from a two-sample t-test on that difference. Pangram itself has no row, because it cannot be split by itself. Attention checks are coded 1 = passed, so their negative differences mean flagged respondents were less attentive.

Two rows should not be read as validation. Pangram enters the feature set from which the PCA and LASSO indices are fitted — in the Training no-copy-paste LASSO it carries the largest coefficient of any feature — so those rows compare an aggregate against one of its own heaviest components, a consistency check rather than independent corroboration. The pre-specified rule counts a Pangram flag as one of its sufficient conditions, so it equals 1.000 among flagged respondents by construction and carries no information; it is kept only so the row set matches Table 3. Self-report and the two attention checks are the only rows genuinely independent of Pangram. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Figure A.1: **The fitted AI-use index is spread across the whole range, with no natural cutoff separating human from AI-assisted responses.**



NOTES: The figure plots the distribution of the LASSO predicted probability of AI assistance across field respondents, pooled over Prolific, CloudResearch and MTurk. The unit of observation is a respondent who completed the main survey. Each panel is one index, labeled by the contrast in the labeled sample it was trained on: human versus AI-assisted in CELSS on the left, and the same contrast in the Training sample on the right. Both panels share an axis, with bin width 0.02 and dashed lines at 0 and 1. The indices are linear-probability fits, so they are not bounded to $[0, 1]$; 1.0 percent of fitted values fall outside the window shown and are omitted from the plot only, not from any estimate. The spread of the index is what motivates reweighting rather than classifying in Section 5.2: a cutoff would fall where the distribution is thin, so the classified set would move a great deal for a small change in it. On the pooled sample `index_lasso` is the common-slot variant by construction, since not every block is fielded on every platform, so the panels are not a simple average of what each platform contributes.

Table A.4: Base-Rate Correction Diagnostics

Quantity	Value
Mean score among AI answers, lab	0.8580
Mean score among human answers, lab	0.1338
Detector separation	0.7242
Lab AI rate	0.5000
Mean score in the field	0.2123
Field AI rate, BBSE, full sample (used for the weight)	0.1083
	(0.0178)
Pure control arm only (reported prevalence)	0.1136
N in that cell	131
Mean re-based posterior	0.0771
Gap against the field AI rate	0.0312
Field N	2955
Lab N	414
Bootstrap replications	500

NOTES: The table reports the three quantities the base-rate correction of Equations (3) and (4) rests on, estimated on the Prolific sample with CELSS as the labeled sample. Separation is $\mu_1 - \mu_0$, the gap in mean detector score between known AI and known human answers, and is the denominator of the correction, so a small value would make $\hat{\pi}_{\text{field}}$ fragile. $\hat{\pi}_{\text{field}}$ is the implied field AI rate; its bootstrap standard error is in parentheses and carries the sampling error in the laboratory components as well as in the field mean.

Appendix B Results

Table B.1: Treatment Effects on AI Use

	(1) Pre-spec	(2) PCA (CELSS)	(3) LASSO (CELSS)	(4) PCA (Training)	(5) LASSO (Training)	(6) Self-Report	(7) Pangram
Base Pay	0.00257	-0.00163	0.00950	-0.00431	-0.00132	-0.00782	-0.00371
	(0.0118)	(0.00748)	(0.00931)	(0.00735)	(0.00425)	(0.00683)	(0.00801)
Bonus Amount	0.0199***	0.0115***	0.00838**	0.0112***	0.00412**	0.00391	0.00401
	(0.00512)	(0.00326)	(0.00404)	(0.00319)	(0.00187)	(0.00295)	(0.00343)
Sanction	-0.0576***	-0.0378***	-0.0465***	-0.0377***	-0.0228***	-0.0263***	-0.0206**
	(0.0118)	(0.00750)	(0.00901)	(0.00736)	(0.00426)	(0.00680)	(0.00802)
Prevention	-0.0726***	-0.0450***	-0.0445***	-0.0451***	-0.0477***	-0.0124*	-0.0824***
	(0.0119)	(0.00747)	(0.00928)	(0.00734)	(0.00428)	(0.00682)	(0.00805)
Control mean	0.418	0.481	0.290	0.475	0.390	0.111	0.217
N	5568	5568	5568	5568	5568	5568	5568
R-squared	0.200	0.103	0.081	0.116	0.201	0.220	0.337

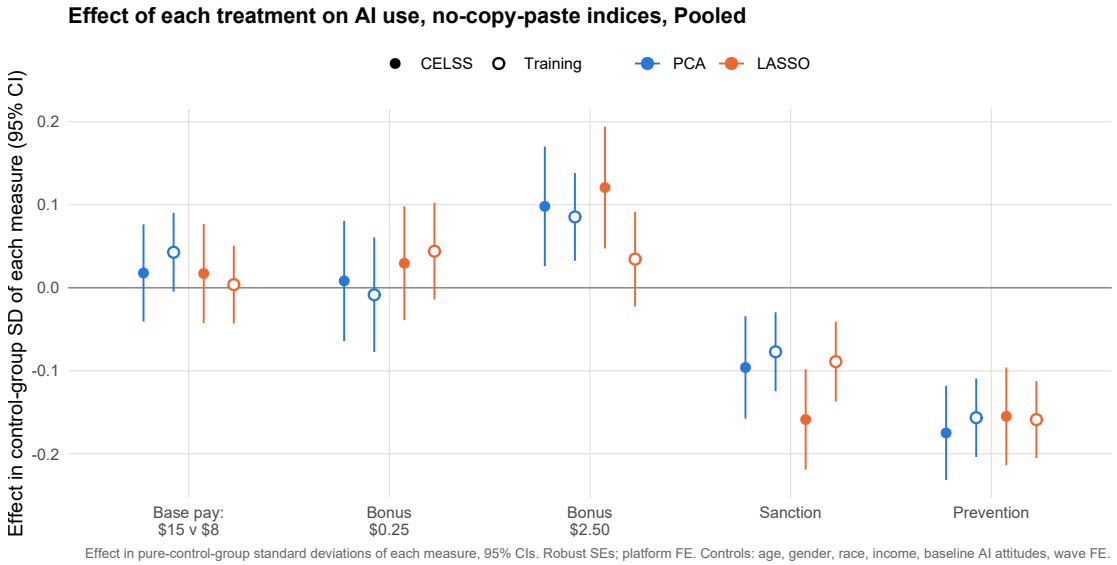
NOTES: The table reports treatment effects on each AI-assistance measure. The unit of observation is a respondent who completed the main survey. Each column is a separate regression on one AI-use measure: column (1) is the pre-specified binary index, columns (2) to (5) the four data-driven indices, and columns (6) and (7) self-report and Pangram. Base pay is \$15/hr against \$8/hr and enters as an indicator, and the bonus enters linearly in dollars, so its coefficient is the effect per dollar offered. Estimates are pooled across Prolific, CloudResearch, and MTurk with platform fixed effects, and the indices are constructed from questions common to all platforms. Controls, not displayed, are age, gender, race, income, and the four baseline AI attitudes, with wave fixed effects. The control mean is the mean of the outcome in the pure control cell. Robust standard errors in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table B.2: Treatment Effects on AI Use, Saturated in Bonus

	(1) Pre-spec	(2) PCA (CELSS)	(3) LASSO (CELSS)	(4) PCA (Training)	(5) LASSO (Training)	(6) Self-Report	(7) Pangram
Base Pay	0.00261	-0.00155	0.00952	-0.00425	-0.00128	-0.00784	-0.00373
	(0.0118)	(0.00748)	(0.00931)	(0.00735)	(0.00425)	(0.00682)	(0.00801)
LowBonus	0.0139	0.0199**	0.00628	0.0169*	0.00826	-0.00465	-0.00385
	(0.0163)	(0.0101)	(0.0109)	(0.0101)	(0.00573)	(0.00995)	(0.0115)
HighBonus	0.0524***	0.0339***	0.0222**	0.0323***	0.0125**	0.00805	0.00853
	(0.0136)	(0.00868)	(0.0110)	(0.00846)	(0.00495)	(0.00765)	(0.00898)
Sanction	-0.0576***	-0.0379***	-0.0466***	-0.0377***	-0.0228***	-0.0263***	-0.0206**
	(0.0118)	(0.00750)	(0.00900)	(0.00736)	(0.00426)	(0.00680)	(0.00802)
Prevention	-0.0726***	-0.0450***	-0.0445***	-0.0451***	-0.0477***	-0.0124*	-0.0824***
	(0.0119)	(0.00746)	(0.00928)	(0.00734)	(0.00428)	(0.00682)	(0.00805)
Control mean	0.418	0.481	0.290	0.475	0.390	0.111	0.217
N	5568	5568	5568	5568	5568	5568	5568
R-squared	0.200	0.103	0.081	0.116	0.202	0.220	0.337

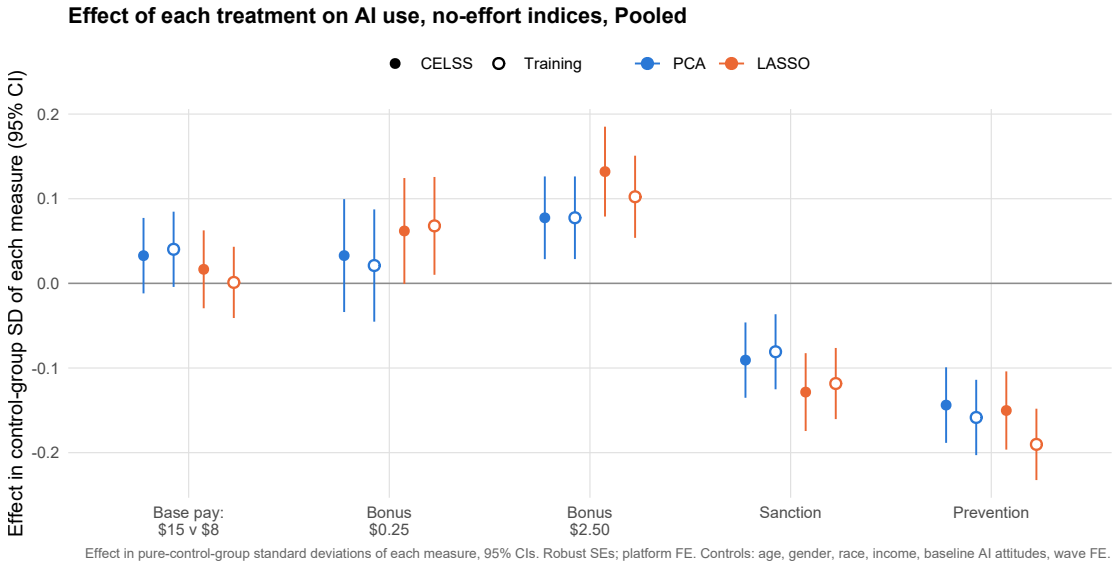
NOTES: The table reports treatment effects on each AI-assistance measure, with the bonus entered as two indicators, Low and High Bonus, each against the No Bonus arm. The unit of observation is a respondent who completed the main survey. Each column is a separate regression on one AI-use measure: column (1) is the pre-specified binary index, columns (2) to (5) the four data-driven indices, and columns (6) and (7) self-report and Pangram. Base pay is \$15/hr against \$8/hr and enters as an indicator. Estimates are pooled across Prolific, CloudResearch, and MTurk with platform fixed effects, and the indices are constructed from questions common to all platforms. Controls, not displayed, are age, gender, race, income, and the four baseline AI attitudes, with wave fixed effects. The control mean is the mean of the outcome in the pure control cell. Robust standard errors in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Figure B.1: **Prevention still lowers AI use when the copy-paste features are removed from the index.**



NOTES: The figure repeats Figure 1 using the no-copy-paste variants of the four model indices, which are refitted after dropping the clipboard features, so the only remaining information is typing rhythm, page engagement and cursor behavior. This is the variant the prevention result has to survive, since prevention blocks copy-paste and would mechanically lower any index built on it. Effects are in standard deviations of the pure control group's distribution of each measure, with 95% confidence intervals, on the same scale as Figure 1. Estimates are pooled across Prolific, CloudResearch and MTurk with platform fixed effects, so every contrast is estimated within platform. Marker color denotes the index method (PCA or LASSO) and marker fill the labeled sample the index was trained on (CELSS or Training). There is no pre-specified series, because the pre-specified rule is a fixed binary and has no no-copy-paste variant. Controls: age, gender, race, income, the four baseline AI attitudes, and wave fixed effects. Robust standard errors.

Figure B.2: **The treatment effects are not an effort artifact: they survive dropping every effort feature from the index.**



NOTES: The figure repeats Figure 1 using the no-effort variants of the four model indices, which are refitted after dropping keystroke counts, response timing, time on page and mouse activity. Because those features also move with how hard a respondent is working, an index containing them could in principle register a change in effort as a change in AI use; this variant removes that possibility. Scale, sample, encoding, controls and standard errors are as in Figure B.1. There is again no pre-specified series.

Table B.3: Treatment Effects on Individual Detection Variables

	(1) Left page	(2) Copy	(3) Char-edit	(4) Paste	(5) Empty keylog	(6) Input jump
Base Pay	0.0298**	0.00827	-0.00209	-0.00173	-0.00461	-0.0117
	(0.0125)	(0.0121)	(0.00919)	(0.0112)	(0.00283)	(0.0102)
Bonus Amount	0.0161***	0.0122**	0.00398	0.0120**	-0.000473	0.00887**
	(0.00540)	(0.00525)	(0.00396)	(0.00486)	(0.00119)	(0.00445)
Sanction	-0.0541***	-0.0382***	-0.0164*	-0.0257**	0.00396	-0.0279***
	(0.0125)	(0.0122)	(0.00918)	(0.0113)	(0.00281)	(0.0102)
Prevention	0.00975	-0.0502***	-0.196***	-0.0866***	0.00569**	-0.108***
	(0.0125)	(0.0122)	(0.00908)	(0.0113)	(0.00287)	(0.0103)
Control mean	0.650	0.324	0.268	0.334	0.010	0.275
N	5568	5568	5568	5568	5568	5568
R-squared	0.075	0.063	0.169	0.155	0.065	0.174

NOTES: The table reports treatment effects on each detection marker separately, so that the index results of Figure 1 can be read against the individual behaviors they aggregate. The unit of observation is a respondent who completed the main survey, pooled across Prolific, CloudResearch and MTurk with platform fixed effects. Each column is a separate regression. Markers are recorded across all task blocks and coded so that the value shown is the outcome for any question; "Char-edit" counts character-level edit events and "Self-report" is the respondent's own statement that they used AI. Table A.1 defines each marker. Coefficients are in each outcome's own units rather than z-scored, and the binary outcomes are linear-probability effects. Controls, not displayed: age, gender, race, income, the four baseline AI attitudes, and wave fixed effects. Robust standard errors in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table B.4: Treatment Effects on the Effort and Performance Indices

	(1) Performance	(2) Effort
Base Pay	-0.00782	0.0384
	(0.00944)	(0.0255)
Bonus Amount	0.0222***	0.0643***
	(0.00415)	(0.0117)
Sanction	-0.0373***	-0.0204
	(0.00942)	(0.0256)
Prevention	-0.0242**	0.0624**
	(0.00947)	(0.0259)
Control mean	0.176	-0.471
N	5568	5568
R-squared	0.116	0.298

NOTES: The table reports the estimates plotted in Figure 2. The unit of observation is a respondent who completed the main survey, pooled across Prolific, CloudResearch and MTurk with platform fixed effects. Column (1) is performance and column (2) is effort, both defined in the note to Figure 2. The control mean is the mean of the outcome in the arm receiving none of the four interventions. Performance coefficients are percentage points; an effort coefficient of 1 is one standard deviation of the no-AI arm of the training sample. Controls, not displayed: age, gender, race, income, the four baseline AI attitudes, and wave fixed effects. Robust standard errors in parentheses. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.5: Treatment Effects on the Effort Markers

	(1) Keystrokes	(2) Edits	(3) Median IKI (ms)	(4) Time (sec)	(5) Scroll switches	(6) Mouse jerk
Base Pay	36.96*	52.23	9.469	10.90	-1.520	-5.11e-08
	(22.07)	(37.69)	(48.78)	(22.88)	(1.446)	(0.000000614)
Bonus Amount	36.13***	70.38***	-1.615	42.38***	1.422**	-0.000000397
	(9.682)	(16.80)	(19.69)	(11.08)	(0.682)	(0.000000307)
Sanction	6.539	8.495	15.31	-54.73**	-0.822	-0.000000578
	(20.59)	(34.78)	(47.70)	(23.53)	(1.496)	(0.000000637)
Prevention	78.66***	150.7***	-133.6***	36.76	1.708	-0.000000276
	(22.64)	(37.90)	(46.34)	(23.53)	(1.455)	(0.000000710)
Control mean	933.062	1676.206	479.688	1169.669	51.934	0.000
N	2955	2955	2950	2955	2955	2955
R-squared	0.042	0.051	0.013	0.084	0.038	0.007

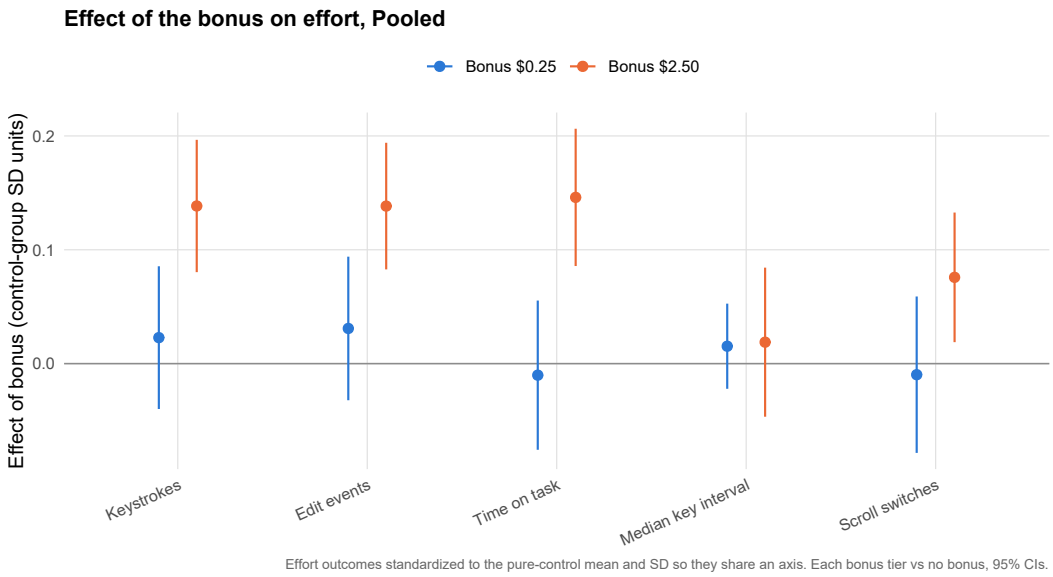
NOTES: The table reports treatment effects on each effort marker separately, so that the summary index of Figure 2 can be read against its components. The unit of observation is a respondent who completed the main survey. The sample is Prolific only, because no pooled version of this table is built. Each outcome is aggregated across the task blocks and left in its own units rather than standardized, so keystroke counts are counts and the inter-keystroke interval is in milliseconds; binary outcomes are linear-probability effects. Controls, not displayed: age, gender, race, income, the four baseline AI attitudes, and wave fixed effects. Standard errors clustered at the listing level in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table B.6: Treatment Effects on Task Performance

	(1)	(2)	(3)	(4)
	Prob. (hard)	Video	Share correct	Any honeypot
Base Pay	-0.00782	0.0261***	0.0185***	0.000404
	(0.00944)	(0.00937)	(0.00685)	(0.00684)
Bonus Amount	0.0222***	0.00612	0.0133***	0.000351
	(0.00415)	(0.00411)	(0.00299)	(0.00291)
Sanction	-0.0373***	-0.000703	-0.0133*	-0.00197
	(0.00942)	(0.00937)	(0.00685)	(0.00680)
Prevention	-0.0242**	-0.0122	-0.0305***	-0.0148**
	(0.00947)	(0.00948)	(0.00690)	(0.00678)
Control mean	0.176	0.836	0.526	0.090
N	5568	5568	5568	5568
R-squared	0.116	0.033	0.075	0.183

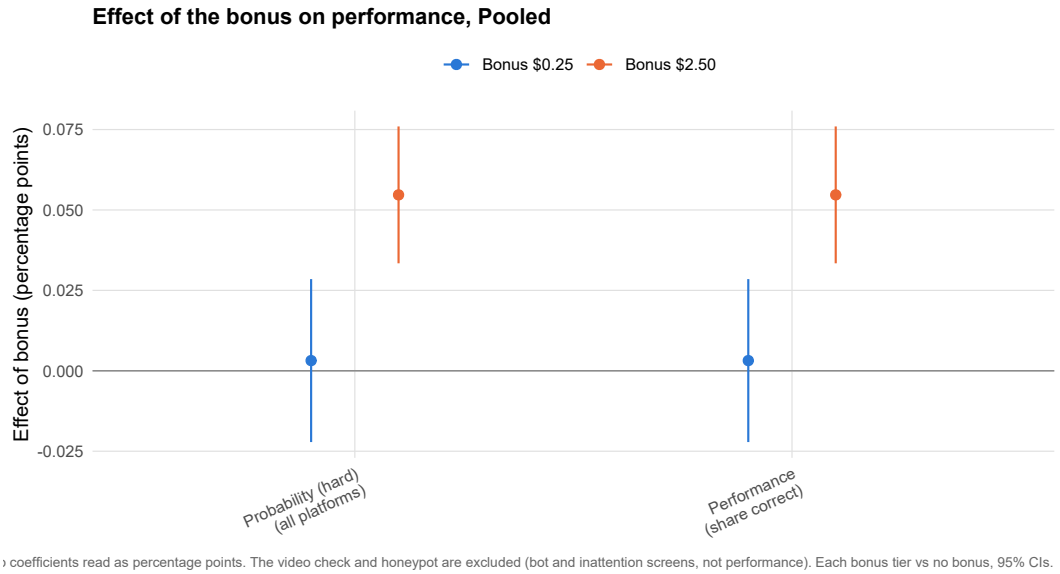
NOTES: The table reports treatment effects on each scored task separately. The unit of observation is a respondent who completed the main survey, pooled across Prolific, CloudResearch and MTurk with platform fixed effects. Each column is a separate regression and each outcome is an indicator for a correct answer, so coefficients read as percentage points. "Prob. (easy)" and "Prob. (hard)" are the two probability-reasoning questions, "Labelling" is the entity-extraction task, and "Video" is the video comprehension check. Controls, not displayed: age, gender, race, income, the four baseline AI attitudes, and wave fixed effects. Robust standard errors in parentheses. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Figure B.3: **Higher bonuses raise effort on every component.**



NOTES: The figure reports the effect of each bonus tier on each effort component, relative to the no-bonus arm, with 95% confidence intervals. Effort outcomes are standardized to mean 0 and standard deviation 1, so the axis is in standard deviation units. Estimates are pooled across Prolific, CloudResearch and MTurk with platform fixed effects, so every contrast is estimated within platform; robust standard errors. Figure 2 reports the effect on the summary index that aggregates these components.

Figure B.4: **Higher bonuses raise performance, and the gain is concentrated on the hard tasks.**



NOTES: The figure reports the effect of each bonus tier on each scored task, relative to the no-bonus arm, with 95% confidence intervals. Outcomes are indicators for a correct answer, so coefficients read as percentage points. Estimates are pooled across Prolific, CloudResearch and MTurk with platform fixed effects, so every contrast is estimated within platform; robust standard errors. The hard probability question is the ball-and-urn problem described in Section 6.5, on which a language model has the clearest advantage over an unaided respondent.

Table B.7: Heterogeneity by Gender

	Pre-spec	PCA (CELSS)	LASSO (CELSS)	PCA (Training)	LASSO (Training)	Self-Report	Pangram
Panel A: Women / non-binary							
Base Pay	0.000567	-0.00604	0.00653	-0.00757	-0.00615	-0.00312	0.00385
	(0.0167)	(0.0105)	(0.0131)	(0.0103)	(0.00577)	(0.00977)	(0.0116)
Bonus Amount	0.00719	0.00573	0.00333	0.00484	0.00110	-0.00148	0.00302
	(0.00724)	(0.00459)	(0.00543)	(0.00451)	(0.00253)	(0.00428)	(0.00497)
Sanction	-0.0585***	-0.0215**	-0.0310**	-0.0225**	-0.0133**	-0.0245**	-0.00574
	(0.0167)	(0.0105)	(0.0128)	(0.0104)	(0.00579)	(0.00976)	(0.0115)
Prevention	-0.0730***	-0.0349***	-0.0367***	-0.0374***	-0.0451***	-0.0204**	-0.0754***
	(0.0168)	(0.0105)	(0.0127)	(0.0103)	(0.00583)	(0.00969)	(0.0116)
Control mean	0.437	0.463	0.278	0.457	0.387	0.111	0.234
N	2772	2772	2772	2772	2772	2772	2772
R-squared	0.219	0.118	0.105	0.131	0.268	0.183	0.377
Panel B: Men							
Base Pay	0.00386	-0.000367	0.00922	-0.00406	0.00227	-0.0103	-0.0122
	(0.0168)	(0.0106)	(0.0134)	(0.0104)	(0.00623)	(0.00943)	(0.0110)
Bonus Amount	0.0302***	0.0165***	0.0129**	0.0166***	0.00685**	0.00961**	0.00380
	(0.00724)	(0.00457)	(0.00569)	(0.00447)	(0.00274)	(0.00406)	(0.00467)
Sanction	-0.0581***	-0.0553***	-0.0634***	-0.0538***	-0.0328***	-0.0278***	-0.0345***
	(0.0169)	(0.0106)	(0.0122)	(0.0104)	(0.00625)	(0.00938)	(0.0111)
Prevention	-0.0716***	-0.0545***	-0.0532***	-0.0527***	-0.0501***	-0.00189	-0.0877***
	(0.0170)	(0.0106)	(0.0129)	(0.0103)	(0.00624)	(0.00954)	(0.0111)
Control mean	0.400	0.499	0.301	0.492	0.392	0.112	0.200
N	2796	2796	2796	2796	2796	2796	2796
R-squared	0.194	0.119	0.080	0.132	0.161	0.282	0.314

NOTES: The table reports the main specification estimated separately within each gender group, stacked as two panels: Panel A is women and non-binary respondents, Panel B is men. The heterogeneity is the difference in a coefficient between the panels; the table reports no interaction term, so that difference carries no standard error of its own. The unit of observation is a respondent who completed the main survey, pooled across Prolific, CloudResearch and MTurk with platform fixed effects. Columns are the AI-use measures on their common 0–1 scale, in the same order as Appendix Table B.1: the pre-specified rule, the four model indices, self-report and Pangram. The PCA indices are percentile ranks, so their coefficients are rank points, while the pre-specified rule, the LASSO indices, Pangram and self-report coefficients are probability or rate points. Controls, not displayed: age, gender, race, income, baseline AI attitudes, and wave fixed effects. Robust standard errors in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table B.8: Heterogeneity by Income

	Pre-spec	PCA (CELSS)	LASSO (CELSS)	PCA (Training)	LASSO (Training)	Self-Report	Pangram
Panel A: Below-median income							
Base Pay	-0.0000984	-0.00138	0.0143	-0.00381	0.000459	-0.00335	0.00620
	(0.0154)	(0.00965)	(0.0117)	(0.00943)	(0.00542)	(0.00887)	(0.0102)
Bonus Amount	0.0144**	0.00646	0.00154	0.00582	0.00315	0.00370	0.00437
	(0.00671)	(0.00426)	(0.00521)	(0.00415)	(0.00239)	(0.00388)	(0.00441)
Sanction	-0.0438***	-0.0299***	-0.0425***	-0.0316***	-0.0225***	-0.0216**	-0.0200*
	(0.0154)	(0.00964)	(0.0115)	(0.00942)	(0.00539)	(0.00882)	(0.0102)
Prevention	-0.0645***	-0.0421***	-0.0401***	-0.0416***	-0.0465***	-0.0154*	-0.0704***
	(0.0157)	(0.00975)	(0.0120)	(0.00955)	(0.00551)	(0.00880)	(0.0104)
Control mean	0.456	0.483	0.287	0.475	0.399	0.101	0.232
N	3216	3216	3216	3216	3216	3216	3216
R-squared	0.227	0.119	0.099	0.132	0.264	0.188	0.405
Panel B: Above-median income							
Base Pay	0.000987	-0.00731	-0.00350	-0.0118	-0.00666	-0.0139	-0.0243*
	(0.0183)	(0.0117)	(0.0147)	(0.0115)	(0.00675)	(0.0105)	(0.0125)
Bonus Amount	0.0282***	0.0180***	0.0178***	0.0184***	0.00543*	0.00266	0.00354
	(0.00788)	(0.00501)	(0.00609)	(0.00491)	(0.00295)	(0.00447)	(0.00530)
Sanction	-0.0826***	-0.0508***	-0.0547***	-0.0486***	-0.0251***	-0.0283***	-0.0245*
	(0.0184)	(0.0117)	(0.0145)	(0.0115)	(0.00675)	(0.0105)	(0.0126)
Prevention	-0.0775***	-0.0449***	-0.0491***	-0.0457***	-0.0478***	-0.00430	-0.0937***
	(0.0184)	(0.0115)	(0.0145)	(0.0113)	(0.00678)	(0.0106)	(0.0125)
Control mean	0.369	0.479	0.294	0.475	0.378	0.124	0.198
N	2352	2352	2352	2352	2352	2352	2352
R-squared	0.191	0.131	0.091	0.150	0.161	0.299	0.283

NOTES: The table reports the main specification estimated separately within each income group, split at the sample median and stacked as two panels. Sample, columns, scales, controls and standard errors are as in Table B.7. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.9: Heterogeneity by Age

	Pre-spec	PCA (CELSS)	LASSO (CELSS)	PCA (Training)	LASSO (Training)	Self-Report	Pangram
Panel A: Below-median age							
Base Pay	-0.00400	0.00650	0.0184	0.00526	0.00174	-0.00299	-0.00204
	(0.0169)	(0.0107)	(0.0135)	(0.0106)	(0.00632)	(0.0101)	(0.0124)
Bonus Amount	0.0247***	0.0107**	0.0135**	0.0108**	0.00430	0.00338	0.00579
	(0.00736)	(0.00475)	(0.00589)	(0.00467)	(0.00283)	(0.00436)	(0.00534)
Sanction	-0.0375**	-0.0275**	-0.0435***	-0.0290***	-0.0191***	-0.0128	-0.0104
	(0.0169)	(0.0108)	(0.0134)	(0.0106)	(0.00638)	(0.0101)	(0.0123)
Prevention	-0.0401**	-0.0393***	-0.0457***	-0.0398***	-0.0415***	-0.000460	-0.0768***
	(0.0172)	(0.0107)	(0.0136)	(0.0106)	(0.00644)	(0.0102)	(0.0125)
Control mean	0.451	0.511	0.322	0.505	0.414	0.105	0.276
N	2798	2798	2798	2798	2798	2798	2798
R-squared	0.217	0.092	0.089	0.107	0.210	0.190	0.373
Panel B: At or above median age							
Base Pay	0.00283	-0.0135	-0.000923	-0.0177*	-0.00600	-0.0118	-0.00754
	(0.0165)	(0.0104)	(0.0127)	(0.0101)	(0.00561)	(0.00909)	(0.00994)
Bonus Amount	0.0163**	0.0123***	0.00369	0.0118***	0.00396	0.00427	0.00319
	(0.00707)	(0.00447)	(0.00565)	(0.00434)	(0.00247)	(0.00391)	(0.00427)
Sanction	-0.0801***	-0.0496***	-0.0503***	-0.0474***	-0.0270***	-0.0409***	-0.0319***
	(0.0165)	(0.0104)	(0.0125)	(0.0101)	(0.00565)	(0.00901)	(0.0101)
Prevention	-0.103***	-0.0504***	-0.0444***	-0.0506***	-0.0535***	-0.0230**	-0.0888***
	(0.0165)	(0.0103)	(0.0127)	(0.0101)	(0.00565)	(0.00902)	(0.0100)
Control mean	0.382	0.449	0.254	0.442	0.364	0.118	0.152
N	2770	2770	2770	2770	2770	2770	2770
R-squared	0.180	0.118	0.053	0.136	0.167	0.286	0.274

NOTES: The table reports the main specification estimated separately within each age group, split at the sample median and stacked as two panels. Sample, columns, scales, controls and standard errors are as in Table B.7. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.10: Treatment Effects on Attentiveness

	(1) Video check	(2) IMC	(3) Both
Base Pay	0.0261***	0.0244**	0.0370***
	(0.00937)	(0.0121)	(0.0123)
Bonus Amount	0.00612	-0.00377	-0.00188
	(0.00411)	(0.00530)	(0.00537)
Sanction	-0.000703	-0.0111	-0.0131
	(0.00937)	(0.0121)	(0.0123)
Prevention	-0.0122	-0.00856	-0.0138
	(0.00948)	(0.0122)	(0.0123)
Control mean	0.836	0.539	0.463
N	5568	5568	5568
R-squared	0.033	0.190	0.165

NOTES: The table reports treatment effects on the two attention checks. The unit of observation is a respondent who completed the main survey, pooled across Prolific, CloudResearch and MTurk with platform fixed effects. Each column is a separate regression. "Video check" is an indicator for answering the video comprehension question correctly, "IMC" for passing the classic instructional manipulation check, and "Both" for passing both, missing where either was not asked. All three outcomes are coded 1 = attentive, so a negative coefficient means the arm reduced attentiveness, and coefficients are linear-probability effects in percentage points. Base pay is \$15/hr against \$8/hr and the bonus enters linearly in dollars. Controls, not displayed: age, gender, race, income, the four baseline AI attitudes, and wave fixed effects. Robust standard errors in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table B.11: Treatment Effects on Submission Attrition

	(1) Approved	(2) Returned	(3) Timed-out	(4) Rejected	(5) Awaiting
Base Pay	0.0226**	-0.0119	-0.00922**	-0.00149*	0.0000133
	(0.0112)	(0.0107)	(0.00410)	(0.000811)	(0.000655)
Bonus Amount	0.000890	-0.00366	0.00201	0.000527	0.000228
	(0.00507)	(0.00493)	(0.00192)	(0.000417)	(0.000395)
Sanction	0.00630	0.000524	-0.00604	-0.000884	0.0000964
	(0.0121)	(0.0115)	(0.00382)	(0.000922)	(0.000733)
Prevention	0.0163	-0.0221*	0.00685	-0.00192*	0.000941
	(0.0121)	(0.0116)	(0.00430)	(0.00103)	(0.000661)
Control mean	0.880	0.110	0.010	0.000	0.000
N	3881	3881	3881	3881	3881
R-squared	0.076	0.065	0.014	0.003	0.004

NOTES: The table reports treatment effects on how a submission ends. The unit of observation is a Prolific submission, and unlike every other results table the sample is all submissions rather than completers only, so returned and timed-out respondents are included. The sample is Prolific only, because the five outcomes are Prolific submission statuses and have no MTurk or CloudResearch analogue. Each column is a separate regression on one status share, so coefficients are linear-probability effects. Wave fixed effects are included; no demographic controls, because those are post-selection and unavailable for non-completers. Standard errors clustered at the listing level in parentheses. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table B.12: Attrition by Treatment Arm and Platform

Platform	Contrast	Control	Treated	Difference	p
Prolific	Base pay: \$15 v \$8	0.224	0.212	-0.012	0.388
	Bonus \$0.25 v none	0.231	0.216	-0.015	0.357
	Bonus \$2.50 v none	0.231	0.208	-0.023	0.178
	Sanction v none	0.224	0.212	-0.011	0.402
	Prevention v none	0.207	0.229	0.023	0.092
MTurk	Base pay: \$15 v \$8	0.193	0.177	-0.015	0.441
	Bonus \$0.25 v none	0.204	0.198	-0.005	0.821
	Bonus \$2.50 v none	0.204	0.177	-0.027	0.271
	Sanction v none	0.202	0.185	-0.017	0.401
	Prevention v none	0.116	0.275	0.159	0.000
CloudResearch	Base pay: \$15 v \$8	0.209	0.165	-0.045	0.015
	Bonus \$2.50 v none	0.196	0.181	-0.015	0.411
	Sanction v none	0.191	0.186	-0.005	0.784
	Prevention v none	0.203	0.176	-0.026	0.151

NOTES: The table reports the share of consented starters who did not complete the main survey, by treatment arm and platform. The unit of observation is a consented starter, so unlike most tables in this section the sample is not restricted to completers. Each row is one randomized contrast: "Control" is the mean non-completion rate in the omitted arm, "Treated" the mean in the named arm, "Difference" the gap, and p the two-sided p-value of a two-sample t-test on that difference. The two bonus tiers are each compared against the no-bonus arm, so the two rows share a control group. CloudResearch has no \$0.25 row because it did not field that tier. Base pay is entry-stamped on Prolific, assigned by intention to treat on MTurk, and taken from the assignment lists on CloudResearch. Because each row is a separate test, the p-values are not adjusted for multiple comparisons.

Table B.13: Follow-up Completion by Main-Survey Arm, Prolific

Main-survey arm	Invited	Completed	Completion (%)	p
Base pay: low	1851	1649	89.1	0.066
Base pay: high	1874	1633	87.1	
Bonus: none	1216	1076	88.5	0.064
Bonus: \$0.25	1313	1173	89.3	
Bonus: \$2.50	1196	1033	86.4	
Sanction: no	1923	1685	87.6	0.346
Sanction: yes	1802	1597	88.6	
Prevention: no	1853	1644	88.7	0.250
Prevention: yes	1872	1638	87.5	

NOTES: The table reports completion of the follow-up survey by the arm a respondent was assigned in the main survey. The unit of observation is a respondent invited to the follow-up; the sample is Prolific only, because the follow-up was fielded there alone. "Invited" is the number of respondents in the arm who were invited back and "Completed" the number who finished the follow-up, so "Completion" is the ratio of the two. Each block is one randomized intervention, and the p-value printed on a block's first row is a single test across all of that block's arms rather than a test of the row it sits on. Because the randomizations are cross-cutting, every respondent appears once in each block, so the blocks are four views of the same sample and their p-values are not independent. The effect of base pay on follow-up completion is estimated separately with controls and answers a different question from the main-survey randomization tested here.

Table B.14: Treatment Effects on AI-Automated Responses

	Prolific		MTurk		CloudResearch	
	(1)	(2)	(3)	(4)	(5)	(6)
	No-signal = 0	No-signal dropped	No-signal = 0	No-signal dropped	No-signal = 0	No-signal dropped
Base Pay: \$15 v \$8	-0.00873	-0.0103	0.0373**	0.0462**	-0.0128	-0.0146
	(0.00611)	(0.00699)	(0.0185)	(0.0222)	(0.0100)	(0.0112)
Bonus \$0.25	-0.00943	-0.0110	-0.00984	-0.0168		
	(0.00759)	(0.00868)	(0.0229)	(0.0277)		
Bonus \$2.50	-0.00281	-0.00427	-0.0312	-0.0444	-0.00154	-0.00303
	(0.00787)	(0.00899)	(0.0228)	(0.0274)	(0.00982)	(0.0110)
Sanction	0.00791	0.00866	0.00803	0.00823	-0.00939	-0.00990
	(0.00616)	(0.00708)	(0.0185)	(0.0222)	(0.00991)	(0.0110)
Prevention	0.00198	0.00326	-0.0655***	-0.0560**	-0.0192*	-0.0220*
	(0.00578)	(0.00659)	(0.0184)	(0.0221)	(0.0101)	(0.0113)
Constant	0.0344***	0.0396***	0.173***	0.203***	0.0656***	0.0742***
	(0.00782)	(0.00891)	(0.0237)	(0.0279)	(0.0119)	(0.0134)
Control mean	0.030	0.034	0.197	0.227	0.070	0.081
Pre-spec not computable (attriters)	430	0	211	0	180	0
N	3462	3031	1488	1217	1692	1517
R-squared	0.002	0.002	0.012	0.011	0.004	0.005

NOTES: The table reports treatment effects on the AI-automated pre-specified binary, extended with the in-survey screens. The unit of observation is a consented starter who carried a treatment stamp and reached the randomized task blocks, so unlike the AI-assisted tables the sample is not restricted to completers. The extended rule flags a respondent who stopped at the captcha, stopped at the video, or completed the survey and was flagged by the pre-specified binary; answering the video incorrectly is deliberately excluded, because it captures inattention rather than automation. The pre-specified binary needs the three core typed tasks, so attriters who did not stall at a screen carry no signal: the first column for each platform codes them zero and the second drops them. Base pay is entry-stamped on Prolific, assigned intention-to-treat on MTurk, and taken from the assignment lists on CloudResearch, which did not field the \$0.25 tier. Standard errors are clustered at the listing level on Prolific and robust elsewhere. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table B.15: The Base-Rate Fallacy, Raw and Reweighted

	Unweighted	Human-only	Difference	N
Correct posterior (57%)	0.129	0.102	-0.027	2955
		(0.006)		
Base rate (25%)	0.219	0.233	0.014	2955
		(0.009)		
Likelihood, Jar A (80%)	0.090	0.095	0.005	2955
		(0.006)		
Likelihood, Jar B (20%)	0.098	0.102	0.004	2955
		(0.005)		
Fifty-fifty (50%)	0.079	0.082	0.002	2955
		(0.005)		
Any other answer	0.384	0.386	0.002	2955
		(0.009)		

NOTES: The table reports the share of respondents giving each answer category on the hard probability question, raw against reweighted to human respondents. The unit of observation is a respondent who completed the main survey; the sample is Prolific only, for the same reason as Table 8. The weight is $w_i = 1 - p_i^{\text{field}}$, using the same base-rate corrected posterior as that table. The question is a ball-and-urn problem with $P(\text{Jar A}) = 0.25$, $P(\text{green} \mid A) = 0.80$ and $P(\text{green} \mid B) = 0.20$, so the correct posterior is 57%; 25% is the base rate, 80% and 20% are the two likelihoods, and 50% is the uninformative answer. A respondent answering 25 has anchored on the base rate and ignored the evidence; one answering 80 or 20 has anchored on a likelihood and ignored the base rate. Bootstrap standard errors in parentheses, from 500 replications resampling both the labeled and the field sample.

Table B.16: AI Use by Platform and Measure

		Naive		Corrected AI rate (%)	
Platform	N	mean score	Separation	Full sample	Control arm
AI-assisted, CELSS-trained common-slot no-effort LASSO					
Prolific	2955	0.226	0.726	13.0 (1.9)	13.4 (4.4)
CloudResearch	1456	0.214	0.726	11.2 (1.9)	17.5 (5.4)
MTurk	1169	0.374	0.726	33.3 (1.8)	36.7 (6.0)
Pooled	5580	0.254	0.726	16.8 (1.7)	19.2 (3.2)
CELSS lab (human/AI)		*0.132 / 0.859*	0.726	*lab base rate 50.0%*	

NOTES: The table reports implied AI use by platform and detection measure, naive and corrected for the shifted base rate by the procedure in Section 3.7. Each measure is summarized by its average on its own scale rather than by the share above a cutoff: the flag rate for the binary measures and the mean predicted probability for the LASSO and Pangram measures. Each panel is one index variant; the lab row of a panel gives the two type-conditional means the correction is identified from, together with the training sample's own AI rate. The no-effort index is not built on CloudResearch, which is why the common-slot panels follow it. Bootstrap standard errors in parentheses, from 500 replications resampling both samples, clustered at the listing level on Prolific.